THÈSE

En vue de l'obtention du

DOCTORAT de L'UNIVERSITE de TOULOUSE

Délivré par : l'Université Toulouse III – Paul Sabatier

Spécialité : Génétique des Populations

Présentée et soutenue par Natacha NIKOLIC

Le 15 juillet 2009

Diversité génétique et taille efficace chez les populations de poissons sauvages : le cas du Saumon atlantique un poisson migrateur amphihalin menacé.

JURY

P. BERREBI	Professeur, Université de Montpellier II, Montpellier	Rapporteur
E. GARCIA-VAZQUEZ	Professeur, Université d'Oviedo, Espagne	Rapporteur
M. GARDES	Professeur, Université Paul Sabatier, Toulouse	Présidente du Jury
H. HARTMANN	Maître de Conférences, Université de La Rochelle	Examinateur
S. LAUNEY	Chargée de Recherche, INRA, Rennes	Examinatrice
J. RIQUET	Chargée de Recherche, INRA, Toulouse	Examinatrice
C. CHEVALET	Directeur de recherche, INRA, Toulouse	Directeur de thèse

---

Ecole Doctorale Sciences Ecologiques, Vétérinaires, Agronomiques, et Bioingénéries

UMR444, INRA, chemin de Borde Rouge BP 52627 Auzeville, 31326 Castanet-Tolosan, France.

THESE

Diversité génétique et taille efficace chez les populations de poissons sauvages : le cas du Saumon atlantique un poisson migrateur amphihalin menacé

PHD THESIS

Genetic diversity and effective size of wild fish populations: The case of Atlantic salmon a diadromous threatened migratory fish

Natacha NIKOLIC, 2009

Thèse de doctorat
Ecole Doctorale Sciences Ecologiques, Vétérinaires, Agronomique et Bioingénieries (SEVAB)
Université Paul Sabatier

Financement :
Bourse du Ministère de la Recherche

Encadrant principal :
Claude CHEVALET, Directeur de Recherche au Laboratoire de Génétique Cellulaire – Institut National de la Recherche Agronomique, Chemin de Borde Rouge, 31326 Castanet Tolosan, France
E-mail : claude.chevalet@toulouse.inra.fr

Auteur:
Natacha NIKOLIC, Institut National de la Recherche Agronomique, Chemin de Borde Rouge, 31326 Castanet Tolosan, France
E-mail : natachanikolic@hotmail.com

Avant tout, je tiens à remercier les membres du jury pour avoir consacré de leur temps à l'évaluation de ce manuscrit. Merci d'une part à Monique Gardes d'avoir accepté de présider cette thèse et d'autre part à Patrick Berrebi et Eva Garcìa-Vazquez de me faire l'honneur d'être rapporteurs.

Tout le travail présenté dans ce manuscrit n'aurait pas été possible sans l'aide, l'encadrement et le soutien de beaucoup de personnes. Ce petit prélude me permet de remercier ceux dont je n'ai pas eu l'occasion de vive voix et ceux dont on ne dit jamais assez à quel point ils sont précieux.

Passionnée de théâtre, je me suis amusée à nommer toutes les personnes qui m'ont aidée, conseillée et épaulée comme une pièce. Je n'oublierai personne sur scène, en coulisse, en régie, dans le trou du souffleur, au jeu d'orgues, dans la corbeille et au parterre.

En priorité d'ailleurs, je voulais exprimer toute ma gratitude à Claude Chevalet grâce à qui ce travail est ce qu'il est aujourd'hui. Comme un metteur en scène il m'a guidée et m'a enseignée à la craie, noir sur blanc, des déductions mathématiques comme un tableau de Guernica.

Cette pièce s'est déroulée en deux actes. Dans le premier se trouvent deux intervenants capitaux sans lesquels je n'aurais pu y arriver. Le régisseur Juliette Riquet pour qui j'ai une affection profonde car humainement on n'en trouve pas deux comme elle. Elle m'a transmis beaucoup de son savoir avec bonne humeur et elle a plus que rempli sa fonction de préparation, de coordination, d'exécution de la partie spécifique « moléculaire ». Heu, moléculaire au théâtre c'est un objet qui permet de s'occuper pendant des heures, jusqu'à très tard la nuit. Puis le souffleur du labo, Katia Fève, qui se rappelle de toutes les réparties, répliques, tirades, monologues, apartés, stichomythies, didascalies, vers, arguments, préambules, comédies, tragédies, farces indispensables à tous les acteurs. Dans mon rôle de cuisinière, elle a su me transformer de Maïté à Marc Veyrat.

J'ai pu également apprécier le rôle de conseiller, côté jardin, de Jean-Luc Baglinière. Merci encore pour toutes ses conversations qui m'ont permis d'avancer sur mon projet et pour ton soutien. Merci également à Alexandre Dewez, Simon Boitard, Bertrand Servin, Guillaume Evannot et Maxime Bonhomme pour leurs lectures et leurs conseils. Ce sont de vrais pédagogues. Sans oublier Matthias Macé qui m'a fait gagner du temps en m'aidant sur certains logiciels.

Grâce à Manuela Ferré et Florence Comayras, nous avons pu avoir une mise en place du matériel et des techniques d'éclairage, SPECTACULAIRE pour que l'ambiance soit lumineuse. Parmi les autres techniciens, il ne faut pas oublier une dame débordant de gentillesse, Solange Cassette, que je remercie amplement pour toutes ces recherches de manuscrits datant de la Bibliothèque du Congrès.

J'appelle maintenant à la cour, le machiniste, Eddie Iannuccelli qui grâce à ces talents italiens tel que Giacomo Torelli a su équiper et activer le décor. Ainsi que le magicien Patrice Dehais. Merci pour vos débogages et votre patience à réparer nos bêtises. Je dis « nos » car je pense que je ne suis pas la seule, oups.

Je remercie également Joël Gellin, qui voulait tenir le rôle de cabotin, mais par son talent et son inspiration dramatique ne peut dans aucun cas tenir ce rôle. Merci pour ta bonne humeur.

Merci au directeur artistique, Philippe Mulsant, de m'avoir permis de réaliser ce projet passionnant et au Département de Génétique animale de m'avoir soutenu pour finaliser mes travaux au sein du Laboratoire. Merci également à Denis Milan, Monique Falières et à la Plateforme Génomique de Toulouse pour leur aide surtout ces derniers mois de thèse.

Je me tourne maintenant vers le parterre pour remercier ma mère Danielle, mon père François et ma grand-mère Emilienne de m'avoir toujours soutenue et laissée faire mes propres choix sans jamais me juger. Merci également à mon frère Nicolas, ma marraine Géraldine, mes tantes (Zoritza, Monique, Henriette, Annie), mes oncles (Gilles, Michel, Gérard, Etienne, Serge, Guy), mes cousines (Fanny, Laura, Christelle, Vanessa, Laetitia, Laurence, Valérie, Karine, Amandine) et mes cousins (les deux Freddy, Mickael, Manivone, Marco, Vincent et Stéphane) pour m'avoir remontée le moral et fait comprendre que dans la famille « groseille » on n'est jamais seul, on est une baie bien soudé.

Pour finir, je remercie la corbeille, eh oui c'est comme ça que l'on dit au théâtre. Ça n'a peut-être pas l'air élégant mais pourtant sans la corbeille le chœur n'est plus le même. Mille mercis à Thierry, Isabelle et Daniel qui ont été très présents dans ma vie ces trois dernières années. Merci à tous mes amis, Laurent, Emilie, Julie, Fabrice, Cécile, Yannick, Laetitia, Sandrine, Caroline, Fanny, Irina, Sonia, Mathieu, Guillaume et Grégory pour leur affection et leur réconfort sans limite.

Sans oublier les spectateurs, avec qui j'ai passé des moments inoubliables et qui m'ont aidée à ne pas craquer, Mireille, Jérôme (après Wonder Woman a choisi le rôle de Charlie et ses drôles de dames), avec Sophie (la pile sur pattes), Marion (rougail, cocktail et soleil), Caroline (l'amour est dans les prés), Manuela (ça se discute) et Maguy (animatrice santé et bien-être). Après la pile, voici Flavie la mobylette, Marine (ennemie de Brigitte Bardot depuis sa tuerie sur les poussins), David Croquette, Nathalie (la speakeuse des z'amours), l'autre Nathalie (présentatrice dans Télé-Matin des produits de couleurs inattendues pour cheveux), Marie & Marc (Futurs présidents d'ARTE), Laurette (devenu couturière suite à un incident très embarrassant de pantalon), Aurélie (actrice dans massacre à la tronçonneuse avec Roger Rabbit), Valérie (trente millions d'amies), Alain (Brian is in the kitchen), Annie, Pitou (et oui ! complexement), Sophie (accessoiriste), Charles (cheveux-au-vents dans Danse avec les Loups), Laurence (le petit routard) et Philippe (vendeur de rêve pour piscine).

Merci encore à vous tous.

A ma grand-mère Véra,

Ne znam ako je uzbudljivije od Columbo epizoda ali nadam se da ste ponosni na mene.
Volim vas bakica, zelim da ste bili kod mene.

Préambule

Cette thèse s'intègre dans le cadre d'une obtention du titre de Docteur en Sciences au sein de l'Ecole Doctorale SEVAB (Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries). Les travaux réalisés au cours de cette thèse se sont portés sur la diversité génétique et la taille efficace chez les populations de poissons sauvages avec pour cas d'étude des populations de Saumon atlantique migrateurs amphihalin menacées. Ces études m'ont permis de développer des compétences pluridisciplinaires tel que la biologie moléculaire lors de l'établissement du panel de microsatellite et des génotypages, la génétique quantitative et des populations lors de l'analyse de la diversité et structure des populations, la biostatistique et la bioinformatique lors du développement des modèles de diversité génétique et d'estimation de taille efficace.

Ce travail a été réalisé au sein du Laboratoire de Génétique Cellulaire au Département de Génétique Animale de l'INRA de Castanet Tolosan, sous la responsabilité de Claude Chevalet entre Octobre 2005 et Mai 2009. Il n'a été précédé d'aucun travail sur le sujet mais mon travail de Master Recherche m'a permis de me familiariser avec les outils en diversité et différenciation génétique (voir Nikolic et al. 2009c).

Il n'aurait pas été possible sans l'investissement des équipes de l'Unité mixte de recherche Ecobiologie et qualité des hydrosystèmes continentaux (INRA-AgroCampus Rennes), Spey Fishery Board Research Office (Ecosse) et Fisheries Research Services Perthshire (Ecosse) qui conservent depuis plus de vingt ans les écailles de Saumon atlantique récoltés dans plusieurs rivières et affluents de leur région.

Les résultats issus de cette recherche ont été valorisés par trois articles (Nikolic et al. 2009a, b et Chevalet et Nikolic 2009). Un quatrième article est en cours et sera l'aboutissement de ce travail, puisque les données génétiques des populations étudiées seront analysées par le modèle d'estimation de l'évolution des tailles efficaces développé au cours de ces travaux.

Résumé

Ces travaux de thèse se sont intéressés principalement à la diversité génétique et à la taille efficace (*Ne*) de stocks de poissons sauvages amphihalins migrateurs en danger, le Saumon atlantique (*Salmo salar*). Pour cela, quatre populations ont été choisies en Europe du Nord (Ecosse) et en Europe du Sud (France) pour leurs différences de structure et de mode de gestion.
Ces travaux ont abouti au développement de deux nouveaux modèles. *DemoDivMS* prédit la diversité génétique d'une population dont le scénario évolutif, donné par l'utilisateur, décrit des variations de *Ne* au cours des générations antérieures. *VarEff* cherche à estimer les tailles efficaces, au moment de l'échantillonnage et passées, en s'appuyant sur des calculs analytiques directs raccourcissant les temps de calculs.

Abstract

This thesis was concerned mainly with the genetic diversity and the effective size (*Ne*) of wild fish stocks of diadromous migratory endangered Atlantic salmon (*Salmo salar*). For this, four populations were chosen in the northern (Scotland) and southern (France) part of Europe for their differences in their structure and management.
This work led to the development of two new models. *DemoDivMS* predicted genetic diversity of a population whose evolutionary scenario, given by the user, describes variations of *Ne* during the previous generations. *VarEff* tries to estimate the effective sizes, sampling time and pasts, based on direct analytical calculations to shorten the time of calculations.

PLAN

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

D’une manière générale, l’état des populations de poissons ne se résume pas seulement à leur abondance mais également à leur viabilité vis-à-vis des pressions exercées à leur égard. Cependant à ce jour, les régulations de l’exploitation des ressources naturelles se basent principalement sur les estimations d’abondance des stocks et négligent les paramètres renseignant sur leurs évolutions. Les indices de captures et d’effort de pêche pour juger de la durabilité d’un stock sont les plus utilisés pour en déduire des totaux autorisés de captures (TAC) (captures maximales annuelles) en mer par des modèles tel que celui de Graham et Shaefer (Graham 1935, Shaefer 1954, 1957). Si une année 40 bateaux de pêche ont été nécessaires pour prélever 2000 tonnes de poisson, contre 60 bateaux l’année suivante, alors ils en déduisent que l’effort de pêche est plus important et de ce fait que le stock est surexploité, et ainsi réévaluent la TAC à la baisse. Toutefois, il apparaît que ces estimations sont un mauvais indicateur de l’abondance des stocks (Cury et Miserey, 2008). La stabilité des captures ne témoignerait pas de la viabilité et de la durabilité d’une ressource. Si nous prenons l’exemple des morues au Canada, les politiques de gestion jugeaient les stocks à l’équilibre, entre leur exploitation et leur renouvellement, avant leur effondrement brutal en 1992. Il semble donc impératif d’estimer des paramètres reflétant à la fois la taille des stocks et la viabilité des populations.

L’abondance des populations de poissons, notamment migratoires, est difficile à estimer compte tenu de la taille des habitats colonisés et de leurs parcours migratoires complexes dans le temps et dans l’espace. Face à ces difficultés, les scientifiques travaillent sur des paramètres alternatifs qui permettraient d’évaluer la taille ainsi que l’état de santé de ces populations. Grâce aux avancées technologiques, informatiques et moléculaires, l’information contenue dans les gènes peut être exploitée pour renseigner sur la viabilité d’une population. Certaines séquences génétiques sont utilisées comme horloge moléculaire pour retracer l’histoire évolutive des populations. Un des paramètres importants conditionnant cette histoire évolutive est celui de la taille efficace (*Ne*) actuelle et ancestrale renseignant sur l’évolution globale des populations et leur potentiel évolutif. Dans la littérature ce paramètre est mentionné comme primordial en conservation pour comprendre et prédire le devenir des populations, tant sur le court terme que sur le long terme. Comme son nom l’indique ce paramètre représente le nombre d’individus qui « agissent » (qui sont « efficaces ») pour faire perdurer la population. Il peut être considéré comme le nombre de géniteurs d’une population transmettant effectivement leurs gènes à leurs descendants. La taille efficace peut être estimée à partir de données génétiques mais également démographiques qui sont plus coûteuses et

difficiles à obtenir. Malgré l'importance et la commodité d'estimer la taille efficace à partir des données génétiques, ce paramètre est très peu exploité dans les programmes de suivi et de conservation de populations en danger, telles que le Saumon atlantique (*Salmo salar*) sauvage.

Ainsi, il paraissait intéressant de développer un travail utilisant des outils en génétique des populations en vue de faciliter l'estimation des états des populations de poissons. L'intérêt de travailler sur ces nouveaux outils était de s'abstraire d'une certaine façon des données démographiques pas toujours évidentes à collecter chez les populations aquatiques. Ainsi, le travail de thèse s'est déroulé selon deux volets.

Le premier volet s'est focalisé sur les potentialités et les sensibilités des méthodes d'estimations de taille efficace à partir de données génétiques de populations de Saumon atlantique intégrées dans des programmes de conservation. Nous avons fait le choix de travailler sur des populations de Saumon sauvage et anadrome, tout d'abord, parce que les stocks s'effondrent et ensuite, parce que leur phylopatrie (retour dans leur rivière natale pour se reproduire) permet l'identification des populations. A chaque bassin hydrographique correspond une population isolée c'est-à-dire un stock. En plus de pouvoir identifier les populations, l'utilisation des écailles rend possible l'estimation de l'âge individuel, ce qui permet de définir les cohortes et ainsi éviter les générations chevauchantes qui peuvent fortement limiter les analyses en génétique des populations.

Les estimateurs de taille efficace étudiés au sein de ces travaux de thèse sont des méthodes bayésiennes de coalescence utilisant des séquences microsatellites. La coalescence fait référence à la description de la réunion des lignées phylogénétiques de séquences orthologues dans une population. La généalogie des séquences étant décrite dans un arbre, les évènements de coalescence font référence aux noeuds des arbres. Les méthodes phylogénétiques sont plus efficaces que les méthodes non-phylogénétiques (Felsenstein 1992), principalement en raison du complément d'information fourni par la structure de l'arbre. De plus, la modélisation généalogique a grandement facilité l'estimation des paramètres démographiques et de mutations en utilisant la longueur des différences de microsatellite (Chakraborty et Kimmel 1999, Feldman et al. 1999, King et al. 2000 par Storz et Beaumont 2002). Il paraissait ainsi évident de tester des estimateurs de taille efficace utilisant la coalescence et l'information apportée par les microsatellites. Analyser ces méthodes sur des données réelles demande un effort dans le choix des populations de Saumon

et dans l'élaboration du panel de marqueurs microsatellites. Les populations ont ainsi été choisies très différentes en termes de taille, de diversité et de structure génétique. Il s'agit de deux populations Françaises (Oir et Scorff) et deux Ecossaises (Shin et Spey). Devant la quantité de marqueurs microsatellites développés chez le Saumon atlantique, un ensemble de marqueurs a été sélectionné et testé pour aboutir à un large panel polymorphe permettant des évaluations fines de la variabilité et de la structure des populations étudiées. L'ensemble de ces travaux fait l'objet du premier volet de cette thèse et permet de fournir des propositions d'utilisations d'estimateurs de taille efficace à partir de données génétiques pour les programmes de conservation.

Néanmoins ces résultats ne permettent pas de conclure sur le devenir des populations. Afin de retracer l'histoire évolutive entre le moment de l'échantillonnage et l'ancêtre commun, le deuxième volet de cette thèse s'est consacré au développement d'un nouveau modèle de coalescence, *VarEff*, cherchant à repérer les fluctuations ancestrales de la taille efficace des populations. Un point fort de ce modèle est d'offrir des temps de calculs courts car les estimations s'appuient en partie sur des résolutions analytiques, sans passer par la reconstruction des arbres de coalescence par simulation. Certaines résolutions analytiques de ce modèle ont été utilisées au sein du premier volet de thèse pour développer un autre modèle, *DemoDivMS*, qui permet de déduire la diversité génétique (en terme d'hétérozygotie) selon l'histoire démographique d'une population. Ce modèle a ainsi permis de comprendre le maintien de la forte diversité génétique observée au sein des petites populations (Oir et Scorff) pour le premier volet de la thèse.

Le développement du modèle, *VarEff*, a pour objectif de retracer l'histoire évolutive des populations étudiées pour comprendre leur diminution en comparant les fluctuations observées avec celles de données abiotiques comme la pression de pêche ou les variations des pressions atmosphériques Nord Atlantique NAO (North Atlantic Oscillation). A la suite de tests sur les performances de ce modèle avec des données simulées, nous espérons l'appliquer aux données microsatellite des populations de Saumons étudiées pour mieux comprendre l'origine de leur diminution ces dernières décennies.

Ce mémoire de thèse présente des études sur la diversité et la structure de populations chez le Saumon atlantique, en portant une attention particulière sur la taille efficace. Ces analyses ont soulevé des points importants dans l'utilisation des méthodes pour leur

application en conservation et exposé les perspectives offertes par le nouveau modèle d'estimation de taille efficace développé au cours de ces travaux de thèse.

I

CONTEXTE SCIENTIFIQUE ET CONCEPTUEL

Le Saumon atlantique (*Salmo salar*) est une espèce migratrice amphihaline anadrome et sténotherme d'eau froide (MacCrimmon & Gots 1979 ; Porcher & Baglinière 2001). Cette espèce est considérée comme en danger du fait de la diminution de son aire de répartition et de l'abondance de ses stocks (ICES, 2001 & 2003 ; Caron et Fontaine, 2003). Elle est donc inscrite depuis le début de la décennie 1990 sur la liste rouge des espèces menacées en France et en Europe (Porcher & Baglinière 2001). Le déclin du Saumon résulte de causes multiples, cumulatives, liées aux stress anthropiques : multiplication des barrages, dégradation de la qualité de l'eau et de l'habitat en rivière entraînant une modification de la productivité des cours d'eau et taux d'exploitation inadapté (Saunders 1981 ; Baglinière *et al*. 1990; Parrish *et al*. 1998 ; Schindler 2001). En Europe, cela s'est traduit par l'éradication de l'espèce de la plupart des grands bassins fluviaux tels que l'Elbe, la Tamise, le Rhin, la Seine, la Garonne et la Dordogne (MacCrimmon & Gots 1979 ; Thibault 1994). De fait, des programmes de restauration ont été mis en place sur de nombreux grands fleuves de l'aire de répartition originelle de l'espèce (Baglinière *et al.* 1990 ; Ayllon *et al.* 2006).

A ces menaces se sont ajoutées des menaces plus récentes liées à l'impact du changement climatique qui ont pu se cumuler aux modifications des usages des bassins versants. Ces impacts se sont traduits, pour la majorité des stocks de saumons, par une très forte diminution de la composante d'individus de plusieurs hivers de mer (ICES, 2001). Il s'est également traduit par une contraction du cycle de vie (diminution du temps de séjour en rivière et en mer) entraînant un taux de renouvellement plus rapide des populations de saumons (Baglinière *et al.* 2004 ; Aprahamian *et al.* 2008).

Cette espèce présente des intérêts et des enjeux à la fois économique, patrimonial et écologique. Le saumon a été largement exploité dans le passé par la pêche commerciale dont l'activité s'est fortement réduite suite à la mis en place d'une réglementation internationale pour conserver l'espèce. Il reste néanmoins toujours très attractif pour la pêche à la ligne. Par ailleurs, l'espèce a fait l'objet d'un très fort développement de l'aquaculture avec un tonnage actuel de plus de 1 millions (Knockaert 2006). Ces élevages marins qui se sont multipliés à travers le monde ne sont pas sans conséquences sur la viabilité des stocks sauvages (introgression génétique, pathologies). Enfin, le saumon est considéré comme un symbole de la qualité de l'eau et un indicateur du changement global, compte tenu de son cycle biologique qui se déroule à la fois dans le milieu continental et le milieu marin.

Le Saumon atlantique a fait l'objet d'études depuis plus d'un siècle (Roule, 1920). Ces premières études et celles qui ont suivi, ont permis de connaître son cycle de vie et d'aborder la caractérisation des populations (Prévost 1987 ; Mills 1989 ; Shearer 1992 ; Gueguen et Prouzet 1994) et enfin d'analyser les traits et les stratégies d'histoire de vie de l'espèce (Klemetsen *et al.* 2003). A ces études biologiques et écologiques se sont ajoutées plus récemment des approches génétiques qui se sont multipliées en raison de la performance et de la précision des outils de génétique moléculaire utilisés permettant notamment d'intégrer le niveau d'organisation biologique qu'est l'individu (Paterson *et al.* 2004). Toutes ces études ont permis d'aborder le fonctionnement, la dynamique des populations de l'espèce et les mécanismes de régulation et d'évolution des stocks sous contraintes anthropiques (Prévost & Chaput 2001 ; Dumas & Prouzet 2003 ; Baglinière *et al.* 2005).

Ainsi, cette première partie présente la biologie du Saumon atlantique, l'état et le statut de ces populations en développant d'une manière plus particulière certains facteurs d'impact. Elle fait également le point sur les études génétiques réalisées sur l'espèce en abordant la notion de biologie de la conservation.

Partie 1.

CYCLE BIOLOGIQUE ET ÉVOLUTION ACTUELLE DES POPULATIONS

Le Saumon atlantique a été nommé et classé *Salmo salar* par Linné qui signifie en latin « le sauteur ». C'est un Téléostéen de la Famille des Salmonidés et de l'ordre des Salmoniformes.

1.1. Cycle biologique (Figure I-1)

Les populations de Saumon atlantique peuvent être anadrome ou non-anadrome. Cependant la forme anadrome est celle qui représente l'archétype du Saumon atlantique. Elle est également la forme qui produit les stocks les plus abondants. Les saumons Atlantiques sont des poissons migrateurs, amphibiotiques, potamotoques (voir le Glossaire) et phylopatriques, c'est-à-dire qu'ils naissent en rivière, y passent leur période juvénile, puis partent en mer pour croître avant de revenir dans leur rivière d'origine pour se reproduire (frayer). Contrairement au saumon du Pacifique qui est sémelpare (mort après le frai), le Saumon atlantique peut retourner en mer après le frai. La reproduction a lieu en rivière, sur zones caillouteuses dans lesquelles la femelle enfouie ses œufs. Les oeufs vont rester en incubation pendant deux ou trois mois suivant la température de l'eau (Carnac 1988). À l'éclosion, les larves (alevins) sont pourvues d'une vésicule vitelline qui assure leur survie jusqu'à ce qu'ils se nourrissent progressivement de larves d'insectes et de vers. La phase juvénile s'étend du stade alevin jusqu'au stade smolt en passant par le stade tacon.

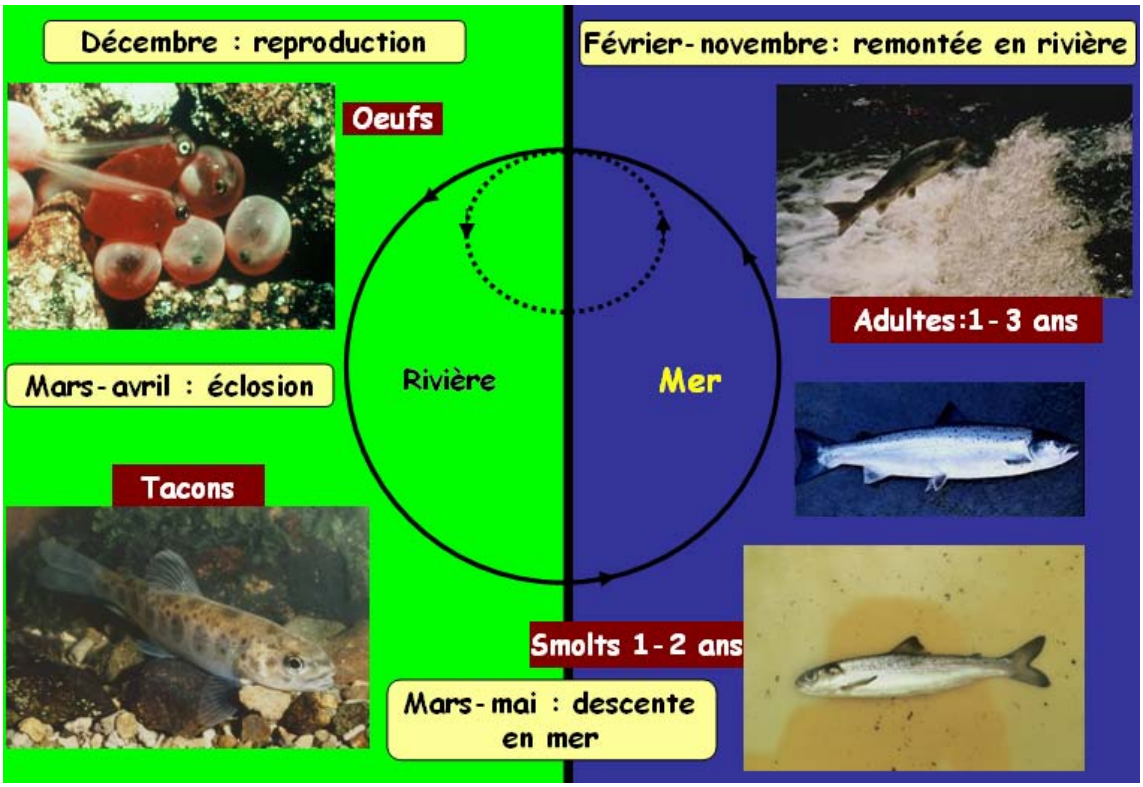


Figure I-1. *Cycle biologique du Saumon atlantique (Salmo salar) (Source modifiée Jean-Luc Baglinière, INRA).*

Le stade tacon dure jusqu'au début de la smoltification soit de 1 à 7 ans en Europe et de 1 à 8 ans en Amérique du Nord (Heland et Dumas 1994). En France, la durée est plutôt de 1 à 2 ans. Puis, le tacon parfaitement adapté à l'eau douce va subir un ensemble de modifications internes et externes et devenir un smolt : phase qui le préadapte à la vie dans le milieu marin (Bœuf 1994).

Figure I-2 Tacon de Saumon atlantique (Source INRA et ORE PFC).

Le stade smolt apparaît au printemps lors de la migration vers la mer La smoltification correspond à un ensemble de modifications cytologiques, morphologiques, physiologiques, biochimiques, hormonales et comportementales (Bœuf, 1994) déclenchées principalement par la température et la photopériode (Hoar 1976 ; Hoar 1988 ; Bœuf 1992) et permettant aux saumoneaux de s'adapter à la vie marine.

Figure I-3. Smolt de Saumon atlantique (Source Jean-Luc Baglinière, INRA).

Le Saumon atlantique est l'espèce qui effectue les plus longs déplacements en mer, plus de 14 000 km (Bœuf 1994). A ce jour quatre zones d'engraissement marine ont été identifiées : Groenland-Labrador (détroit de Davis et mer du Labrador), îles Féroé, mer de Norvège et Baltique (Figure I-4) Après un temps de séjour de 1 (castillons ou grisles) à 3 ans (poisson de plusieurs hivers de mer sur ces zones d'engraissement les adultes reviennent dans la rivière où ils sont nés. Cet instinct de retour ou « homing » est fortement influencé par des paramètres visuels et olfactifs (Hasler 1996 ; Dittman et Quinn 1996). Cependant, la mémoire

olfactive (odeurs environnementales et phéromones) apparaît être dominante dans la dernière partie de leur voyage (Hasler 1966 ; Stabell 1984 ; Hasler 1996 ; Dittman et Quinn 1996 par Dukes *et al.* 2004). En règle générale, il est considéré que cet instinct de retour à la rivière natale est très fort chez le Saumon atlantique puisque le pourcentage de divagants (strayers) varie entre 2 et 6% (Stabell 1984 ; Quinn 1993 ; Youngson *et al.* 1994 ; Altukhov *et al.* 2000 ; Jonsson *et al.* 2003).

A la différence avec le saumon du Pacifique, les mécanismes de leur survie en mer n'ont toujours pas été bien définis (Webb *et al.* 2007). Dans tous les cas, les études s'accordent à dire que les taux de survie en mer varient beaucoup plus dans le temps et dans l'espace que ceux en eau douce et que la phase post smolt est certainement celle la plus critique au cours du séjour marin (Shearer 2002 ; Friedland *et al.* 2003 ; Klemetsen *et al.* 2003).

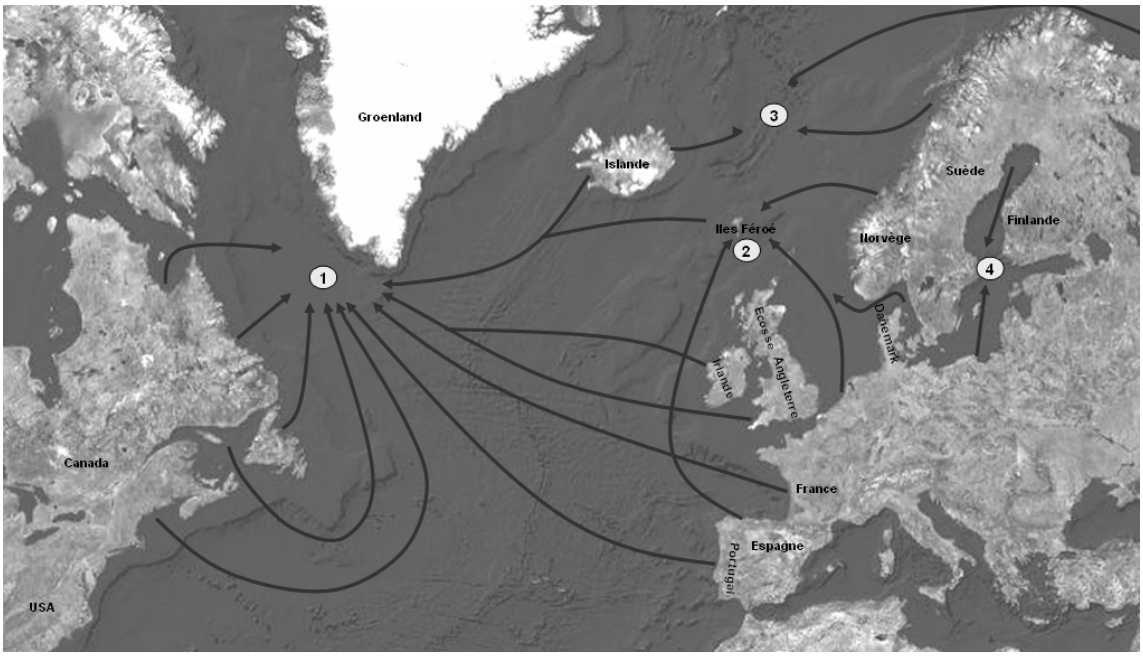


Figure I-4. *Migration du Saumon atlantique vers ses aires d'engraissements : l'ouest du Groenland (1), les îles Féroé (2) et la mer de Norvège (3) et la mer baltique (4).*

1.2. Évolution des populations et statut actuel de l'espèce

1.2.1. Aire de répartition originale et actuelle

Le Saumon atlantique est présent dans les pays dont les rivières se déversent dans l'Atlantique Nord et dans la mer Baltique. En Europe, son aire de répartition comprend des bassins versants de l'Atlantique, depuis l'Islande jusqu'au nord du Portugal. En Amérique du Nord, celle-ci s'étend du bassin versant de la rivière Hudson, dans l'État de New York, jusqu'à la baie d'Ungava, au Québec (Knockaert 2006). Il existe des formes non migratrices ou non anadromes de saumons atlantiques dans plusieurs zones glaciaires de l'Europe, de la Scandinavie et de l'Amérique du Nord. En moins d'un siècle, les saumons ont disparu en Allemagne, en Suisse, aux Pays-Bas, en Belgique, en république tchèque et en Slovaquie et sont sur le point de disparaître en Estonie, au Portugal, en Pologne, aux États-Unis et dans certaines parties du Canada (MacCrimmon & Gots, 1979 ; NASCO 2001 ; Jonsson and Jonsson 2004). Il a disparu des grands bassins européens tels que la Tamise, la Seine, la Garonne, le Rhin et l'Elbe, suite notamment à la construction de barrages (Prévost, 2001 par Knockaert 2006). De même, le saumon n'est plus présent que dans certaines parties du territoire français et plus particulièrement dans le grand ouest (Figure I-5).

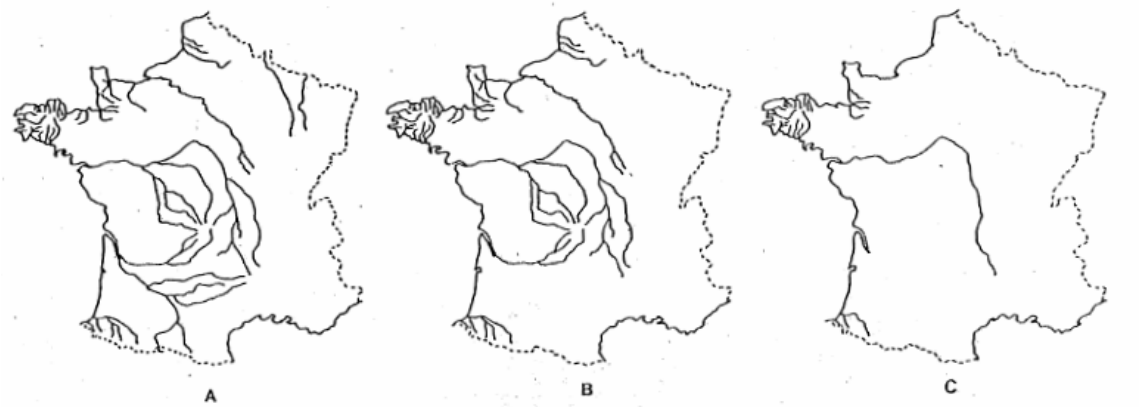


Figure I-5. *Évolution du nombre de rivières fréquentées par le Saumon atlantique depuis le milieu du XVIIIème siècle (A) jusqu'aux XIXème (B) et XXème siècles (C) (Source Thibault 1985).*

1.2.2. Évolution des stocks

Outre une diminution drastique de l'aire de répartition, l'abondance des stocks a fortement diminuée soit en raison d'une réduction des zones accessibles favorables et fonctionnelles sur les bassins, soit d'une baisse de productivité des stocks. Enfin, il faut

ajouter plus récemment une modification de l'histoire de l'espèce sur son aire de répartition actuelle. Cette modification se traduit par un renouvellement plus rapide des populations (Baglinière *et al.* 2004 ; Aprahamian *et al.* 2008) ce qui reflète généralement une adaptation d'une population animale à un milieu devenant de plus ou plus instable voir défavorable que ce soit en mer ou en rivière. Cet état de fait est la conséquence de tout un ensemble de contraintes anthropiques : érection de barrages dépourvus de dispositifs de franchissement efficaces, pollution et industrialisation, dégradation de l'habitat, exploitation et changement climatique. Chez le saumon la première cause de déclin de l'espèce reste la mise en place des barrages qui, en diminuant l'abondance et la productivité des stocks les a rendu très sensible à des conditions d'exploitation inadaptée. De fait, nous ne développerons pas les premières causes qui ont fait l'objet de nombreuses publications (Thibault, 1994 ; Mills, 1989) mais nous insisterons sur deux points d'impact plus récent l'exploitation et le changement climatique

1.2.2.1. L'exploitation marine sur les aires d'engraissement

Depuis un quart de siècle, l'identification des aires d'engraissement et des axes de migrations du saumon et l'utilisation d'engins de pêche de plus en plus performants l'ont rendu vulnérable face aux pêches professionnelles modernes. Des quotas annuels de captures (TAC) de saumons ont été mis en place progressivement sur les aires d'engraissement de l'ouest du Groenland et des îles Féroé par les pays producteurs de saumons, lors du 2ème Symposium international sur le Saumon atlantique en 1978. Ce n'est qu'au 3ème Symposium en 1986 que des résolutions ont été adoptées à l'unanimité pour la protection et la survie de l'espèce. Celles-ci consistent à réduire la pêche commerciale et développer la pêche sportive (Mills 1987).

Ainsi Le premier saumon pêché le long de la côte sud-ouest du Groenland date de 1956 ; c'était un smolt, marqué en Écosse. A partir de 1960, les captures augmentent de 50 tonnes pour atteindre un pic de 2 689 tonnes en 1971 (Davaine et Prouzet 1994). À partir de 1972, les prises pour l'ensemble de la pêcherie sont fixées à 1 100 tonnes, puis à 1 191 tonnes en 1974, pour finir à 850 tonnes en 1984 (Davaine et Prouzet 1994). C'est en 1976 que les navires groenlandais ont été les seuls autorisés à pêcher le saumon (Davaine et Prouzet 1994 ; Figure I-6). La majorité des saumons capturés sont de futurs saumons de printemps (2 ans et plus de mer) au Groenland et 75% sont des femelles (Davaine et Prouzet 1994). Grâce à la campagne internationale de marquage, les pays qui contribuent à alimenter les stocks

groenlandais ont pu être identifiés : Canada, Écosse, Etats-Unis, Islande, Norvège, Suède, Danemark, Finlande, France, Angleterre, Irlande et Espagne (Davaine et Prouzet 1994).

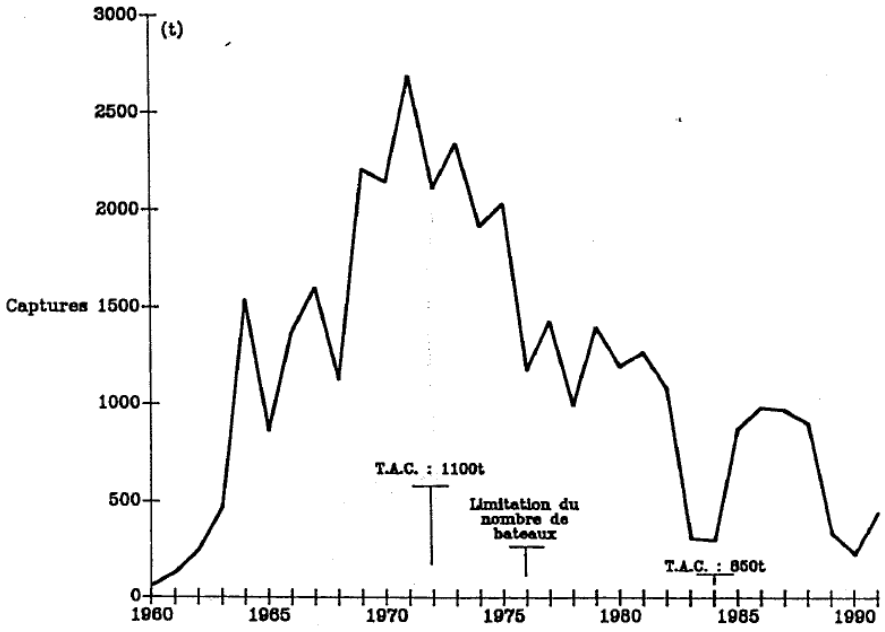


Figure I-6. *Évolution des captures de saumons du Groenland de 1960 à 1991. D'après Davaine et Prouzet (1994).*

Le premier saumon capturé dans les eaux féringiennes date de 1958. De 1967 à 1977 la production ne dépasse pas 40 tonnes. Elle atteint un maximum de 1 025 tonnes en 1981, avant la mise en place du TAC (Total Admissible de Capture), fixé à 627 tonnes en 1987, sous le contrôle de l'OSCAN (Organisation pour la Conservation du Saumon atlantique Nord) (Figure I-7). Avant 1980, les captures étaient constituées majoritairement (62-91%) de saumon ayant passé un hiver en mer et, après 1980, deux hivers en mer (75-95%) (Davaine et Prouzet 1994).

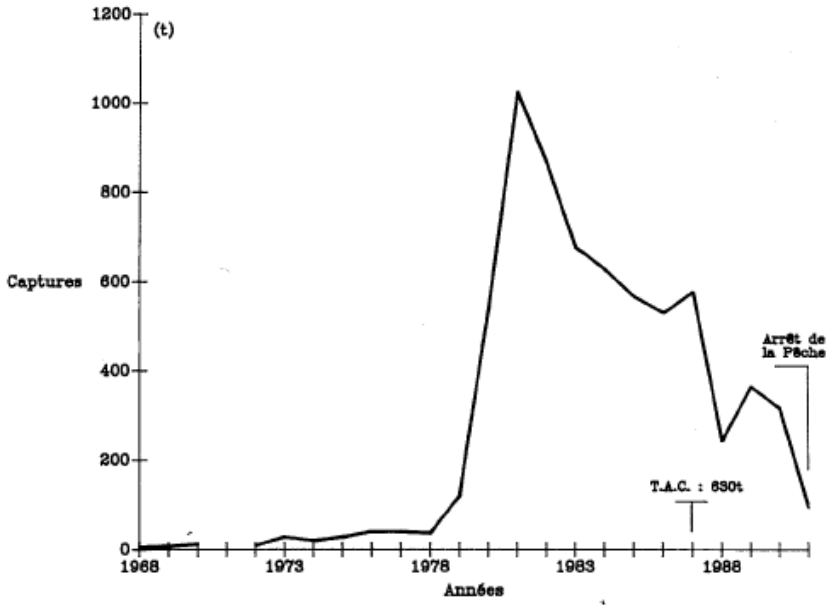


Figure I-7. *Évolution des captures de saumons aux îles Féroé de 1968 à 1991. D'après Davaine et Prouzet (1994).*

1.2.2.2. Changements climatiques

Globalement, la température a augmenté ce dernier siècle mais cette augmentation n'a pas été uniforme dans les deux hémisphères. Dans l'hémisphère Sud, l'augmentation est générale, alors que dans l'hémisphère Nord, certaines aires se sont refroidies (Nord-est de l'Atlantique) et d'autres se sont réchauffées. La température au Nord de l'Atlantique a varié durant ce siècle dernier mais des relevés effectués, depuis 1893, montrent plusieurs périodes de faibles températures et de salinités, entraînant des eaux pauvres en plancton et en larves de poissons :

En Atlantique Nord-est :

- 1906-1912
- 1974-1978

Au Groenland :

- 1919-1924
- 1982-1984

La diminution des températures des eaux de surface au niveau des aires d'engraissement du Saumon atlantique pourrait être attribuée aux variations des pressions atmosphériques

Nord Atlantique NAO (North Atlantic Oscillation) et aux variations de la dérive nord atlantique (Gulf Stream). La NAO est le mode atmosphérique dominant de la variabilité climatique et représente les variations de pression atmosphérique entre l'anticyclone des Açores et les basses pressions subpolaires (Islande). L'indice de la NAO noté NAOI (North Atlantic Oscillation Index), mesurant ces différences de pressions entre les deux centres d'actions (l'Islande et les Açores), est positif lorsque les deux centres se renforcent et est négatif lorsque les deux centres s'affaiblissent simultanément. Plus cet indice est élevé et plus les vents d'ouest sont dominants en Atlantique Nord, ce qui apporte un air humide plus chaud au-dessus du continent européen avec des précipitations accrues, provoque des hivers maritimes plus doux, et repousse l'air froid venu de Sibérie. Ceci entraîne des températures plus froides dans les deux zones principales d'engraissement du saumon (ouest du Groenland et îles Féroé).

Depuis 1974, après une longue période négative, la NAOI est majoritairement positive ce qui entraîne des eaux froides dans les zones d'engraissement. La corrélation entre les variations du NAOI et la variation des stocks des saumons atlantique écossais, montrée par l'étude de Dickson et Turrell (1999) (Figure I-8), suggère que le refroidissement des aires d'engraissement par la NAO est en dessous du rang optimal du saumon fixé entre 4 et 10°C. Cette situation empêcherait les saumons de croître car leur ressource en phytoplancton est affaiblie, voire elle causerait leur mort. Les travaux d'Irigoien *et al.* (2000) ont mis en évidence la corrélation entre la composition et la présence du phytoplancton avec les variations du NAO et il apparaît une diminution du phytoplancton avec les NAOI positifs. Depuis 1950 le phytoplancton et le zooplancton ne cessent de diminuer en Atlantique Nord avec un minimum relevé en 1980. Nous pouvons espérer une amélioration de la situation puisque depuis 1995 on assiste à un changement graduel du NAO vers le négatif. Cependant à la différence d'El Nino qui revient périodiquement tous les 3 à 7 ans, les oscillations de la NAO ne sont pas régulières et aucune prédiction ne peut être faite sur la durée de ses modes.

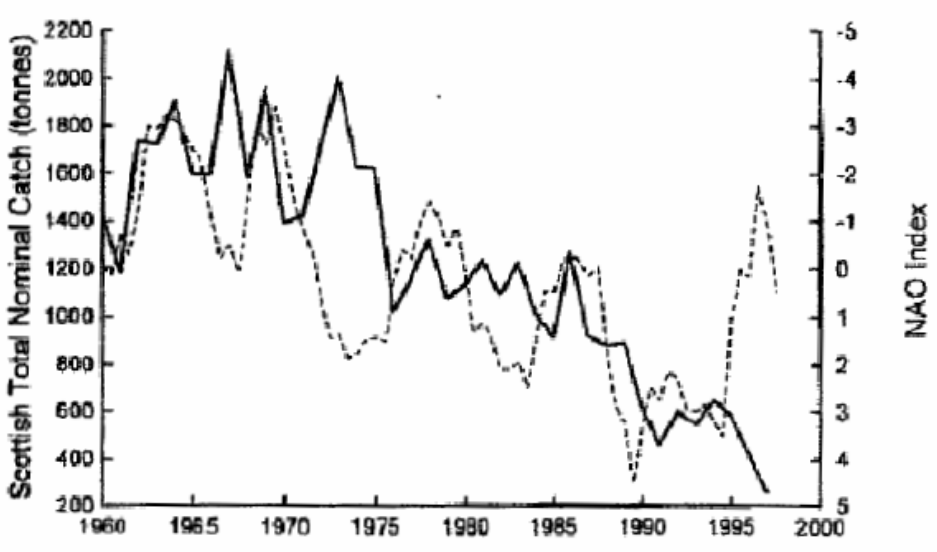


Figure I-8. *Évolution des stocks de saumon écossais (trait solide) depuis 1960 comparés aux indices moyen sur deux ans de la NAO (trait en pointillé). D'après Dickson et Turrell 1999.*

Ainsi, cette modification du milieu marin pourrait être à l'origine de la baisse de la survie du saumon lors de son écophase marine en impactant directement l'individu suite à la baisse des ressources trophiques. Mais ce changement des conditions marines peut impacter également directement la survie du saumon en augmentant son taux de prédation par certaines populations de mammifères marins qui sont en augmentation (Middlemas *et al.* 2003).

1.2.3. Statut des populations étudiées

Dans le cadre de ce travail nous nous sommes intéressés à des populations françaises et écossaises protégées par la Convention de Berne (Annexe III); Arrêté du 8 décembre 1988 ; Directive « Habitats-Faune-Flore »), « populations ateliers » (suivies et étudiés) depuis au moins 20 ans. Elles évoluent dans des habitats distincts et sont de tailles globales très différentes, les plus petites étant en France et les plus grandes en Écosse. Leurs statuts au sein de l'Union Internationale de Conservation ne sont pas réellement définis car il semble difficile de prédire l'évolution future de ces populations. Malgré une tendance actuelle à la diminution, les populations de Saumon atlantique françaises et écossaises semblent ne pas avoir subi le même impact. Selon Parrish et ses collaborateurs (1998) les populations d'Europe du Sud, telles que les populations françaises, doivent être classées parmi les plus vulnérables et en déclin alors que celles d'Europe du Nord, l'Écosse, notamment, sont plutôt stables. Un rapport du WWF fait apparaître les chiffres suivants : les populations vulnérables ne sont que de 6% ; 30% sont éteintes ; 32% sont dans un état critique et 21% sont en danger

d'extinction. Le statut de l'espèce semble plus menacé d'extinction que vulnérable. La situation des populations Françaises est plus alarmante que les populations Ecossaises avec 63% de leur population en bonne santé contre 0% en France (Figure I-9). L'Écosse a, malgré tout, développé des structures « Fishery Board » (en vertu de la législation à la pêche au saumon de 1860, comme indiqué dans la loi sur le saumon « Salmon Act » de 1986). Ces structures sont responsables de la gestion, de la protection, de l'amélioration et de la conservation du Saumon atlantique dans ses rivières.

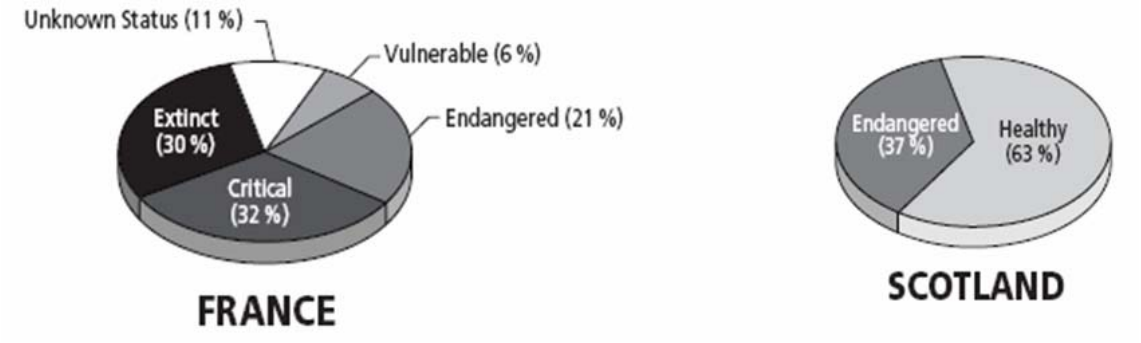


Figure I-9. *Pourcentages de rivières en fonction de leurs statuts (inconnu, vulnérable, en danger, critique ou éteint) (Rapport du WWF, « The Status of Wild Atlantic Salmon: A River by River Assessment », 2001).*

Partie 2.

CARACTÉRISATION GÉNÉTIQUE

Cette partie permet d'apprécier l'apport de la génétique pour comprendre les mécanismes et les régulations des populations de Saumon atlantique et de l'espèce.

2.1. L'ADN chez le Saumon atlantique

L'ADN (Acide Désoxyribonucléique) est considéré actuellement comme la molécule qui porte l'information génétique responsable du développement, de la survie et de la reproduction d'un organisme. La quantité totale d'ADN par cellule haploïde (gamète) chez le Saumon atlantique est de 5.7 picogrammes (pg), l'équivalent de 6000 Mb (Millions de paires de bases), soit deux fois le contenu du génome humain. Les populations issues de l'Atlantique Est présentent 29 chromosomes avec 58 paires et 74 bras chromosomiques alors que ceux de l'Atlantique Ouest ont 27 chromosomes avec 54 paires et 72 bras chromosomique (Hartley et Horne 1984 ; Hartley 1987 ; voir Figure I-10). Il existe donc une réelle différence entre le Saumon atlantique d'Europe et celui d'Amérique du Nord. Comme nous le verrons ci-dessous les évidences les plus marquantes sont basées sur des divergences phylogénétiques provenant d'études sur la variation moléculaire.

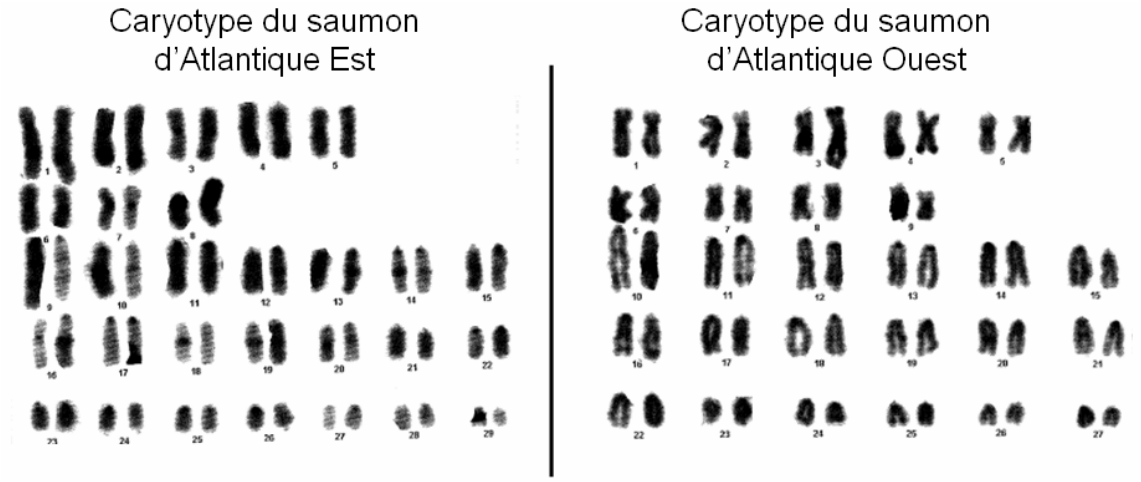


Figure I-10. *Photo du caryotype du Saumon atlantique sur les deux côtés de son aire de répartition (Est et Ouest) (Source modifié Hartley 1987).*

2.2. Concepts de base

La variabilité génétique d'une espèce se répartit à l'intérieur et entre les populations. Celle-ci se mesure généralement par l'hétérozygotie et la richesse allélique. L'importance de maintenir la variabilité génétique a été démontrée sur un grand nombre d'espèces différentes en mettant en évidence la corrélation entre la fitness et la variabilité génétique (Reed et

Frankham 2003). La perte de diversité génétique aurait des conséquences sur la survie et un niveau de diversité trop faible pourrait conduire la population à perdre sa capacité d'adaptation lors d'un changement environnemental brutal dû à la perte d'allèles adaptatifs et l'accumulation d'allèles délétères. Une population dont la variabilité génétique diminue est généralement une population dont la taille diminue. Après un goulot d'étranglement il y a une augmentation temporaire du taux d'hétérozygotie causé par la perte des allèles rares. Comme nous le verrons dans le chapitre 2 cette augmentation d'hétérozygotie permet de détecter la réduction de la taille efficace d'une population mais il existe également beaucoup d'autres modèles qui utilisent aussi les variations génétiques pour accéder à différents paramètres évolutifs.

Les variations génétiques sont créées par des mutations ou des recombinaisons mais leur distribution est définie par des forces évolutives telles que la migration, la sélection et la dérive génétique. L'identification des facteurs intervenant dans cette variabilité permet de comprendre la structuration et l'évolution des populations voire de l'espèce. Les études qui ont testé l'importance de la sélection dans l'évolution des locus, indiquent toutes que les forces principales agissant sur la différenciation entre populations de Saumon atlantique sont la dérive génétique et les flux géniques et pas la sélection (Verspoor 1994 ; Jordan *et al.* 1997).

Parce que les nombres de reproducteurs et d'immigrants dans la population déterminent la variation génétique, il est usuellement admis que l'estimation de ces paramètres est primordiale pour statuer sur l'état des populations. La tendance du saumon sauvage anadrome atlantique à exercer un fort homing laisse à penser que seule une estimation globale du nombre de reproducteurs nous renseignera sur la population. La taille efficace est l'un des indicateurs les plus importants en conservation. Il peut être défini comme la taille d'une population idéale c'est-à-dire une population qui possède un sexe ratio égal à 1 et une variance du nombre de descendants qui correspond à une distribution de Poisson (voir Frankham *et al.* 2002). Ce paramètre est une mesure du taux de perte de la variabilité génétique en termes de nombre d'individus. Il reflète ainsi la force de la dérive génétique et le risque de consanguinité, et renseigne sur le potentiel évolutif.

2.3. Phylogénie et structure des populations

Le Saumon atlantique est une lignée évolutive bien distincte des autres salmonidés. Cependant, des études de la variation génétique ont également montré qu'elle présentait une diversité phylogénétique intra-spécifique significative. Ce n'est pas surprenant puisque les poissons néarctiques et paléarctiques, tel que le Saumon atlantique, possèdent une histoire faite de contractions et d'expansions pendant la période quaternaire, liées à l'avancée et au retrait des glaciers (Hewitt 1996, 1999). Lorsque la température diminue et que les glaces avancent les populations nordiques disparaissent. A contrario lorsque les glaces se retirent, des populations émergent par la migration d'individus des refuges situés plus au sud (Taberlet *et al.* 1998).

Établir les relations entre les espèces de la famille des Salmonidés reste problématique et certaines relations n'ont toujours pas été résolues (Crespi et Fulton 2004). Ces difficultés ont des raisons multiples mais la principale pourrait être le fait que les groupes proviennent de radiations adaptatives rapides par tétraploïdisation de leur ancêtre commun (50-100 millions d'année) (Allendorf et Thorgaard 1984). Les autres pourraient être les hybridations et les introgressions interspécifiques (Utter et Allendorf 1994) et la rapide évolution des populations dans leur habitat à la suite du dernier retrait des glaces Pléistocènes.

Des études sur la structure actuelle du Saumon atlantique ont permis de mettre en évidence, par calculs des distances génétiques entre populations (polymorphisme de gènes codants - ARN ribosomaux et ADN mitochondriaux- et non-codants - microsatellites), trois sous-ensembles : Ouest-Atlantique, Est-Atlantique et Balte (Guyomard 1994). Les différenciations génétiques des populations américaines et européennes résulteraient de l'isolation de l'espèce dans deux zones refuges lors de la dernière glaciation, vraisemblablement dans la partie sud de l'aire de répartition de l'espèce (Guyomard 1994). Cette différenciation pourrait se poursuivre à l'heure actuelle par l'absence de flux génétique entre les deux sous-ensembles (Guyomard 1994). La divergence entre les populations atlantiques et baltes serait plus tardive ; elle se serait produite lors de la fermeture du détroit assurant la communication entre la mer du Nord et la mer Baltique (Guyomard 1994). Chez le Saumon atlantique la variabilité dans les populations est beaucoup plus élevée qu'entre les populations. Elle tombe à 10% entre les populations atlantiques. Certains chercheurs (Guyomard (1994)) affirment que les populations françaises éteintes se distingueraient très peu des populations espagnoles, irlandaises et norvégiennes. Les résultats actuels vont dans ce

sens et observent de faibles différenciations entre les populations d'un même continent (*Fst=0.176* (allozymes) en Europe; *Fst=0.076* (allozymes) en Amérique du Nord) contre une différenciation quatre fois plus élevée entre les populations des deux continents (*Fst=0.33* (allozymes) *=0.27* (14 microsatellites) Europe et Amérique du Nord) (Verspoor *et al.* 2005 ; King *et al.* 2001). Beaucoup d'exemples ont montré la divergence des saumons de l'Est et de l'Ouest de l'Atlantique avec des marqueurs microsatellites (McConnell *et al.* 1995 ; King *et al.* 2001 ; Koljonen *et al.* 2002 ; Gilbey *et al.* 2005) et il est fort probable que leur ancêtre commun soit situé dans le Sud-est de l'Océan Atlantique (endémique à l'Europe) comme les autres membres du même genre *Salmo* (King *et al.* 2007). La forte différenciation entre les populations européennes et américaines proviendrait du fait qu'elles aient été séparées lors de la dernière glaciation au Pléistocène (115 000 – 10 000 ans) et qu'elles n'aient pas pu échanger de gènes pendant cette période (King *et al.* 2001 ; Nilson *et al.* 2001 ; King *et al.* 2007). Néanmoins à ce jour la date exacte de leur séparation varie énormément en fonction des marqueurs utilisés. Le scénario le plus probable suppose une séparation européenne et américaine il y a environ 100 000 ans vers différents refuges glaciaires puis les populations actuelles se seraient formées indépendamment sur les deux côtés de l'Atlantique à partir de ces refuges il y a environ 10 000 ans (Figure I-12). La Figure I-13 présente un résumé de la chronologie du Saumon atlantique avec ses incertitudes et ses inconnues. Les comparaisons génétiques entre les populations d'Amérique du Nord et d'Europe ont révélé que les saumons d'Amérique du Nord présentaient des allèles uniques qui permettaient de distinguer les saumons des deux côtés de l'Atlantique (King *et al.* 2001). Cependant il semblerait également que les populations d'Amérique du Nord ont une diversité génétique moins élevées et une structuration moins forte que les populations Européenne (King *et al.* 2001). Mais nous ne connaissons toujours pas la signification biologique des différenciations. Est-ce juste le reflet de la dérive génétique et de la dispersion limitée des individus, sans présenter aucune divergence des traits biologiques, ou est-ce que cette divergence reflète les adaptations locales de chaque population ?

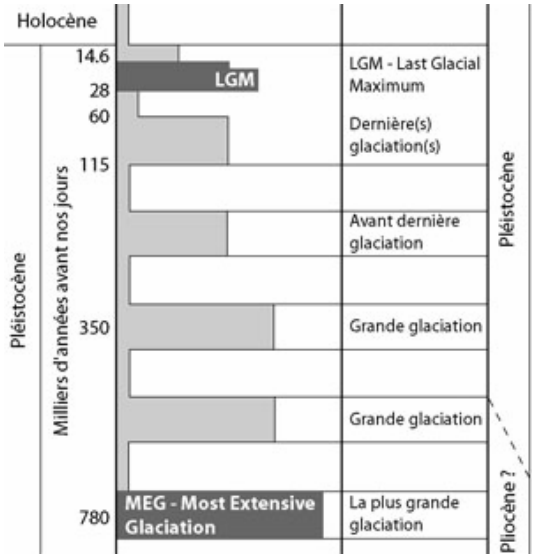


Figure I-11. *Chronologie des grandes glaciations quaternaires (Source Schlüchter et Kelly 2000 modifiée).*

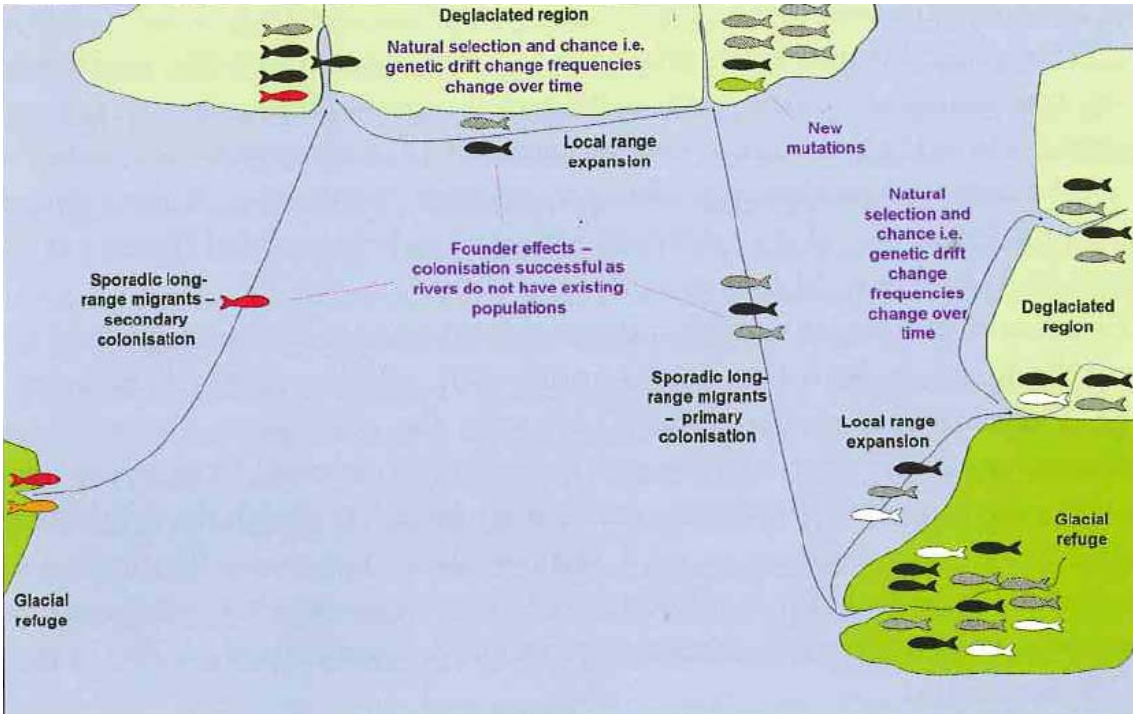


Figure I-12. *Scénario probable de l'origine des Saumons atlantique depuis la dernière glaciation sur la droite (Source King et al. 2007).*

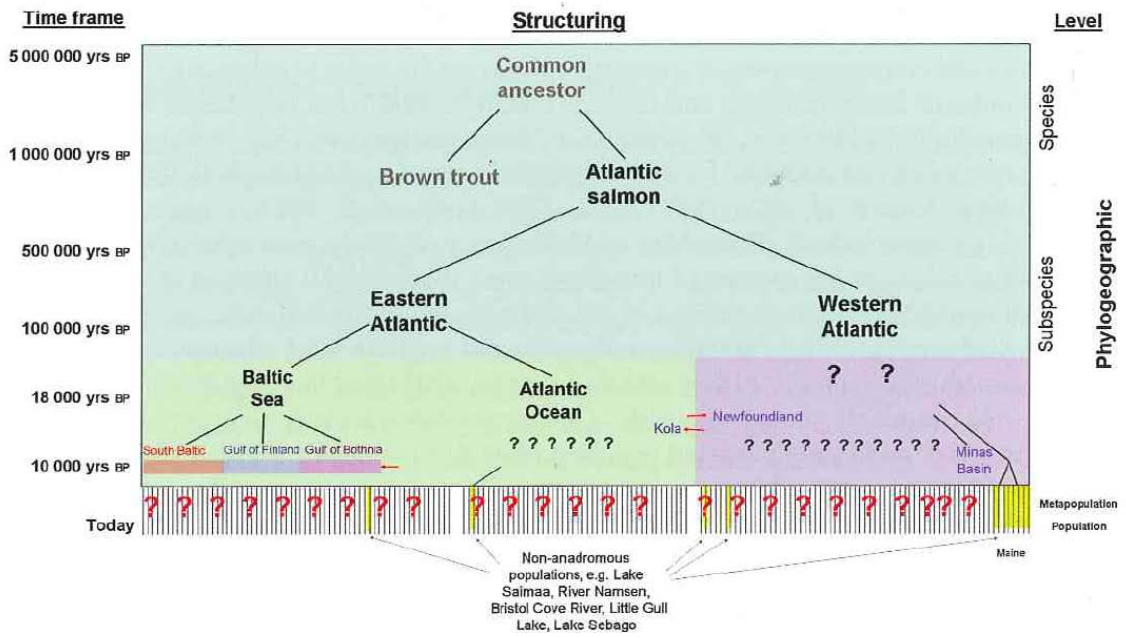


Figure I-13. *Résumé de l'origine du Saumon atlantique en utilisant les marqueurs génétiques, « ? » les structures qui n'ont toujours pas été résolues (Source King et al. 2007).*

2.4. Polyploïdie

A l'heure actuelle, on considère que les salmonidés ont subi au cours de leur évolution un épisode de tétraploïdisation (i.e. duplication génomique) avec retour progressif à un état diploïde. En conséquence, un même marqueur peut, aujourd'hui apparaître à l'état simple ou dupliqué, selon les amorces utilisées.

Les observations d'une quantité d'ADN deux fois supérieure chez les salmonidés que celui de la proche famille des clupéidés (Ohno *et al.* 1968), ont amené les travaux de Ohno et ses collaborateurs (Ohno *et al.* 1968 ; Ohno *et al.* 1969 ; Ohno 1970a ; Ohno, 1070b) à attester l'évènement de tétraploïdisation suggéré par Svärdson (1945). Des travaux se basant sur un évènement de tétraploïdisation unique, ont estimé l'émergence de l'ancêtre tétraploïde à 25-100 millions d'années (Wright *et al.* 1983 ; Allendorf et Thorgaard 1984). À la suite de cet évènement, le génome des salmonidés aurait essentiellement évolué par translocations robertsoniennes (i.e., fusion de deux chromosomes acrocentriques par leur centromère pour former un chromosome métacentrique) (Ohno *et al.* 1969 ; Wright *et al.* 1983 ; Allendorf et Thorgaard 1984 ; Hartley 1987 ; Johnson *et al.* 1987) comme en témoignent les stigmates (marques). Les trois mécanismes susceptibles de donner des cellules ou des individus polyploïdes sont l'endomitose, la production de gamètes non réduits et la polyspermie (Otto et

Whitton 2000, par Gharbi 2001). L'endomitose correspond à la duplication des chromosomes sans division du noyau, résultant en un doublement (ou plus) du nombre de chromosomes à l'intérieur d'une cellule. Le mécanisme de production de gamètes non réduits est une sorte d'endomitose avant la rentrée en méiose des cellules gamétiques (endomitose pré-méiotique) (Gharbi 2001). Le gamète contient le double de chromosomes (diploïde) par rapport à un gamète normal (haploïde). La combinaison d'un gamète diploïde avec un gamète normal produira un œuf triploïde et avec un gamète diploïde un œuf tétraploïde. Un œuf tétraploïde peut également être produit à partir de deux gamètes haploïdes. Pour cela, il suffit que le zygote subisse au début de son développement une endomitose (Gharbi 2001). Pour finir, citons le mécanisme de polyspermie qui correspond à la fécondation du gamète femelle par plusieurs gamètes mâles, ce qui génère un zygote triploïde.

Lorsqu'un individu possède la combinaison de deux génomes de la même espèce, il est nommé *autopolyploïde* et lorsque il possède une combinaison deux génomes d'au moins deux espèces différentes, il est nommé *allopolyploïde*. La fréquence significative des évènements de polyploïdisation dans une population semble être provoquée par des paramètres tels que la température et l'hybridation (Ramsey et Schemske 1998). Des travaux récents sur les relations phylogénétiques attestent du caractère récurrent de ces évènements (Gharbi 2001). Contrairement à la vision classique d'une origine unique pour chaque espèce, d'autres auteurs penchent sur une origine multiple avec différentes populations ancestrales qui auraient contribuées à au patrimoine génétique des espèces actuelles (Soltis et Soltis 1999, par Gharbi 2001).

2.5. La diversité génétique

L'objectif visant à conserver une grande diversité génétique au sein des populations revient à préserver le potentiel d'adaptation de l'espèce pour faire face à de nouvelles conditions dans son milieu. La dégradation de l'habitat (conduisant notamment à la disparition des zones de reproduction) et la pêche sont susceptibles de réduire la diversité des populations de saumon. Plusieurs facteurs contribuent à l'émergence et aux variations de la diversité d'une espèce : mutation, recombinaison génique, migration, sélection et dérive génétique. La mutation, la recombinaison et la dérive génétique sont des facteurs qui augmentent la différenciation entre les populations tandis que la migration tend à en maintenir l'homogénéité. La sélection, quand à elle, peut intervenir dans les deux sens (Guyomard 1994). La diversité génétique dans et entre les populations est très sensible aux remaniements

produits par l'homme, tel que le repeuplement, et ces remaniements doivent être effectués avec une grande vigilance. Il semble préférable, quand la situation le permet, d'utiliser des souches sauvages lors de l'introduction d'individus pour prendre en considération les adaptations locales de la population. Cette stratégie a été mise en œuvre en Écosse et en Suède et il semblerait qu'elle soit mieux adaptée lorsque les populations peuvent supporter des prélèvements de géniteurs (Guyomard 1994). Comme nous le verrons par la suite, le maintien de la variabilité génétique est lié à la taille de la population. Cependant il existe de petites populations de Saumon atlantique qui ont conservé un taux de variabilité génétique fort (Nielsen 1999 ; Säisä *et al.* 2003 ; Consuegra *et al.* 2005). La nature de la tétraploïdie et les générations chevauchantes seraient la cause du maintien de la variation génétique au-dessus d'un certain degré (Allendorf et Thorgaard 1984 ; Waples 1991) mais cela n'a jamais été démontré.

2.6. Taille efficace

Beaucoup de populations de salmonidés sauvages ont considérablement diminué ces dernières décennies (Nielsen *et al.* 1997 ; Consuegra *et al.* 2002 ; Koljonen *et al.* 2002) et les réductions de taille, sur les bases théoriques en génétique des populations, sont supposées avoir un impact direct sur la variabilité génétique : plus une population est petite, moins sa variabilité génétique sera élevée et plus elle aura des risques de s'éteindre (Taylor *et al.* 1994). Pour mieux comprendre comment une variation génétique est perdue dans une population, il faut intégrer le concept de l'efficacité génétique d'une taille de population appelée aussi « taille efficace » (*Ne*). Le concept de taille efficace englobe la mesure de capacité d'une population à passer sa variabilité génétique d'une génération à l'autre (Consuegra et Nielsen, 2007). Lage et Kornfield (2006) donnent un exemple de réduction parallèle entre la diversité génétique et la taille efficace d'une population de Saumon atlantique en danger (rivière Maine) par la réduction du nombre de ses reproducteurs sur le long terme.

Pour conserver une petite population, l'une des problématiques est de déterminer le nombre minimum d'individus nécessaires pour maintenir le taux actuel de diversité génétique. Des estimations qui dérivent de l'équilibre entre la force de dérive génétique et de mutation ont vu le jour, dans les années 1980, à la suite des travaux de Franklin. Dans la littérature qui traite de génétique, nous voyons apparaître régulièrement l'idée de conservation sur la règle 50/500, avec 50 le nombre minimal d'individu efficace pour minimiser la dépression consanguine et 500 pour maintenir le potentiel évolutif (Franklin 1980, Lande 1988, Franklin

et Frankham 1998, Lynch et Lande 1998). De nombreux généticiens s'accordent à penser que plusieurs centaines d'individus à chaque génération sont nécessaires pour maintenir la variabilité sur le long terme (Waples et al. 1990). Selon l'équation de Frankel et Soulé (1981), la diminution de la variation génétique en fonction de la dérive est estimé par $-t/2Ne$, t étant le nombre de génération, ce qui prédit une taille efficace de 50 individus insuffisante pour une conservation à long terme puisque l'hétérozygotie diminue après quelques générations (Figure I-14, Meffe et Carroll 1997). Des expériences sur les mouches ont observé que 54 générations étaient nécessaires pour qu'une population de 50 individus efficaces disparaisse (Reed et Bryant 2000). Ces expériences n'ont jamais été réalisées sur les salmonidés et il est difficile de se prononcer sur le temps et la taille minimale pour que la population ne disparaisse. Néanmoins, quelque soit l'espèce l'idée de préconiser une taille minimale de 500 pour maintenir la variabilité génétique au cours des générations semblerait plus sûre (Lande 1995). Pour le Saumon atlantique avec un temps de génération de 3-5 ans, cette taille correspond à un nombre de reproducteurs compris entre 100 et 167 par an (Consuegra et Nielsen 2007). Ce nombre est identique à celui estimé par Waples (1990) pour réduire la perte d'allèles rares chez le saumon du Pacifique.

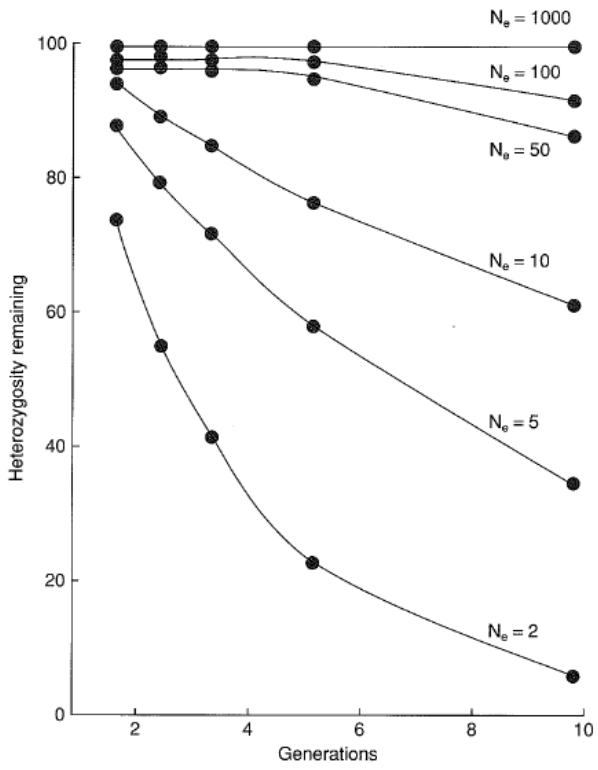


Figure I-14. *Perte de la variabilité génétique simulée après 10 générations en fonction de tailles efficaces différentes. Les populations à large taille efficace maintiennent leur variabilité génétique alors que les petites la perdent. D'après Meffe et Carroll 1997.*

La taille efficace est généralement plus petite que la taille réelle (*N*). Les estimations du ratio *Ne/N* sont rares chez les populations de Saumon atlantique sauvage à cause de la difficulté de leurs estimations dans ces populations naturelles. Cependant quelques estimations sont disponibles et elles sont supérieures à 0,8 (Koljonen *et al.* 2002) chez les populations de Saumon atlantique d'élevage. Si on applique le même ratio pour le Saumon atlantique sauvage, sous la condition de taille minimale, en transformant le nombre de reproducteur par génération (100-167) nous obtenons une taille annuelle réelle entre 1000-2000 individus annuelle nécessaire pour conserver la population (Consuegra et Nielsen 2007).

Comme nous le verrons dans les résultats exposés sur les estimateurs de taille efficace dans ces travaux de thèse (Article 2, Chapitre 3), l'estimation de la taille efficace peut être influencée par différents facteurs. Dans la littérature (Consuegra et Nielsen 2007), il est mentionné que le facteur probable le plus important, sous-estimant les valeurs de *Ne* chez le Saumon atlantique, est la fluctuation de sa taille au cours du temps. Nous savons que la taille des populations varie d'une génération à l'autre en réponse à l'instabilité environnementale. Mais d'autres facteurs peuvent également intervenir tels que la variance du nombre de descendants par individu reproducteur (l'idéal correspond à deux parents donnant deux enfants ; plus la variance augmente et plus *Ne* diminue) et le sexe ratio (l'idéal étant un mâle pour une femelle et tout déséquilibre diminue *Ne*). À ceci, viennent s'ajouter la migration et les modes de reproduction.

2.7. Conclusions

Les scientifiques s'accordent à penser que la variation génétique est nécessaire non seulement pour assurer l'adaptabilité actuelle des espèces mais aussi pour l'évolution continue de l'espèce. Pour conserver cette variabilité, cinq points peuvent être déclinés :

- Maintenir la variabilité génétique dans les petites populations en :

- maintenant une taille efficace minimale de 500 individus.

- évitant la sélection artificielle.

- évitant la migration artificielle.

- Maintenir une large taille réelle en :

- assurant des habitats de reproduction dans les rivières.

- fournissant des passages pour la migration des adultes et des smolts.
- régulant la pêche commerciale et sportive (nombre, sélection spécifique) et en affinant les mesures des taux de captures annuels (TAC) qui semblent, bien souvent, insuffisants pour prévenir le crash d'un stock.

- Eviter l'aquaculture dans les rivières où les stocks des saumons sauvages sont en dessous des seuils minima conseillés (minimum de la taille réelle 2000 individus).
- Conseiller une aquaculture moins sélective, moins polluante et plus sécurisée (milieu de contrôle plus clos).
- Faire de la cryoconservation.

Ces démarches pourraient intervenir après avoir statué sur l'état en danger d'une population. Ce qui nécessite un travail préliminaire sur les estimations d'abondance des stocks et de leurs états de santé. Néanmoins comme je l'ai déjà mentionné dans l'introduction, les outils actuels s'avèrent peu robuste et ne renseignent pas sur la viabilité à long terme des stocks de poissons. Au cours de ce manuscrit, je fais part de l'utilité des outils génétiques tels que ceux estimant la taille efficace, car ils permettent tout d'abord d'avoir une idée de la taille globale puis ensuite de la viabilité des populations. Cependant je suis tout à fait consciente que nous ne pouvons pas négliger les autres paramètres évolutifs, tels que la migration et la sélection, qui interviennent également dans la survie et la pérennité de l'espèce. J'ai ainsi essayé de travailler sur la sensibilité des estimateurs de la taille efficace face à la migration. Le modèle proposé au cours de ces travaux pourrait être modifié en vue d'introduire cette autre force évolutive. Concernant la sélection, elle n'est pas abordée ici puisque le travail s'est effectué sur des séquences neutres. Mais les études en sélection ne doivent pas être négligées et devraient prendre également une place dans les programmes de conservation.

II

OUTILS, MÉTHODES ET PROCÉDÉS

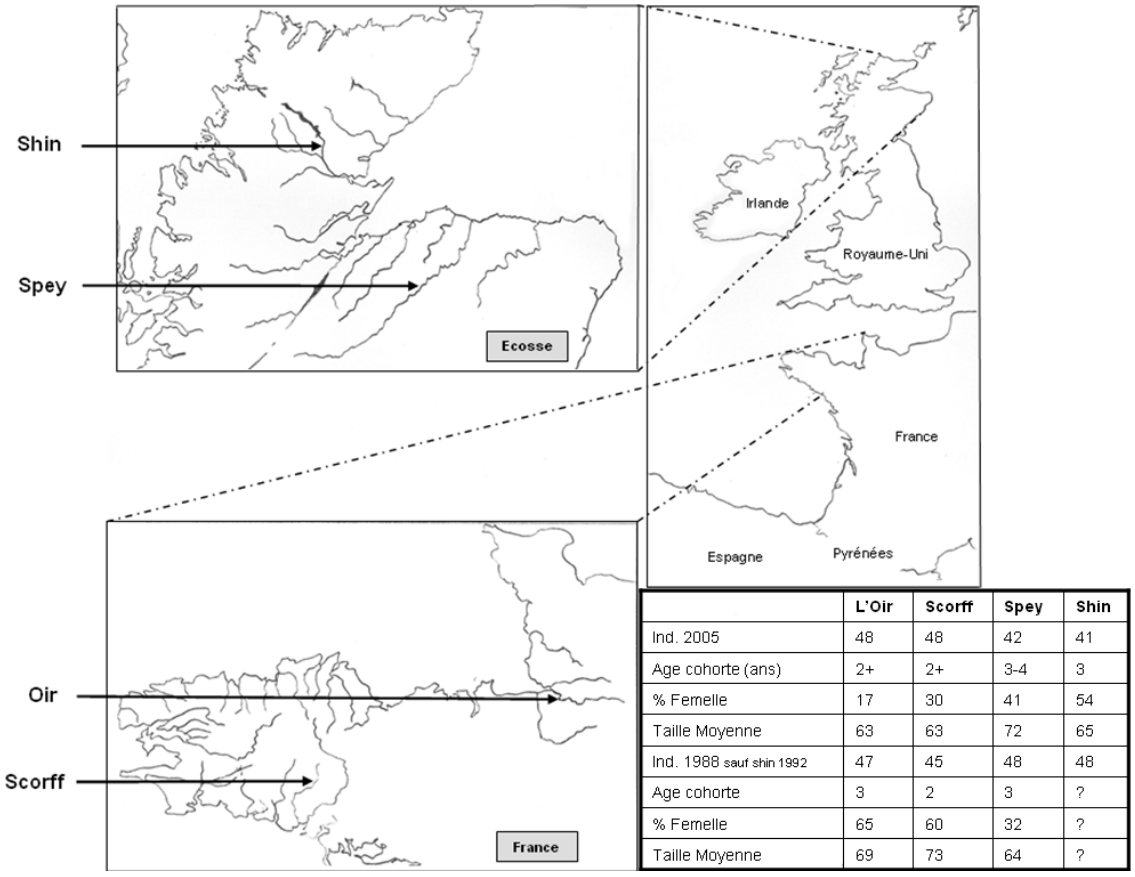
Partie 1.

LES INDIVIDUS ET ÉCHANTILLONS

1.1.Caractéristiques des individus et des populations

Un total de 367 Saumons atlantiques sauvages et anadromes provenant de 4 aires géographiques différentes (Oir et Scorff en France ; Spey et Shin en Écosse) ont été génotypés.

Pour chacune des rivières, 96 individus ont été échantillonnés dans deux périodes différentes : environ 48 individus, en 2005, et 48 individus dans le passé (1988 pour Oir, Scorff et Spey ; 1992 pour Shin) (Figue II-1). Ces individus sont tous des adultes de la même cohorte qui ont été capturés lors de leur remontée en rivière. Avant de les relâcher nous avons effectué des prélèvements d'écailles et quand cela était possible de nageoires adipeuses.



	L'Oir	Scorff	Spey	Shin
Ind. 2005	48	48	42	41
Age cohorte (ans)	2+	2+	3-4	3
% Femelle	17	30	41	54
Taille Moyenne	63	63	72	65
Ind. 1988 sauf shin 1992	47	45	48	48
Age cohorte	3	2	3	?
% Femelle	65	60	32	?
Taille Moyenne	69	73	64	?

Figure II-1. *Localisations géographiques des 4 populations de Saumons atlantique étudiées en France (Oir et Scorff) et en Écosse (Spey et Shin) avec le tableau des caractéristiques des échantillons faits en 2005 et en 1988, pour Oir, Scorff et Spey, et 1992 pour Shin : Nombre d'individus par échantillon (Ind.), âge de la cohorte, pourcentage de femelles et taille moyenne en centimètres des individus échantillonnés.*

Populations	Oir	Scorff	Shin	Spey
Température annuelle moyenne (°C)	3-19	5-20	1-17	1-20
Année de sècheresse	1976, 2003	1976, 2003	1976, 2003	1976, 2003
Drainage (Km^2)	87	480	494.6	3000
Débit cours d'eau (m^3/s)	0,86	5	15.2	65
Longueur (Km)	19.5	75	27.2	172
Pente (‰)	11	3.6	NA	22
Type de sol	Schiste et granite	Schiste et granite	Schiste et gneiss	Schiste et gneiss
Analyse récente du PH	7.1-7.8	7	6-7	6-9
Analyse récente des concentrations de nitrate (mg/l)	Fort : 30	18.4	12.5	27
Végétation	Rare	Abondant: *Ranunculus sp*	?	Présent *Ranunculus Sp*
Faune principale	*Salmo trutta, Anguilla anguilla, Petromyzon marina, Lampetra fluviatilis, Cottus gobio, Neimacheilus barbatula, Phoxinus phoxinus, Lampetra planeri, Gobio gobio*	*Salmo trutta, Petromyzon marinus, Salmo truta L, Phoxinus phoxinus, Petromyzon marina, Lampetra fluviatilis, Alosa alosa, Alosa agone, Esox lucius, Cottus sp, Leuciscus cephalus, Rutilus rutilus, Gobio gobio, Barbatula barbatula, Lepomis gibbosus, Leucis*	*Phoca vitulina, Salmo truta L, Salmo truta, Anguilla anguilla,Margaritifera margaritifera, Lutra lutra*	*Phoca vitulina, Lutra lutra,Salmo trutta L, , Margaritifera margaritifera, Petromyzon marinus, Anguilla anguilla, Lampetra fluvitillis, Esox lucius, Lampetra planeri, Salvelinus alpinus,*
Nombre de barrages	2	16	2	3
Stade majeur des adultes en 2005	Castillons (90%)	Castillons (90%)	Printemps + Castillons	Printemps + Castillons
Stade majeur des adultes en 1988	Castillons	Printemps + Castillons	Printemps + Castillons	Printemps + Castillons
Taille approximative observée de la population en 2005 (individus)	130	1000	3000	<60000
Taille approximative observée de la population en 1988 et 1992 pour Shin	230-260	NA	2000 - 4000	<60000
Début des projets de régulations	1976 (Plan Saumon)	1976 (Plan Saumon)	1986 (Salmon Act)	1986 (Salmon Act)
Statut actuel (IUCN)	Vulnérable	Vulnérable	Préoccupation mineure	Préoccupation mineure
Pêche côtière	Permis	Permis	En déclin	En déclin
Pêche en rivière et estuaire	Mars-Juillet (extension possible jusqu'en Octobre)	Mars-Juillet (extension possible jusqu'en Octobre)	11 Janvier au 30 Septembre	11 Février au 30 Septembre
Pression de pêche	Printemps >> Castillons	Printemps >> Castillons	Printemps >> Castillons	Printemps + Castillons
TAC	30% - 50%	30% - 50%	10-20%	10-20%
Introduction de stocks non natifs	Introduction par la Sée et Sélune	De 1973 to 1975 : Introduction de stocks Ecossais	Echappés de Saumons d'élevage (Shin) possible	Echappés de Saumons d'élevage
Introduction de stocks natifs	Aucune	Aucune	Oui depuis construction hydroélectrique vers 1950	Oui depuis 1970

Table II-1. *Caractéristiques des habitats et des populations.*

Comme nous pouvons le voir dans le tableau résumant les caractéristiques des 4 populations étudiées (Table II-1), celles-ci sont très différentes et la différence est d'autant plus forte entre les deux continents. En plus des différences géophysiques et chimiques de l'habitat, les populations ont des tailles très différentes allant de 130 à 60 000 individus. Cependant, il est à noter que la population la plus grande (Spey) se trouve dans une rivière constituée de plusieurs lits qui suppose des sous populations. Nos échantillonnages ne représentent probablement pas la métapopulation de cette rivière mais une population nettement inférieure à 60 000 individus. Une dernière différence à souligner concerne le phénomène de diminution du retour des Saumons de printemps (plusieurs hivers en mer) accompagnée par l'augmentation des castillons (un hiver en mer). Ce phénomène est plus important dans les rivières françaises que dans les rivières Ecossaises. Dans l'Oir et le Scorff, les individus qui reviennent frayer sont majoritairement des castillons et il y a de moins en moins de Saumons de printemps. Il semblerait que l'Écosse commence à subir ce même phénomène. Malgré un nombre de Saumons de printemps encore nombreux en Écosse, les centres de recherches (Fishery Board Research Office) constatent depuis plusieurs années leur diminution progressive et leur remplacement par des castillons.

1.2.Extraction d'ADN

L'extraction d'ADN a été réalisée à partir d'écailles ou de tissus de nageoires adipeuses d'environ 5mm de diamètre. Ces prélèvements ont été digérés dans 230 µl de solution contenant 10 µl de protéinase K (qui dégrade les protéines natives en hydrolysant les liaisons esters entre acides aminés), 20 µl de TE (Tris/EDTA) 10/0.1 et 200 µl de CHELEX Bio-Rad 5% (2.5 g de CHELEX pour 50 ml d'eau déminéralisée). Le mélange a été ensuite vortexé puis mis en incubation pendant plusieurs heures (de 2h à 24h suivant l'épaisseur du prélèvement) à 55°C et à 105°C pendant 15 minutes (Estoup et al. 1996; S. Launey, communication personnelle). Après repos d'une nuit à 4°C, celui-ci a été centrifugé à 4000 tours/minutes pendant 5 minutes et le surnageant a été récupéré pour être transféré dans un tube propre contenant 200 µl de CHELEX 5%. Ce dernier mélange correspond à une dilution ½ qui peut être conservée à -20°C. L'utilisation d'ADN, pour l'amplification PCR, nécessite une dernière dilution, avec du TE 10/0.1, à 1/60 s'il s'agit d'ADN extrait d'écailles et à 1/240 s'il s'agit de tissus.

Partie 2.

LES MARQUEURS

2.1. Choix et qualité des marqueurs génétiques moléculaires

Un marqueur génétique est un caractère mesurable à hérédité mendélienne (Tagu, 1999) et il est considéré idéal lorsqu'il est polymorphe (variable entre individus), discriminant (différencie les individus apparentés), multiallélique (possède plusieurs allèles sur un même locus), codominant (les hétérozygotes sont visibles), non épistatique (indépendant de l'expression des autres marqueurs), neutre (quel que soit l'allèle présent au locus, la valeur sélective de l'individu est la même), reproductible d'une expérience à l'autre, manipulable à grande échelle et économique. La répartition des marqueurs sur l'ensemble du génome est également un critère à retenir. Les principales sources de marqueurs moléculaires proviennent soit d'un polymorphisme de séquence (par exemple substitution, insertion, délétion), soit d'un polymorphisme de nombre d'unités de répétitions (microsatellites, IMA et minisatellites). Nous avons fait le choix d'utiliser les marqueurs microsatellites pour l'étude de nos populations de Saumons atlantiques car ils constituent d'excellents marqueurs génétiques. Les microsatellites sont des marqueurs constitués de motifs mono-, di-, tri- ou tétranucléotidiques répétés en tandem. En plus d'être généralement hautement polymorphes, les microsatellites possèdent les caractéristiques requises pour être un marqueur idéal. Ils sont neutres, marquant essentiellement les régions non-codantes, codominants, spécifiques d'un locus, avec un coût et une technicité abordable. Les marqueurs microsatellites sont très abondants dans le génome des organismes eucaryotes, ce qui facilite la création de panels répartis dans l'ensemble du génome et le choix de marqueurs indépendants. Ces caractéristiques sont très importantes pour les études en génétique des populations dont les théories reposent généralement sur des concepts de neutralité, d'indépendance et de grande variabilité pour différencier les populations et les individus.

Les différentes amorces microsatellites utilisées dans cette étude proviennent d'une base de données norvégiennes (http://www.salmongenome.no) incluant 850 marqueurs ADN, dans laquelle nous avons sélectionné 73 marqueurs microsatellites qui ont été étudiés sur plusieurs critères de qualité pour aboutir à un panel de 37 marqueurs polymorphes. L'étude et la sélection de ces marqueurs sont décrites dans le premier article de cette thèse intitulé «A set of 37 microsatellite markers for the analysis the genetic diversity and structure of Atlantic salmon (*Salmo salar*) population ». Cependant je profite de ce chapitre pour donner des informations supplémentaires.

Les 73 marqueurs microsatellites (Table II-2) ont été sélectionnés de telle sorte qu'ils soient polymorphes, avec un minimum de 12 répétitions, de préférence dinucléotides, pas de 'N' (région inconnue) dans la séquence, des amorces d'environ 20pb et un pourcentage de GC

avoisinant les 50%. Certaines amorces ont été redéfinies par le logiciel Primer 3, version 0.4.0 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi), pour appliquer les deux derniers critères et permettre ainsi une hybridation plus spécifique. Nous avons également sélectionné les marqueurs les plus utilisés dans la littérature ainsi que les 12 marqueurs du Projet Saumon Arc Atlantique (ASAP) initié en mai 2004. Pour finir, nous avons retenu les marqueurs indépendants : (1) localisés sur des groupes de liaison différents ou (2) distants d'au moins 50cM lorsque deux marqueurs appartenaient au même groupe (Communication personnelle concernant la carte des groupes de liaison chez *Salmo salar*, Høyheim Bjorn, non publié). Il est à noter que parmi les 14 marqueurs regroupés par deux sur un même groupe de liaison, certains sont éloignés sur la carte femelle alors qu'ils sont proches sur la carte mâle (exemple AF256715 et Z48596).

Nous avons également étudié la variabilité génétique en terme de taux d'hétérozygotie et de richesse allélique avec le logiciel GENETIX (Raymond et Rousset, 1995 ; http://www.univ-montp2.fr/~genetix/genetix/genetix.htm), ainsi que leur entropie selon l'équation de Shannon (1948) :

$$Hs(p) = -\sum_{i=1}^{k} p_i \log p_i \quad ,$$

avec *pi* la fréquence de l'allèle *i*, en un locus qui compte *k* allèles.

Grâce au logiciel FDIST2 (Beaumont et Nichols, 1996, http://www.rubic.rdg.ac.uk/~mab/software.html) nous avons détecté les marqueurs sous sélection, à partir des estimations de différenciation entre et dans les échantillons (*Fst* de Cockerham et Weir, 1993) et en simulant les *Fst* attendus sous le modèle en îles et sous un modèle de mutation (IAM ou SMM). Les locus présentant des différenciations trop fortes (+95%) ou trop faibles (-5%) sont supposés subir la sélection naturelle (Beaumont et Nichols, 1996). Dans notre étude, nous avons fait tourner nos données sous le modèle de mutation IAM avec 200 000 simulations. Nous avons également regardé si les marqueurs étaient à l'équilibre Hardy-Weinberg par le logiciel GENEPOP (Raymond et Rousset, 1995). Pour finir, nous avons étudié leur capacité à différencier les populations. Le logiciel WHICHLOCI (Banks et al. 2003 ; http://www-bml.ucdavis.edu/whichloci.htm) a permis de mettre en évidence la combinaison de marqueurs fournissant la meilleure différenciation des populations. Ce logiciel calcule un score pour chaque locus qui correspond au pourcentage de bon assignement de l'ensemble des génotypes, pour ce locus spécifique, soustrait du pourcentage de génotypes mal assignés à la population. Les distances génétiques, obtenues par l'Analyse Factorielle des Correspondances sous GENETIX, à l'intérieur et entre les

populations, nous ont permis d'évaluer également les différenciations de nos populations en fonction de l'ensemble des combinaisons de marqueurs (soit 37*37=1369). Ces résultats ont été comparés avec ceux obtenus par STRUCTURE (Pritchard et al. 2000 ; http://pritch.bsd.uchicago.edu/software/structure2_1.html) et le logiciel de Claude Chevalet (IRTA) qui permet de calculer la probabilité d'assignation des individus à leur population d'origine. Ce dernier exécute des calculs d'assignation en prenant la moitié des individus de la population au hasard pour estimer les fréquences alléliques et en calculant la probabilité d'assignation des autres individus. En répétant ceci 5000 fois au minimum, nous obtenons une distribution des probabilités par individu, ce qui permet de calculer ainsi son maximum. De plus, il possède deux particularités non négligeables :

1. Lorsqu'un individu est trop éloigné de la population la plus probable, on le met dans un groupe extérieur.
2. Lorsque les distributions de deux populations sont proches, les individus se trouvant à leurs jonctions, seront mis préférentiellement dans la population connue par l'utilisateur.

Pour plus de précisions sur les outils et les méthodes qui nous ont permis d'évaluer la qualité génétique des marqueurs microsatellite sélectionnés, veuillez vous reporter au premier article de ce mémoire intitulé «A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon (*Salmo salar*) populations».

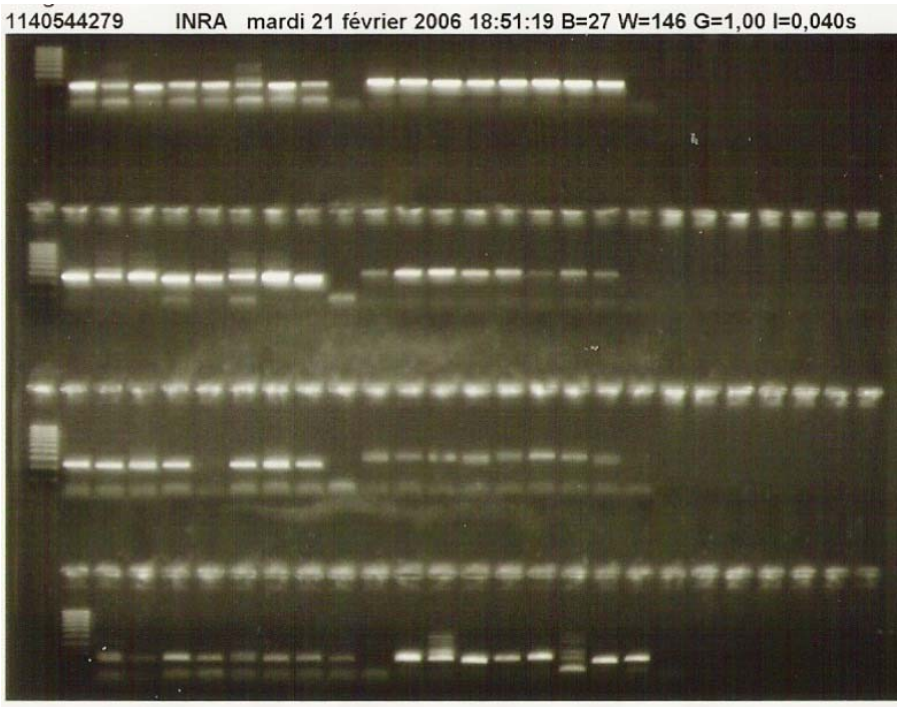


Figure II-2. *Produit d'amplification PCR obtenus avec 8 marqueurs microsatellites à partir d'ADN extrait d'écailles d'individus de Castle Grant (Spey) datant de 1988.*

GenBank	Nom marqueur	Amorce Up	Amorce Down	Rq
AF257049	SSA00149NVH	GTTTTCCCAGTCACGACGTTGGCAGCATAGAGAGTAATGG	TGGAAAAGTCCCTCACCC	
AF256693	SSA0043NVH	**GTTTTCCCAGTCACGACGTTGGATGTAAACATTCACAGGCACA**	**TGAATGAGGGGTGACCTAGC**	Refait
AF256695	SSA0045NVH	GTTTTCCCAGTCACGACGTTGGCCATTTATCACTCGACCC	GCATATACAGTGCCGTCG	
AF256749	SSA0106NVH	GTTTTCCCAGTCACGACGTTGGACCTTTTGGCTGAATGAC	TAACCGAATGACTGTGAG	
AF256719	SSA0071NVH	GTTTTCCCAGTCACGACGTTGGCCCCTGTCAAACGTCTTC	AGCACACTGGATTCAAGG	
AF256735	SSA0090NVH	GTTTTCCCAGTCACGACGTTGGTCTTACATCATACCCCAG	TCAGAAGAAGAGGCGAGC	
Z48596	SSOSL85	GTTTTCCCAGTCACGACGTTGGAGACTAGGGTTTGACCAAG	ATTCAGTACCTTCACCACC	
AF256746	SSA0103NVH	GTTTTCCCAGTCACGACGTTGGCTGTGATTTCTCTCTGC	AAAGGTGGGTCCAAGGAC	
AF256699	SSA0049NVH	GTTTTCCCAGTCACGACGTTGGAGAGGAAATTCCACCGTG	CTTTACCTATTTTTGGCAGG	
AF256756	SSA0116NVH	GTTTTCCCAGTCACGACGTTGGCCCCAAACTTTTGAACGG	AGAGGCATATCAATCCTAC	
AF256674	SSA0023NVH	**GTTTTCCCAGTCACGACGTTGGACACAGATCACATGAAAGACACG**	**AGGAGGAAAAGGGGAGACAC**	Refait
AF256751	SSA0109NVH	GTTTTCCCAGTCACGACGTTGGCCAATGATGTATATGGCG	TTTGTGAATGGGAGACCG	
AF256672	SSA0021NVH	GTTTTCCCAGTCACGACGTTGGACTTGGAGACTCTTTGG	GAGAGGGAGATAGCATCG	
AF256843	SSA0214NVH	GTTTTCCCAGTCACGACGTTGGCTGCCCAACCTCATTACC	GTGAGACTGTTAGAGCTGG	
AF256680	SSA0030NVH	GTTTTCCCAGTCACGACGTTGGCCAAATAACTGACAAGTGAG	CAGAGGTTGATAATGGGG	
AF256845	SSA0216NVH	GTTTTCCCAGTCACGACGTTGGCACTGGGGTTTAATGTC	TGTATAGGGGCAATCAGC	
AF256661	SSA0008NVH	GTTTTCCCAGTCACGACGTTGGTCCTCTATCATACGGCTG	ATCAGATGACCCAGTGGC	
AF256662	SSA0011NVH	GTTTTCCCAGTCACGACGTTGGTTACACAGCCCTGCTCAC	TCCTGTCACACTCACTACC	
AF256667	SSA0016NVH	GTTTTCCCAGTCACGACGTTGGTGAAACTAGGATGCCTGG	TCTGACCCACACACAAGC	
AF256694	SSA0044NVH	GTTTTCCCAGTCACGACGTTGGCACACCTCAGCTCCTC	TTCCCCTGCCACCTAGC	
AF256707	SSA0057NVH	GTTTTCCCAGTCACGACGTTGGTGGCTACAACAGGGATAC	AGTCTCTTACATGGAGGC	
AF256846	SSA0217NVH	GTTTTCCCAGTCACGACGTTGGAGCGAGCTTTCTTTCCAG	AGCTGTCTATTCACGACTC	
AF256683	SSA0033NVH	GTTTTCCCAGTCACGACGTTGGTGGTAGATGTCTTCTCCC	GATCTGGTGACCTGTTCC	
AF256786	SSA0152NVH	GTTTTCCCAGTCACGACGTTGGCTGTTCATTTCTGAGCAG	GACACACCGAATCAGTGC	
AF256686	SSA0036NVH	GTTTTCCCAGTCACGACGTTGGTAGATGAGATCAGGGCAG	CATGTGGAGAAACATTCAC	
AF257058	SSA0146NVH	GTTTTCCCAGTCACGACGTTGGAGTGCGTCTGTGGCTTTG	GGGCAGAGAGAAGAGATAG	
AF271427	Alu005	GTTTTCCCAGTCACGACGTTGGTATGTGATTAGGGCTTGC	CTTGGCGTAGTTTAGTGC	
AF009796.1	Ogo4	GTTTTCCCAGTCACGACGTTGGTCGTCACTGGCATCAGCTA	GAGTGGAGATGCAGCCAAAG	
AF256762	SSA0122NVH	GTTTTCCCAGTCACGACGTTGGTCTAGGTAGCAGCCTCAG	CCCTGTGGTTCAAAGAGG	
U43694	SSA197	GTTTTCCCAGTCACGACGTTGGTGAGTAGGGAGGCTTGTG	TGACATAACTCTTCTATGGC	ASAP
U43695	SSA202	GTTTTCCCAGTCACGACGTTGGAGGTAACTGGCTCAACTC	ATGTTGCGTGAGTGAGTG	ASAP
unknow	SSA289	GTTTTCCCAGTCACGACGTTGGCTTTACAAATAGACAGACT	TCATACAGTCACTATCATC	ASAP
unknow	SSA14	GTTTTCCCAGTCACGACGTTGGCCTTTTGACAGATTTAGGATTTC	CAAACCAAACATACCTAAAGCC	
U86703	SSleei84	GTTTTCCCAGTCACGACGTTGGGTAACAAACACACTGCTC	CTGAGAGATGGACAGAGTG	
U43693	SSA171	GTTTTCCCAGTCACGACGTTGGTGATGGCAGGGTAAGAGG	GGTCAAAACCTCGCTGTG	ASAP
U43692	SSA85	GTTTTCCCAGTCACGACGTTGGACTGCGGGACATTTGAGG	TCAGATTTCTTACACTCTCG	
U86706	SSleen82	GTTTTCCCAGTCACGACGTTGGCCTGCTATCATGGAGAATC	ACATGGTACTCATGCAGAG	
Z48597	SSOSL311	GTTTTCCCAGTCACGACGTTGGGAGGAACTGCATTCTCTG	TCTGAAAGGGCACTAACCG	
Z49134	SSOSL438	GTTTTCCCAGTCACGACGTTGGTGACAACACACAACCAAGG	GTAAAATGGAAGCATCTGTG	

Z48581	SSOSL25	GTTTTCCCAGTCACGACGTTGGTTGGGATCTACACAGCTCC	GGGTCGAGAGAAGTGACAC	
AF256674	SSA0023NVH	GTTTTCCCAGTCACGACGTTGGAAAGACACGGAGCAAGGC	AAGACAGGAGTCTGGGTG	
AF256708	SSA58NVH	GTTTTCCCAGTCACGACGTTGGAACAACTTCAGAACTTGAC	CGCCTCATAGCTGATATTTAAC	
AF019168	SSA224*	GTTTTCCCAGTCACGACGTTGGACAGACAGAACTGTGCATC	**TCGATTTTGGTTGACTGCAT**	Refait
AF019184	SSA65	GTTTTCCCAGTCACGACGTTGGTGTTGTGGCTCGTGACAG	GAACACAGGGTAGAGTGG	
AY081807	SSsp2201	GTTTTCCCAGTCACGACGTTGGTTTAGATGGTGGGATACTGGGAGGC	CGGGAGCCCCATAACCCTACTAATAAC	ASAP
AF525203.	SSAD144	GTTTTCCCAGTCACGACGTTGGTTGTGAAGGGGCTGACTAAC	TCAATTGTTGGGTGCACATAG	ASAP
AY081811	SSsp2216	GTTTTCCCAGTCACGACGTTGGGGCCCAGACAGATAAACAAACACGC	GCCAACAGCAGCATCTACACCCAG	ASAP
AF525204.	SSAD157	GTTTTCCCAGTCACGACGTTGGATCGAAATGGAACTTTTGAATG	GCTTAGGGCTGAGAGAGGAATAC	ASAP
AY081808	SSsp2210	GTTTTCCCAGTCACGACGTTGAAGTATTCATGCACACACATTCACTGC	CAAGACCCTTTTTCCAATGGGATTC	ASAP
AY081813	SSspG7	GTTTTCCCAGTCACGACGTTGGCTTGGTCCCGTTCTTACGACAACC	TGCACGCTGCTTGGTCCTTG	ASAP
AF525208	SSAD486	GTTTTCCCAGTCACGACGTTGGTCGCTGTGTATCAGTATTTTGG	ACTCGGATAACACTCACAGGTC	ASAP
AY081812	SSsp1605	GTTTTCCCAGTCACGACGTTGGCGCAATGGAAGTCAGTGGACTGG	CTGATTTAGCTTTTTAGTGCCCAATGC	ASAP
AF256687	SSA0037NVH	GTTTTCCCAGTCACGACGTTGGCACTAATGCACAGTGTCAG	GCATAAATGGCATGTGTTC	
AF256688	SSA0038NVH	GTTTTCCCAGTCACGACGTTGGCACTAATGCACAGTGTCAG	TGTGGTCGCATAAATGGC	
AF256697	SSA0047NVH	GTTTTCCCAGTCACGACGTTGGTCTGTCACTGTCACCCTG	AGAGCGGCTGGTATAATC	
AF256698	SSA0048/1NVH	GTTTTCCCAGTCACGACGTTGGCAGAACCGTGATCTGAAG	**TCCTCTGTGACACTGCATCC**	Refait
AF256703	SSA0053NVH	GTTTTCCCAGTCACGACGTTGGGAATTGTGTGTAGATGGG	ACTTGGCAAGGAGCAGAC	
AF256705	SSA0055NVH	GTTTTCCCAGTCACGACGTTGGAATAAGAGGGCAGTGGAG	TGCACCAGAGAGAGTAGC	
AF256706	SSA0056NVH	**GTTTTCCCAGTCACGACGTTGGCACCTGATTTACCTGACCACA**	TCTAGGCTATCGCACAGC	Refait
AF256712	SSA0062NVH	**GTTTTCCCAGTCACGACGTTGGAGGTGAAGGAAAGGGTGTTG**	GACTAAAAAGCGTCTGGC	Refait
AF256714	SSA0064NVH	GTTTTCCCAGTCACGACGTTGGCCTGCCATCATCCAACTC	TCCACACCCAACATACTC	
AF256715	SSA0065NVH	GTTTTCCCAGTCACGACGTTGGCAACACAAACACATTTGC	TATGGAGAGGGTTGGTAG	
AF256723	SSA0075NVH	GTTTTCCCAGTCACGACGTTGGTCTGTCCGTCCTGCATAC	CTTATGTTGTGTTGTGCTG	
AF256728	SSA0082NVH*	GTTTTCCCAGTCACGACGTTGGAGAGCGAATACAACAGCC	**AGGTCCTGGATGTTCTGCAA**	Refait
AF256730	SSA0084NVH	GTTTTCCCAGTCACGACGTTGGACCTCAGCACATGAACAC	TGACAGAGCCATAGACCG	
AF256731	SSA0086NVH	GTTTTCCCAGTCACGACGTTGGATGGGTGCTATTGACTC	CCACACAATCACCGTTGC	
AF256729	SSA0083NVH	GTTTTCCCAGTCACGACGTTGGGTAAGTCAAGGTTTCACC	TTACTCCCCAACTCTGAG	
AF256739	SSA0094NVH	GTTTTCCCAGTCACGACGTTGGAAACTAGGGGTGCATAGG	AATGTATTGACCGTAGTAGC	
AF256744	SSA0101NVH	GTTTTCCCAGTCACGACGTTGGCAGCACCAGAACATAACC	AGCCATCAACACTCCCTG	
AF256745	SSA0102NVH	GTTTTCCCAGTCACGACGTTGGATGGATGAAGGAAGGACC	**CTGCTGCAATCTCACCAATC**	Refait
AF256788	SSA0155NVH	GTTTTCCCAGTCACGACGTTGGCACCTGAAGATCAGCATC	GCTGTGGATTTAGGAGAG	
AF019191	SSA86	GTTTTCCCAGTCACGACGTTGGACTAGATGAAGCGTGTGC	CTGAATGCTACTGATGCC	
AF019197	SSA9	GTTTTCCCAGTCACGACGTTGGACAATTACCAGAGCCAGG	TCTGACGACAAATTACCAC	

Table II-2. *Amorces des 73 marqueurs testés avec ajout de la séquence M13 sur l'amorce Up. En caractères gras sont représentées les séquences redéfinies avec Primer 3. Les marqueurs ASAP sont mentionnés en dernière colonne et les 36 marqueurs qui ont été éliminés, après sélection, sont en rouge.*

Nom	Genbank	Autre Nom	Groupe de liaison	Référence
Alu005	AF271427	ALU005	L26-f, L26-m	Hoyheim et Jorgensen, 2000
BHMS111	AF256661	SSA0008NVH	L14-f, L14-m	Hoyheim B, 2000
BHMS155	AF256667	SSA0016NVH	L16-f, L16-m	Hoyheim B, 2000
BHMS176	AF256672	SSA0021NVH	L8-f, L8-m	Hoyheim B, 2000
BHMS181	AF256674	SSA0023NVH	L6-f, L6-m	Hoyheim B, 2000
BHMS241	AF256683	SSA0033NVH	L20-m	Hoyheim B, 2000
BHMS259	AF256687	SSA0037NVH	L16-f, L16-m	Hoyheim B, 2000
BHMS278	AF256693	SSA0043NVH	L2-f, L2-m	Hoyheim B, 2000
BHMS283	AF256695	SSA0045NVH	L23-f, L23-m	Hoyheim B, 2000
BHMS304/1	AF256698	SSA0048/1NVH	L22-f, L22-m	Hoyheim B, 2000
BHMS365	AF256703	SSA0053NVH	L19-f, L19-m	Hoyheim B, 2000
BHMS377	AF256707	SSA0057NVH	L18-f1, L18-m	Hoyheim B, 2000
BHMS386	AF256712	SSA0062NVH	L15-f, L15-m	Hoyheim B, 2000
BHMS404	AF256715	SSA0065NVH	L10-f, L10-m	Hoyheim B, 2000
BHMS429	AF256719	SSA0071NVH	L3-f, L3-m	Hoyheim B, 2000
BHMS328	AF256731	SSA0086NVH	L5-f, L5-m	Hoyheim B, 2000
BHMS212	AF256739	SSA0094NVH	L2-f, L2-m	Hoyheim B, 2000
BHMS331	AF256744	SSA0101NVH	L17-f, L17-m	Hoyheim B, 2000
BHMS179A	AF257058	SSA0146NVH	L24-m	Hoyheim B, 2000
BHMS184B	AF257049	SSA0149NVH	L11-f, L11-m	Hoyheim B, 2000
BHMS217	AF256786	SSA0152NVH	L21-f, L21-m	Hoyheim B, 2000
BHMS230	AF256788	SSA0155NVH	L9-f, L9-m	Hoyheim B, 2000
BHMS235	AF256846	SSA0217NVH	L19-f, L19-m	Hoyheim B, 2000
SSA171	U43693	SSA171	L4-f, L4-m	O'Reilly et al, 1995
SSA197	U43694	SSA197	L8-f, L8-m	O'Reilly et al, 1995
SSA202	U43695	SSA202	L1-f, L1-m	O'Reilly et al, 1995
SSA224	AF019168	SSA224	L28-m	O'Reilly et al, 1997
SSA289	unknow	SSA289	L13-f, L13-m	King et al., 2001
SSA65	AF019184	SSA65	L29-f, L29-m	O'Reilly et al, 1997
SSA85	U43692	SSA85	L18-f1, L18-m	O'Reilly et al, 1995
SSA86	AF019191	SSA86	L11-f	O'Reilly et al, 1997
SSA9	AF019197	SSA9	L25-f, L25-m	O'Reilly et al, 1997
SSAD144	AF525203	SSAD144	?	Eackles et King, 2002
SSAD157	AF525204	SSAD157	?	Eackles et King, 2002
SSOSL438	Z49134	SSOSL438	L9-f, L9-m	Slettan, 1995
SSOSL85	Z48596	SSOSL85	L10-f, L10-m	Slettan, 1995
SSsp2216	AY081811	SSSP2216	L7-f, L7-m	Verspoor et al, 2002

Table II-3. *Informations supplémentaires sur les 37 marqueurs sélectionnés et utilisés pour les analyses du Saumon atlantique avec les appartenances au groupe de liaison chez les mâles (m) et les femelles (f) selon une communication personnelle de B. Hoyheim.*

2.2. PCR

La PCR (Polymerase Chain Reaction) est une méthode d'amplification *in vitro* qui permet d'amplifier exponentiellement la quantité d'un fragment d'ADN cible par répétition de réactions d'élongation grâce à une enzyme dite l'ADN polymérase. Au sein de ce projet, nous avons travaillé à partir d'une méthode de marquage développée par Markus Schuelke (2000). Cette méthode modifie le protocole classique d'amplification PCR puisqu'elle utilise une amorce supplémentaire de courte séquence (GTTTTCCCAGTCACGACGTTG(G)) appelée « M13 » qui est marquée par fluorescence (6-Fam/Ned/Hex) et que l'on notera F_M13. L'utilisation de cette amorce supplémentaire (F_M13) nécessite un allongement de l'amorce Up en 5' pour permettre l'hybridation de F_M13. L'amorce Up modifiée par allongement sera notée M13_U. On se retrouve donc avec 3 amorces au lieu de deux : l'amorce Up allongée (M13_U), l'amorce M13 fluorescente (F_M13) et l'amorce Down (reverse) complémentaire de l'amorce Up. Comme on peut le voir sur la Figure II-3, l'amplification se déroule lors des 2 premiers cycles comme une PCR classique avec :

(i) Étape de dénaturation : Les deux brins d'ADN à amplifier sont séparés par élévation de la température (environ 95°C) afin d'obtenir des molécules monocaténaires (un brin).

(ii) Étape d'hybridation : Les deux amorces (M13_U et Down) sont hybridées sur le brin monocaténaire cible par diminution de la température entre 40°C et 65°C.

(iii) Étape d'élongation : Les extrémités 3'-OH des oligonucléotides servent de points de départ à l'ADN polymérase thermostable qui synthétise les brins complémentaires.

Puis, à partir du 3ème cycle, lorsque le brin d'ADN néosynthétisé a amplifié la séquence complémentaire M13, une réaction de compétition entre l'amorce M13_U et F_M13 apparaît. L'amorce F_M13 sera principalement incorporée dans l'ADN néosynthétisé car celle-ci a été ajoutée en quantité supérieure (1 : 3) ce qui permettra le marquage des produits PCR (amplicons) par le flurochrome qu'elle porte (Figure II-3). À la différence des autres méthodes PCR, il n'y a aucun marquage préalable des deux amorces Up et Down. Ceci permet une grande flexibilité dans l'élaboration des jeux de marqueurs ainsi qu'un coût moindre puisque les amorces non marquées sont moins coûteuses.

Les conditions d'amplification optimales pour chacun des marqueurs (Table I de l'article 1, Chapitre III) ont été fixées en faisant varier la température d'hybridation et la concentration en MgCL2+ puisque ce dernier agit en modifiant les charges portées par les différentes bases nucléotidiques. La température d'hybridation tourne généralement autour de

T_{hyb} qui se calcule en sommant les bases nucléotidiques (Guanine (G), Cytosine (C), Adénine (A) et Thymine (T)) se trouvant sur un brin d'ADN:

$$T_{hyb} = 4(G+C) + 2(A+T) - 5°C\,.$$

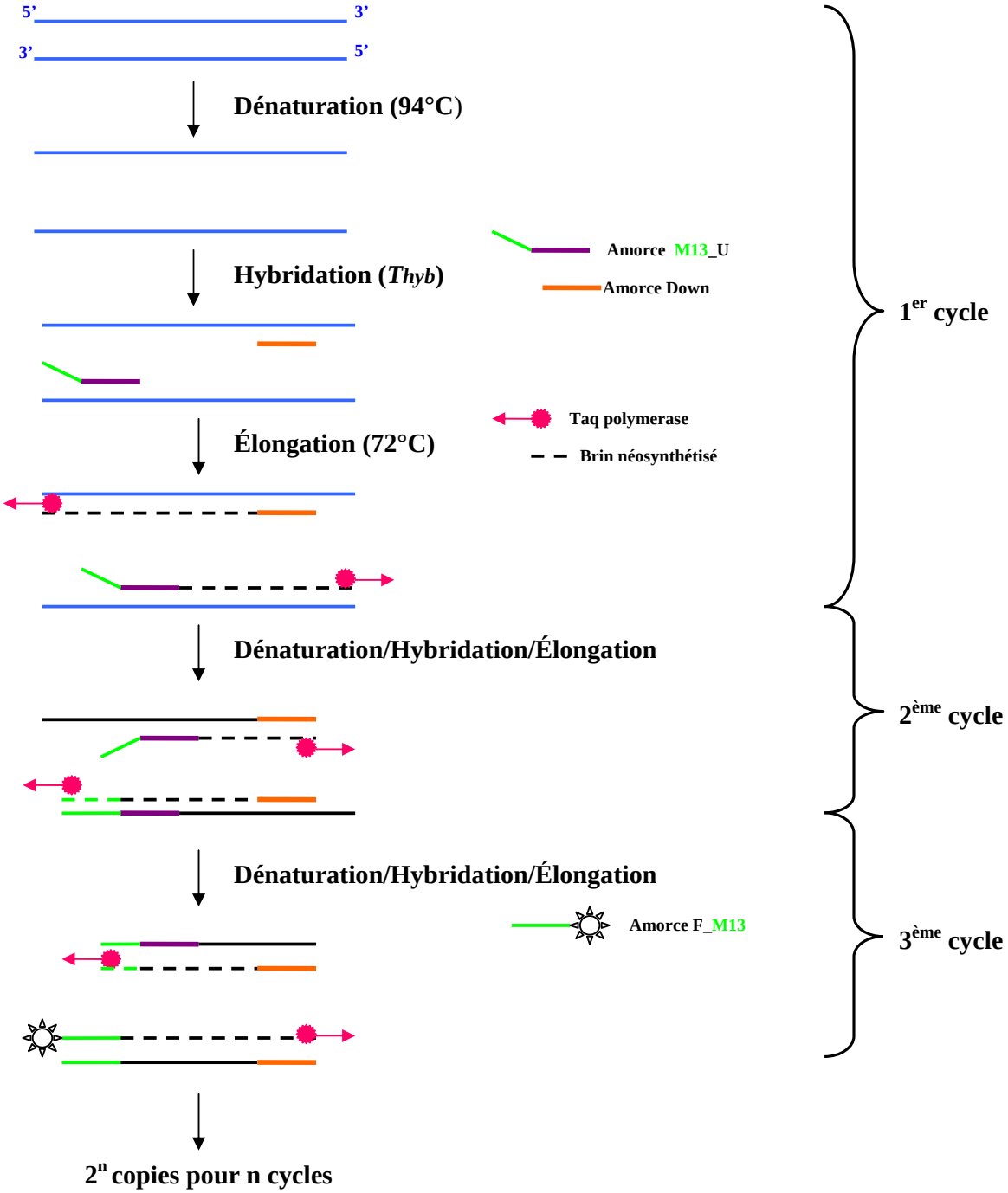


Figure II-3. *Étapes d'amplification PCR avec l'amorce fluorescente M13 (F_M13) et la séquence Up allongé avec la séquence M13 en 5' (M13_U). Le marquage de la séquence amplifiée se fait par simple hybridation de F_ M13 sur la complémentaire.*

2.3. Protocole d'amplification systématique avec la méthode M13

Les réactions ont été effectuées à l'aide des robots de la Plateforme Génomique du GENOPOLE de Toulouse (http://www.genotoul.fr/). Le robot pipeteur, TECAN 200, nous a permis de répartir l'ensemble des individus (367) sur une même plaque 384 échantillons et de distribuer le mix PCR. Le volume final de 10 µL était composé de tampon PCR 1X (GoTaq Promega), 200 µM de dNTP, 1.5 mM MgCl2, 0.1 µM d'amorce Up allongé (M13_U), 0.15 µM d'amorce Down, 0.15 µM d'amorce M13 marqué par Fluorescence (F_M13), 50 ng d'ADN et 0.5 U d'ADN Polymérase (GoTaq Promega). La réaction d'amplification PCR a été réalisée dans un thermocycler GeneAmp PCR System 9700 (Applied Biosystems). Les conditions d'amplification sont les suivantes : 5 min de dénaturation initiale à 94°C puis de 42 à 45 cycles suivant les marqueurs avec 30 s de dénaturation à 94°C, 30 s d'hybridation à la température optimale pour chaque couple d'amorces et 30 s d'élongation à 72°C, enfin 30 min de synthèse finale à 72°C. Pour chacune de nos PCR, nous avons introduit un témoin négatif (tous les composants de la réaction sans l'ADN matrice) pour vérifier qu'il n'y a pas eu de contamination(s) externe(s).

2.4. Génotypages des microsatellites

La migration des fragments d'ADN s'effectue dans un polymère dérivé de l'acrylamide (pop6) dans 48 capillaires du séquenceur ABI PRISM 3730 (Applied Biosytems). Cette méthode permet d'étudier plusieurs microsatellites de même taille mais portant des fluorophores différents, au même moment, en multiplexant les produits d'amplification PCR dans un même puits (Figure II-4). Pour cela, 5 µL de chacun des produits d'amplification ont été mélangés et complétés avec de l'eau milliQ autoclavée pour obtenir un volume final de 50 µL. Par la suite, un volume de 2 µL a été prélevé de ce mélange (maximum 6 produits pour 3 fluorophores avec deux marqueurs par couleur), pour être ajouté à 9,850 µL de formamide et 0,175 µL de standard de taille (GENESCAN 400HD [ROX] d'Applied Biosystems). Les échantillons ont été ensuite dénaturés, 5 min à 95°C, avant d'être placés dans le séquenceur. Lors de la migration des produits PCR dans les capillaires du séquenceur, les fluorophores sont excités par un laser et l'émission de fluorescence est interprétée par le logiciel Genotyper 3.7 NT (Applied Biosystems). Le standard de taille permet d'attribuer une taille aux différents pics d'intensité. Seuls les profils possédant une intensité supérieure à 300 *nM* ont été conservés sur l'ensemble des génotypages. Les données ont été ensuite conservées, stockées, tracées et organisées dans le logiciel GEMMA

(Iannuccelli *et al.* 1996). Il est à noter que 10 individus ont été dupliqués sur l'ensemble des marqueurs pour évaluer l'erreur de génotypage.

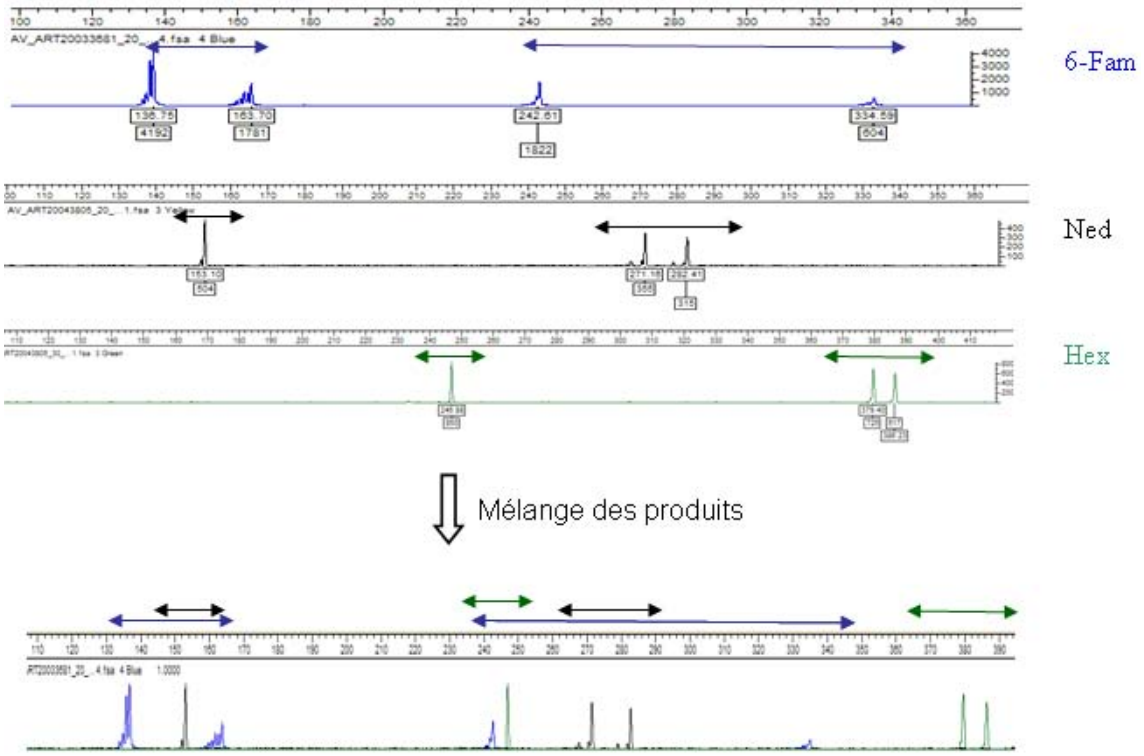


Figure II-4. *Chaque flèche horizontale représente un microsatellite, et la longueur des flèches la fourchette de tailles attendues pour ses allèles. Ici, nous avons représenté deux microsatellites par fluorophore (6-Fam, Ned et Hex) et nous pouvons voir l'importance des tailles dans l'élaboration de nos jeux de marqueurs pour qu'il n'y ait pas de recouvrement. Pour chaque marqueur la visualisation de 2 pics représente un individu hétérozygote et un seul pic un individu homozygote.*

2.5. Présentation des données

Les données de génotypages pour chacune de nos populations se présentent sous forme matricielle avec, en colonne, les marqueurs et, en ligne, les individus. Chaque marqueur est représenté par deux formes alléliques possédant une taille propre en paire de bases (pb). Pour la plupart des analyses en génétique des populations, les données peuvent être utilisées sous forme brute (pb) ou par la nomination de chaque allèle en chiffre distinctif. Cependant les méthodes fonctionnant sous SMM (voir le Glossaire) à partir des données de microsatellite font leurs calculs en fonction du nombre de répétitions (ex : ATATATATATAT = 6 répétitions). Les données ont donc ainsi été transformées en nombre de répétitions à partir de la longueur de la séquence amplifiée (pb).

La matrice de génotypage globale représente les génotypages de l'ensemble des individus par le panel des 37 marqueurs microsatellites sélectionnés. A partir de cette matrice, un classement par hétérozygotie décroissante a été effectué pour créer 9 matrices par population et par échantillon (2005 et passé). La première matrice, utilisée pour toutes les analyses présentées ci-après, concernait les génotypages sur la totalité des marqueurs. Les autres matrices, qui contenaient les 28 marqueurs les plus hétérozygotes (28H+), les 28 marqueurs les moins hétérozygotes (28H-), les 20 marqueurs les plus hétérozygotes (20H+), les 20 marqueurs les moins hétérozygotes (20H-), les 10 marqueurs les plus hétérozygotes (10H+), les 10 marqueurs les plus hétérozygotes (10H-), les 5 marqueurs les plus hétérozygotes (5H+) et les 5 marqueurs les moins hétérozygotes (5H-), ont été utilisées seulement pour les analyses d'estimation de la taille efficace.

Partie 3.

MÉTHODES D'ANALYSE DES DONNÉES

Nous présenterons dans un premier temps les logiciels utilisés pour étudier la taille efficace (*Ne*), la diversité génétique, la consanguinité, les liens de parenté, la phylogénie, la migration, et la présence ou non d'un goulot d'étranglement sur l'ensemble des échantillons de Saumons étudiés.

Concernant les modèles qui ont été élaborés au cours de ces travaux de thèse, leurs méthodologies se trouvent avec la partie des résultats (Chapitre III).

3.1. Logiciels utilisés

3.1.1. Sites Web

BOTTLENECK	http://www.ensam.inra.fr/URLB
CREATE	http://www.lsc.usgs.gov/CAFL/Ecology/Software.html
DIYABC	http://www.montpellier.inra.fr/CBGP/diyabc/
GENALEX 6	http://www.blackwell-synergy.com/doi/abs/10.1111/j.1471-8286.2005.01155.x
GENECLASS2	http://www.montpellier.inra.fr/URLB/index.html
GENEPOP	http://ftp.cefe.cnrs.fr/PC/MSDOS/GENEPOP
GENEPOP on the web	http://wbiomed.curtin.edu.au/genepop
GENETIX	http://www.univ-montp2.fr/~genetix/genetix/genetix.htm
IMa	http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#IM
LAMARC	http://evolution.gs.washington.edu/lamarc/lamarc_prog.html
MSVAR	http://www.rubic.rdg.ac.uk/~mab/software.html
TM3	http://www.rubic.rdg.ac.uk/~mab/software.html
STRUCTURE	http://pritch.bsd.uchicago.edu/software/structure2_1.html

Table II-4. *Sites informatiques des logiciels en génétique des populations utilisés.*

3.1.2. Fonctions des logiciels

LOGICIELS	AUTEURS	FONCTIONS
BOTTLENECK	Cornuet J-M et Luikart G	Détecte la réduction de taille récente à partir des fréquences alléliques.
DIYABC	Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ, Guillemaud T, Estoup A,	Évalue la taille efficace actuelle et ancestrale des populations, le temps qui sépare ces deux tailles, le taux de mutation et les scénarios évolutifs entre populations depuis l'ancêtre commun avec un échantillon ou plusieurs échantillons par population. Méthode avec approximation Bayésienne.
CREATE	Coombs et al.	Transforme des fichiers très utilisés en génétique des populations (GENETIX, GENEPOP..) sous des formats d'autres logiciels de génétique des populations.
FDIST2	Beaumont MA et Nichols RA	Détecte les marqueurs soumis à la sélection: *Fst* trop grand ou trop faible.
FSTAT	Goudet J	Programme complet sur les indices en génétique des populations pour évaluer la diversité et la divergence génétique.
GENALEX	Peakall R, et Smouse P.E.	Calcule la diversité et la divergence génétique. Teste l'équilibre d'Hardy-Weinberg et les déséquilibres de liaison.
GENECLASS2	Piry S, Alapetite A, Cornuet, J-M, Paetkau D, Baudouin L, Estoup, A	Détecte les migrants et assigne les individus à leur population d'origine
GENEPOP	Raymond M et Rousset F	Calcule la diversité et la divergence génétique. Teste l'équilibre d'Hardy-Weinberg et les déséquilibres de liaison.
GENETIX	Belkhir K. et al.	Programme complet sur les indices en génétique des populations pour évaluer la diversité et la divergence génétique.
IM	Hey J et Nielsen R	Évalue la taille efficace de deux populations actuelles, la taille de leur ancêtre commun et leur migration par un modèle IM (Isolation with Migration model). C'est un modèle en île.
LAMARC	Kuhner MK, Yamato J, Beerli P, Smith LP, Rynes E, Walkup E, Li C, Sloan J, Colacurcio P, Felsenstein J	Évalue la taille efficace des populations, le taux de mutation, le taux de croissance et de migration avec un échantillon. Méthode Bayésienne et MCMC.
MICRO-CHEKER	Van Oosterhout CV, Hutchinson WF, Wills M, Shipley P	Détecte les allèles nuls au sein de marqueurs microsatellites.
MSVAR	Beaumont MA	Évalue la taille efficace actuelle et ancestrale des populations, le temps qui sépare ces deux tailles, et le taux de mutation avec un échantillon. Méthode Bayésienne et MCMC.
STRUCTURE	Pritchard JK, Stephens M, Donnelly PJ	Identifie des groupes et assigne les individus aux groupes les plus probables.
TM3	Berthier P, Beaumont MA, Cornuet JM, Luikart G	Évalue la taille efficace actuelle et ancestrale des populations avec deux échantillons. Méthode Bayésienne et MCMC.
WHICHLOCI	Banks MA,Eichert W et Olsen JB	Détermine la puissance de discrimination d'un ensemble de marqueurs microsatellites.

Table II-5. *Caractéristiques des logiciels en génétique des populations utilisés.*

3.1.3. Références bibliographiques des logiciels

BOTTLENECK

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.

CREATE

Coombs JA, Letcher BH, Nislow KH. 2007. *CREATE 1.0 – Software to create and convert codominant molecular data*.

DIYABC

Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ, Guillemaud T, Estoup A. 2008. *Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation*. Bioinformatics, 24(23): 2713-2719.

FDIST2

Beaumont MA, Nichols RA. 1996. *Evaluating loci for use in the genetic analysis of population structure*. Proceedings of the Royal Society B 263: 1619–1626.

FSTAT

Goudet J. 1995. *Fstat version 1.2: a computer program to calculate Fstatistics*. Journal of Heredity. 86(6): 485-486.

GENALEX

Peakall R, Smouse PE. 2006. *GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research*. Molecular Ecology Notes 6: 288–295.

GENETIX

Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F. 1998. *GENETIX, Logiciel sous WindowsTM pour la Génétique des Populations*. Laboratoire Génome et Populations, CNRS UPR 9060. Université de Montpellier II, Montpellier, France.

GENECLASS

Piry S. *et al*. 2004. GeneClass2: *A software for genetic assignment and first- generation migrant detection*. J. Hered. 95: 536–539.

GENEPOP

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

IM

Hey J, R Nielsen. 2004. *Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis*. Genetics 167: 747–760.

LAMARC

Kuhner MK. 2006. *"LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters."* Bioinformatics 22(6): 768-770.

MICRO-CHEKER

Van Oosterhout CV, Hutchinson WF, Wills DPM, Shipley P. 2004. *Micro-checker: software for identifying and correcting genotyping errors in microsatellite data*. Molecular Ecology Notes, 4: 535–538.

MSVAR

Beaumont MA. 1999. *Detecting population expansion and decline using microsatellites*. Genetics 153: 2013–2029.

STRUCTURE

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

Falush D, Stephens M, Pritchard JK. 2003. *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies*. Genetics 164: 1567–1587.

TM3

Berthier P, Beaumont MA, Cornuet JM, Luikart G. 2002. *Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach*. Genetics. 160(2): 741-51.

WHICHLOCI

Banks MA, Eichert W, Olsen JB. 2003. *Which genetic loci have greater population assignment power?*. Bioinformatics 19: 1436–1438.

3.2. Principes des analyses faites en génétique des populations

3.2.1. La diversité génétique et la consanguinité

La diversité génétique est généralement appréciée à partir des estimations sur l'hétérozygotie (voir le Glossaire) et de la richesse allélique (voir le Glossaire). L'estimation de ces deux paramètres permet de renseigner sur la structure des populations et leur viabilité en observant les différences entre les hétérozygoties observées et attendues sous l'équilibre de Hardy-Weinberg (voir le Glossaire) ainsi que la présence des allèles rares. Les valeurs de diversité génétique permettent d'estimer le coefficient de parenté, appelé aussi coefficient de consanguinité. Parce que la reproduction entre individus fortement apparentés entraîne l'accumulation d'allèles délétères et l'apparition de maladies ou de malformations phénotypiques, estimer la parenté entre individus permet d'avoir une idée sur l'état de santé de la population voir de l'espèce. Le coefficient de consanguinité intra population, F_{IS}, a été estimé sur les populations de Saumons étudiées. Il mesure le déficit en hétérozygotes par rapport à la valeur attendue sous l'hypothèse de croisement au hasard des gamètes :

$$F_{IS} = 1 - \frac{Ho}{He},$$

avec *Ho* et *He*, les fréquences respectives observées et attendues d'hétérozygotes. Cette valeur, ainsi estimée, correspond également à la probabilité que 2 allèles observés à un locus soient identiques « par descendance ».

3.2.2. Différenciation génétique

Dans un premier temps, une analyse factorielle des correspondances (AFC) a été menée pour évaluer le niveau de différenciation entre populations. Cette analyse consiste à projeter les distances génétiques entre individus sur un jeu d'axes qui maximisent les différences entre populations. Elle a été effectuée à l'aide du logiciel GENETIX v4.05.2 (Belkhir *et al.* 2004).

À partir des coordonnées individuelles obtenues par ces résultats, nous avons calculé des distances Euclidiennes (voir Glossaire) intra (*Dw*) et inter (D_B) populations et un indice de différenciation qui n'est autre que le rapport D_B/Dw (voir l'article 1 au Chapitre III).

Un autre paramètre de différenciation entre populations, usuellement utilisé, a été mesuré ici. Il s'agit de l'estimateur θ du F_{ST}, proposé par Weir et Cockerham (1984) (voir Glossaire) calculé par le logiciel GENETIX v4.05.2 mais qui peut être également calculé par le logiciel FSTAT (Goudet 2002). Ce paramètre mesure l'écart entre le taux d'hétérozygote attendu globalement, si les populations échangent librement des gènes (panmixie), et le taux d'hétérozygotes moyen observé dans les populations. Cette mesure peut également être vue comme la proportion de diversité totale expliquée par l'écart des fréquences alléliques entre populations :

$$\theta = \frac{\sigma^2 a}{\sigma^2 a + \sigma^2 b + \sigma^2 w},$$

$\sigma^2 a$ étant la variance inter populations, $\sigma^2 b$ la variance intra population et $\sigma^2 w$ la variance intra individus.

3.2.3. Migration et taux de croissance

Lorsque des populations échangent entre elles des reproducteurs et sont à l'équilibre (modèle en îles), le taux de migration (*Nm*) peut être relié à la formule de F_{ST} de Wright (1969) :

$$Nm = \frac{1 - FST}{4FST}.$$

Dans notre étude, les calculs de *Nm* ont été faits à partir du θ proposé par Weir et Cockerham (1984) par le logiciel GENETIX. Mais, le *Nm* peut également être calculé à partir du nombre d'allèles privés (Slatkin, 1985 ; Barton and Slatkin, 1986), qui permet d'éliminer les problèmes d'homoplasies, par le logiciel GENEPOP. En 1985, Slatkin releva une relation linéaire entre la moyenne des fréquences d'allèles privés d'une population ($\overline{p}(1)$) et *Nm*:

$$\log_{10}\left[\overline{p}(1)\right] = a \log_{10}(Nm) + b,$$

avec *a* et *b* des variables qui dépendent du nombre d'individus des échantillons par population. Néanmoins les valeurs de *Nm* obtenues par ce dernier modèle sont plus faibles que celles obtenues par F_{ST} ou R_{ST} sans que nous puissions répondre pourquoi.

Comme, les calculs de migration, basés sur *Nm*, dépendent de la différenciation, ils peuvent êtres biaisés suivant l'histoire des populations, si celle-ci ne correspond pas au modèle en île supposé. Par exemple, si des populations proviennent d'une séparation récente à

partir d'un ancêtre commun et si le taux de mutation est faible, une faible différenciation génétique peut exister alors qu'aucun échange migratoire n'est présent entre ces populations.

Nous avons donc estimé également le taux d'immigration *m* dans chaque population par le logiciel LAMARC (Kuhner, 2006) qui est une estimation plus complexe ne faisant pas intervenir les différences entre populations. Ce sont des calculs sous un modèle de coalescence (identité par descendance) avec migration et mutation. Néanmoins les résultats n'ont pas été satisfaisants en raison du grand nombre de marqueurs utilisés rendant les temps de calculs très longs (3-6 semaines). Il y avait une incohérence entre nos connaissances et les résultats obtenus, notamment sur les paramètres « thêta » (*4Nu*), laissant des doutes sur les valeurs proposées des migrations. Ainsi nous avons préféré considérer les taux de migration obtenue par le logiciel IM (Hey et Nielsen, 2004) après 500 000 burning et 5000 000 itérations.

3.2.4. La taille efficace

Le concept de taille efficace (*Ne*) est apparu pour la première fois en 1931 avec le généticien américain Sewall Green Wright connu pour ces travaux sur la théorie de l'évolution et sa découverte du coefficient de consanguinité avec John Scott Haldane. Elle peut être estimée à partir de données démographiques ou génétiques. Les estimations de *Ne* par données démographiques sont présentées brièvement ici, juste pour rendre compte de la nécessitée d'utiliser des données de qualité.

3.2.4.1. Méthodes démographiques

Selon Nunney et Elam (1994), la taille efficace sur une période de temps donnée correspond à la moyenne harmonique du nombre d'individus qui interviennent dans la reproduction au cours des générations correspondantes, c'est-à-dire la moyenne des tailles efficaces à chaque g-ème génération (Ne_g) :

$$Ne = t / \sum (1/ Ne_g) .$$

L'effet de la variance de la taille familiale (V_k) est généralement estimé par l'équation suivante, où *N* est le nombre d'individus naissant à chaque génération :

$$Ne = (4N - 2)/(V_k + 2) .$$

L'effet par le sex-ratio peut être estimé par cette équation qui fait intervenir le nombre de mâles (N_M) et de femelles (N_F) dans la population :

$$Ne = 4N_F N_M /(N_F + N_M).$$

3.2.4.2. Méthodes génétiques

Notre étude s'est portée exclusivement sur les estimations de *Ne* par les données génétiques qui sont plus faciles à obtenir que les données démographiques. Cependant, comme nous le verrons par la suite dans les résultats obtenus, l'utilisation de ces données et de leurs estimateurs doivent être pris avec précaution.

Il existe plusieurs estimateurs génétiques de *Ne* (voir les revues de Caballero 1994 ; Beaumont 2001) que l'on pourrait résumer en deux types de méthodes :

- Une méthode s'appuyant sur les changements génétiques entre deux échantillons, tels que les fréquences alléliques, la perte d'allèles, la perte d'hétérozygotie, le taux de déséquilibre de liaison et l'augmentation du coefficient de consanguinité. Cette méthode estime plus précisément la variance induite par la dérive génétique au cours de l'intervalle de temps séparant les deux échantillons. La dérive génétique correspond aux fluctuations aléatoires de fréquences alléliques liées à l'échantillonnage d'un nombre fini de reproducteurs, fluctuations pouvant aller jusqu'à la perte de certains allèles dans la population. La méthode tire partie du fait que l'importance de ces fluctuations est fonction de la taille efficace recherchée.

- L'autre utilisant la théorie de la coalescence de Kingman (1982) qui cherche à retracer la généalogie des gènes de l'échantillon jusqu'à l'ancêtre commun (modèle de coalescence). Cette méthodologie relie la longueur des branches de coalescence à *Ne*. Plus la longueur des branches sera longue et plus la taille sera supposée grande. Les méthodes usuelles reposant sur ce modèle cherchent à reconstruire par simulation des arbres possibles des gènes en fonction du modèle de mutation. Les calculs correspondants sont cependant très longs. Comme nous le verrons dans le chapitre des résultats par l'un des modèles développé au cours ce cette thèse (Chapitre III), nous pouvons calculer les temps de coalescence en fonction des allèles observés et l'histoire démographique sans simuler ces arbres. Cette manière numérique permet un calcul plus rapide de la vraisemblance de la taille efficace et de sa variation dans le passé.

Au cours de nos travaux, nous avons utilisé trois méthodes pour estimer la taille efficace : TM3 (Berthier et al. 2003), MSVAR (Beaumont 1999) et DIY ABC (Cornuet et al. 2008). Toutes les trois sont des méthodes de coalescence sous dérive et mutation, c'est-à-dire que la migration n'intervient pas. La première a seulement la particularité d'utiliser deux échantillons de la même population pour faire ses calculs. Elle est en quelque sorte un mélange entre les deux types de méthodes, citées au-dessus, puisqu'elle fait intervenir également les différences entre les deux échantillons. Quant à MSVAR et DIY ABC, ils se différencient sur deux points majeurs. Alors que MSVAR fait ses calculs sur une population, DIY ABC fait ses calculs sur l'ensemble des populations données pour la même espèce. Ainsi, DIY ABC estime une seule taille ancestrale (*Na*) et un seul temps ancestral (temps entre les deux tailles, *Tf*) pour l'ensemble des populations. Le second point réside dans l'algorithme qui est du Métropolies Hastings pour MSVAR et de l'ABC pour DIY ABC.

Partie 4.

NOUVEAUX MODÈLES EN DIVERSITÉ GÉNÉTIQUE ET TAILLE EFFICACE

Au cours de ces travaux deux modèles ont été développés sur les bases de la théorie de la coalescence : *DemoDivMS* (Diversity Simulations) et *VarEff* (Variable Effective size). Leur développement a nécessité un ensemble de résolutions analytiques qui sont présentées dans le chapitre III « Résultats et Discussions » de ce mémoire.

Le premier modèle *DemoDivMS*, proposé en libre accès sur le site https://qgp.jouy.inra.fr/, permet de simuler des variations de taille « thêta » ($4N_e\mu$) d'une population à des intervalles de temps définis par l'utilisateur. De ces simulations en découle un calcul de la diversité génétique en terme de taux hétérozygotie au temps zéro (moment actuel). Ce modèle permet de rendre compte du comportement des variations de la diversité en fonction de l'histoire de la population.

Le second modèle *VarEff* est un modèle qui estime la taille efficace d'une population à partir d'un jeu de données constitué de génotypages microsatellites référencés en nombre de répétitions. À la différence des modèles déjà existant, ce modèle ne reconstruit pas les arbres pour estimer ces paramètres et passe directement par des calculs analytiques tout en ayant une approche bayésienne. Le développement de cette méthode est présenté dans le chapitre III « Résultats et Discussions » au sein de l'article III et sera très prochainement proposé en libre accès sur le site https://qgp.jouy.inra.fr/.

Ces modèles faisant l'objet de résultats nouveaux, seules les simulations faites sur ces modèles seront présentées dans cette partie. Il est prévu d'étendre les simulations, pour le second modèle, en créant différentes populations à l'équilibre et sous dérive et en faisant varier le nombre de locus.

4.1. Théorie de coalescence

La théorie de la coalescence cherche à retracer la généalogie des gènes de l'échantillon jusqu'à leur ancêtre commun. C'est-à-dire qu'elle décrit le processus de fusion (coalescence) des copies alléliques présentes dans une population jusqu'à l'ancêtre commun. Cette théorie dérive de la théorie de l'identité par descendance de Malécot (1941) développée par Kimura & Crow (1964) et généralisée par Kingman (1982) pour l'appliquer à un échantillon de plusieurs gènes dont on considère l'ancêtre commun. Ce processus s'applique sur un échantillon de taille n issu d'une population diploïde de taille N ou haploïde de taille $2N$ en supposant $n<<N$, de telle sorte qu'il ne peut y avoir qu'un seul évènement de coalescence par génération. Cela signifie que le travail se fait sur l'arbre des gènes observés dans l'échantillon, et non pas celui de la population entière. Pour visualiser cette généalogie, il faut ignorer les individus et ne s'intéresser qu'aux gènes. Les relations de ces gènes d'une

génération à l'autre sont des lignes ascendantes qui sont communément appelées des lignages (*l*). Lorsque deux lignages se rejoignent, on dit que les gènes « coalescent ». La probabilité que deux lignages proviennent d'une même copie de la génération précédente (identique par ascendance) est égale à 1*/2N*. Cependant, on doit considérer qu'un lignage peut coalescer avec n'importe quel autre lignage avec la même probabilité. On peut donc former *l(l*-1*)/*2 paires différentes. La probabilité d'évènements de coalescence de 2 lignages parmi *l* lignages est égale à :

$$P(l) = \frac{l(l-1)}{4N}.$$

La probabilité qu'il n'y ait aucun évènement de coalescence est donc *1-P(l).*

Le temps de coalescence *Ti* peut être considéré comme le nombre de générations écoulées jusqu'à ce que l'on ait un évènement de coalescence. Ces temps de coalescence dépendent donc de *P(l),* ce qui en fait des variables aléatoires dont la distribution de probabilité peut s'écrire :

$$P(Ti = t) = \left[1 - P(l)\right]^{t-1} P(l).$$

Ce qui revient à dire que pendant *t-1* génération, il n'y a pas eu de coalescence et à la *t-jème* génération, il y a coalescence.

Comme les temps de coalescence sont proportionnels à la taille efficace de la population, elle-même liée à la diversité génétique, on peut entrevoir par cette présentation sommaire le principe de la coalescence pour en déduire la taille efficace ou la diversité génétique. Comme il a été dit précédemment, la méthodologie des deux modèles, développés dans ces travaux, n'est pas présentée dans ce chapitre mais au chapitre suivant en tant que résultats.

4.2. Simulations *DemoDivMS*

Le premier modèle DemoDivMS a été utilisé pour comprendre la forte diversité génétique des plus petites populations de Saumon atlantique (Oir et Scorff). Ce modèle utilise les résolutions théoriques développées pour le deuxième modèle : *P(D|N,M,T)* avec *D* la différence de taille en nombre de répétitions entre deux allèles, *N* la taille efficace, *M* le taux de mutation et *T* le temps de coalescence souscrit en génération par *g=T/No* (voir Article 3, Chapitre III). Plusieurs scénarios ont été simulés faisant varier 3 tailles efficaces : actuelle

(*N0*) et passées (*N1, N2*) sur des paliers de temps en générations variables (de *t0* à *t3*) (Figure II-5) avec des taux de mutation à 3.10^{-4} et 9.10^{-4}. En fonction des résultats que nous avions obtenus par les analyses en génétiques de populations pour ces deux populations, nous avons fait varier :

- *N0* de 200 à 1000
- *N1* de 50 à 50 000
- *N2* de 10 000 à 50 000
- *t1* de 2 à 1000
- *t2* de 2000 à 4000 (temps ancêtre commun donné par DIY ABC et MSVAR)
- *t3* pour avoir la population à l'équilibre c'est-à-dire *4N/4*, soit de 10 000 à 50 000.

Puis nous avons retenu les combinaisons offrant des diversités génétiques comparables à celles observées (76-81%).

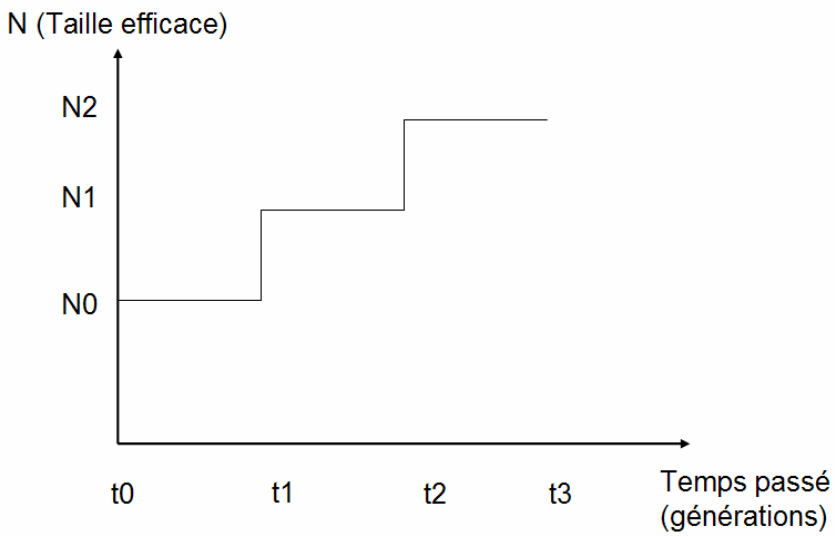


Figure II-5. *Schéma évolutif des tailles efficaces (N) sur des paliers de temps correspondants à des intervalles de deux temps de générations (t) antérieurs.*

Les taux de migrations par population ont été obtenus à partir du logiciel IM. Comme notre modèle ne permet pas d'intégrer les migrations, nous avons fait les mêmes estimations en considérant ce taux de migration (0.004 et 0.005) comme des mutations, en l'additionnant au taux de mutation (0.0003 et 0.0005).

4.3. Simulations *VarEff*

4.3.1 Distributions des *P(D=k|N,M,T)*

Afin de tester le comportement des fréquences des différences alléliques (*D*) théoriques en fonction de la taille efficace (*N*), du taux de mutation (*M*) et du temps de coalescence (*T*), ici exprimé en générations (*g=T/No*), nous avons testé plusieurs scénarios avec différentes intensités de goulot d'étranglement. Pour des raisons de visualisation, nous avons choisi de tester un goulot d'étranglement sur une grande taille, comme celle de la population Spey (10 000). Nous avons fait varier l'intensité du goulot en augmentant la durée (de 50 à 1000 générations) et en diminuant la taille (de 5000 à 500) avec un taux de mutation de 0.0003.

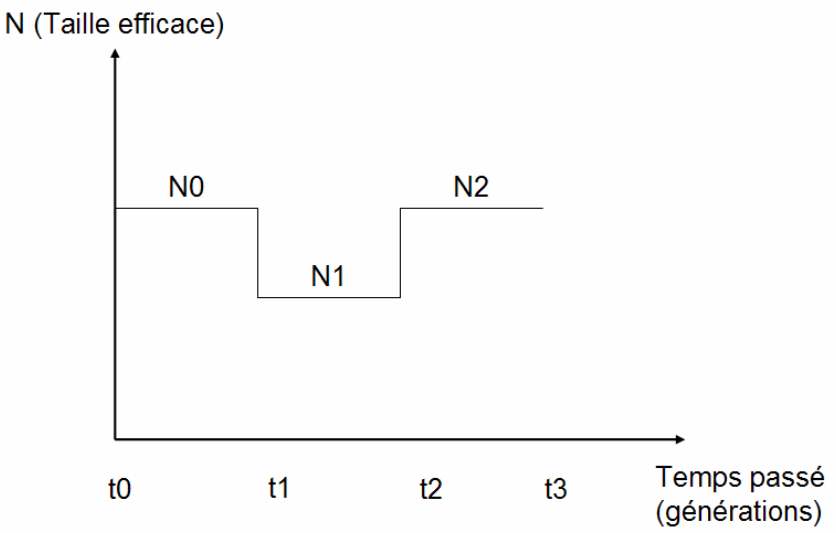


Figure II-6. *Schéma évolutifs décrivant un goulot d'étranglement avec les tailles efficaces (N) sur des paliers de temps correspondants à des intervalles de deux temps de générations (t) antérieurs.*

En raison des grandes similarités entre les résultats et pour faciliter la lecture et la compréhension, vous trouverez dans la dernière partie du Chapitre III uniquement les simulations avec les données extrêmes (Figure II-6):

N0 et *N2* égal à 10 000

N1 égal à 5000 ou 500

Intervalle *t0* - *t1* égal à 2500 (temps ancestral trouvé au sein des populations étudiées)

Intervalle *t1* - *t2* égal à 50 ou 1000

4.3.2 Distributions des *P(T|D,N,M)*

Des simulations présentant le comportement des temps de coalescence (*T*) par rapport au *D*, *N* et *M* sont présentées au sein du troisième article intitulé « *Distribution of coalescence times and distances between microsatellite alleles with changing effective population size* » (Chapitre III).

III

RÉSULTATS ET DISCUSSIONS

Le Saumon atlantique est une espèce dont les stocks s'effondrent, ce qui nécessite des analyses complémentaires urgentes pour mieux connaître son abondance et sa viabilité. Les données démographiques et/ou génétiques sont un moyen efficace pour estimer ces stocks. Toutefois, entre ces deux modes d'investigation, les données génétiques présentent des avantages certains. Tout d'abord, leurs récoltes d'échantillons sont moins difficiles et moins coûteuses. En second lieu, elles permettent de projeter l'histoire des populations sur des intervalles de temps beaucoup plus longs. À ce jour, on constate qu'en dépit de ces avantages, la plupart des populations de salmonidés sont étudiées et suivies avec des données démographiques. Ces travaux de thèse sont les premiers à fournir des analyses approfondies, riches en marqueurs, en génétique des populations de salmonidés dans l'Oir et le Scorff, en France, et à Spey et Shin, en Écosse.

La diversité génétique, la consanguinité, les flux génétiques, la détection de goulot d'étranglement et la taille efficace sont des paramètres issus de données génétiques qui renseignent sur la structure et le potentiel évolutif des populations. Tous ces indices ont été estimés au cours de ces travaux, mais un intérêt majeur s'est porté sur la taille efficace (*Ne*). Je me suis intéressée à ce paramètre, car il est l'un des plus importants en thématique de conservation par sa capacité à renseigner sur la variabilité génétique, la fixation des allèles délétères et la consanguinité. Ces dernières décennies, des estimateurs de *Ne*, de plus en plus performants, ont vu le jour. Parmi les plus utilisés et qui ont, semble-t-il, un avenir prometteur, il faut noter ceux qui se basent sur le modèle de coalescence. Les performances de tels estimateurs sont généralement testées par simulations. Très peu de travaux les ont évalués à partir de données réelles. L'objectif de cette thèse a donc été d'évaluer certains d'entres de ces estimateurs à partir de données de populations de Saumon atlantique, afin de proposer des recommandations dans des programmes de conservation. Devant les résultats obtenus et, face à certaines problématiques qu'engendre leur utilisation, une seconde partie de cette thèse s'est consacrée à élaborer une nouvelle méthode d'estimation de taille efficace sous coalescence. Celle-ci a été construite dans un souci de réduction des temps de calculs. Pour cela, l'approche de ce modèle a été abordée de manière innovante puisqu'elle ne s'appuie pas sur la reconstruction par simulation des arbres de coalescence. Elle calcule directement les probabilités d'observer *Ne* en analysant les différences alléliques.

Ce chapitre s'articule donc en 3 parties. La première comprend l'élaboration des données génétiques avec la sélection et l'analyse des marqueurs. La seconde comprend

l'ensemble des analyses faites pour déterminer la diversité et la structure génétique, et estimer la performance des certains estimateurs de la taille efficace. La troisième partie concerne le modèle proposé pour estimer la taille efficace, de manière analytique, avec son comportement vis-à-vis des données simulées. Ce modèle servira ainsi à retracer l'histoire évolutive des populations de Saumons étudiées, afin de comprendre leur diminution actuelle. Chacune de ces parties est composée d'un article scientifique avec, en amont, son résumé en français et, en aval, un paragraphe sur les données complémentaires. Une discussion de synthèse sur l'ensemble de ces travaux se poursuivra dans le prochain chapitre.

Partie 1.

ÉLABORATION DES DONNÉES GÉNÉTIQUES

ARTICLE 1

A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon *Salmo salar* populations

NATACHA NIKOLIC
KATIA FEVE
CLAUDE CHEVALET
BJORN HØYHEIM
JULIETTE RIQUET

Journal of Fish Biology (2009) 74, 458–466

Résumé

Un panel de marqueurs ADN microsatellites du Saumon atlantique (*Salmo salar*) a été analysé et sélectionné à partir d'une large base de données déjà existante sur des critères techniques, économiques et génétiques. Nous nous sommes attachés à optimiser un panel en vue de travaux sur la diversité génétique et la structure de Saumon atlantique sauvage et anadrome. Ainsi, nous avons pu identifier un ensemble de 37 marqueurs microsatellites polymorphes qui sont faciles à utiliser et qui fournissent un haut niveau de variabilité génétique et de différenciation entre et parmi les populations.

Nous avons choisi les marqueurs microsatellites car ils ont démontré leur grande capacité à différencier des populations apparentées, par rapport à d'autres marqueurs communément utilisés en génétique des populations tels que les SNP (Polymorphisme d'un Nucléotide Simple) (Narum et al. 2008).

Journal of Fish Biology (2009) **74,** 458–466
doi:10.1111/j.1095-8649.2008.02094.x, available online at http://www.blackwell-synergy.com

A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon *Salmo salar* populations

N. Nikolic*, K. Fève*, C. Chevalet*, B. Høyheim‡ and J. Riquet*

**INRA, UMR444 Laboratoire de Génétique Cellulaire, Chemin de Borde Rouge BP 52627, 31326, Castanet Tolosan, France and ‡Norwegian School of Veterinary Science, BasAM-Genetics, PO Box 8146, 0033 Oslo, Norway*

(Received 26 May 2008, Accepted 17 September 2008)

Atlantic salmon *Salmo salar* microsatellite markers from a large database were analysed and selected with technical, economic and genetic criteria to provide an optimized set of polymorphic DNA markers for the analysis of the genetic diversity and the structure of anadromous Atlantic salmon populations. A set of 37 microsatellite markers was identified that are easy to use and provide a high level of differentiation power.

Key words: anadromous fish; differentiation; genetic diversity; microsatellite.

During the past decade, microsatellite genetic markers have been extensively developed in several species to address population genetics and demographic questions (Estoup & Angers, 1998; Dieringer & Schlötterer, 2003). Analysis of neutral molecular markers has proven to be a robust method for bringing an evolutionary perspective to conservation and management (King *et al.*, 2001). Over recent decades, Atlantic salmon *Salmo salar* populations have declined or have been extirpated in many parts of its distribution L. (Parrish *et al.*, 1998; ICES, 2002; Knockaert, 2006). Many microsatellite markers have been developed on *S. salar*, but it is necessary to select a panel that allows genetic surveys to be carried out efficiently. Following the recommendations on the number and the quality of markers used in the field of domestic animal resources (Barker *et al.*, 1993; SanCristobal *et al.*, 2006), a set of 37 markers was selected from a large data set (Norwegian Salmon Genome Project's database, http://www.salmongenome.no), including *c.* 850 markers. Markers were selected on technical and genetic criteria and tested for their efficiency using samples from four European rivers contrasting in environment and demographic structure.

†Author to whom correspondence should be addressed at present address: Tel.: +33 561285568; fax: +33 561285308; email: natacha.nikolic@toulouse.inra.fr

Genomic DNA was extracted from *S. salar* fin tissues and scales by heating samples in 230 μl solution [proteinase K, TE (TrisEDTA) buffer and chelex] at 55° C several hours and then at 105° C for 15 min (Estoup *et al.*, 1996; S. Launey, pers. comm.). An economic approach, called the M13 method (Schuelke, 2000), was used to label DNA polymerase chain reaction (PCR) amplifications. This method uses three primers: a forward primer with a M13-specific tail at the 5′ end, a specific reverse primer and a universal fluorescent-labelled M13 primer. As this method does not need the specific primer pairs to be prelabelled, it reduces the costs significantly. This method is as easy as a classical amplification method and is flexible, enabling three or four fluorescent dyes to be used for a set of markers. Several modifications to the original labelling protocol of Schuelke (2000) were applied. PCR was carried out in a 10 μl reaction volume containing 1·5 mM $MgCl_2$, 200 μM dNTPs, 0·1 μM forward primer, 0·15 μM reverse primer, 0·15 μM M13-Fluo, 25–50 ng DNA and 0·5 U Taq DNA polymerase. The amplification conditions were as follows: an initial denaturation for 5 min at 94° C, then 42–45 cycles for 30 s at 94° C, 30 s at Tm° C (annealing temperature), 30 s at 72° C and to finish at a final synthesis for 30 min at 72° C. For a sub-set of markers, primers had to be redefined to obtain satisfactory results (Table I) using Primer 3 0.4.0 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi).

The assessment of a marker's quality was based on technical and molecular criteria, and level of genetic variability [allelic richness (A), observed (H_o) and expected (H_e) heterozygosity and entropy (E); Table I]. First, 73 *S. salar* microsatellite markers were selected for high levels of polymorphism, diploidy, spacing across genome (unpubl. data) and their use in the scientific literature or projects from *c.* 850 markers in the database (King *et al.*, 2001; Landry & Bernatchez, 2001; Vasemägi *et al.*, 2001; Tiira *et al.*, 2003; Withler *et al.*, 2005). These 73 markers were tested on 367 wild adult anadromous *S. salar* from four European rivers (L'Oir and Scorff in France, Shin and Spey in Scotland), and 37 markers were selected for their high quality of amplification, easy peak identification and high level of genetic variability. Finally, the efficiencies of these markers to differentiate populations and to assign individuals to their river were evaluated.

Most of the 37 markers had a large number of alleles per locus (A), ranging from 5 to 41 alleles. In fact, 30 markers had at least 10 alleles, and eight markers had 30 or more alleles in the sample. Allelic size range was larger than mentioned in the Norwegian Salmon Genome Project's database and varied from 15 bp(*Ssa86*) to 119 bp (*SsaD157*). The mean H_o in 367 individuals varied from 0·37 to 0·93, and the mean within-population H_e varied from 0·35 to 0·93. These results are concordant with a previous study of 11 microsatellites markers (King *et al.*, 2005). Most markers (30) showed heterozygosities larger than 0·70. The Shannon entropy index ($E = -\sum p_i log\, p_i$ where p_i's are allele frequencies) provided the same ranking of markers, with 30 markers showing large mean values between 1·5 and 3·0. Only three markers (*Ssa86*, *BHMS179A* and *BHMS304/1*) had a lower diversity with respect to the three indicators (H_o, A and E). Exact tests of deviation from Hardy–Weinberg expectations (HWE) were calculated using GENEPOP 3.4 (Raymond & Rousset, 1995). Six markers (*Alu005*, *BHMS283*, *SSsp2216*, *BHMS259*, *BHMS217* and *SsaD157*)

TABLE I. Characteristics of 37 *Salmo salar* microsatellite loci and measures of information content derived from four different wild populations. Markers are sorted as in Fig. 2 in decreasing classification efficiencies. Markers selected as the most efficient ones according to WHICHLOCI programme are indicated in bold. *A*, total number of alleles; C, number of amplification cycles; Class, classes of efficiency; D_B/D_W, ratio of the mean Euclidian distances between individuals from different populations (D_B) to the mean distances between individuals from the same populations (D_W); *E*, mean of entropy; H_e, mean of expected heterozygosity; H_o, mean of observed heterozygosity; Ind, number of individuals genotyped; Repeat, type of motif repeat (dinucleotide or tetranucleotide); *S*, allele size range; S_d, difference between the minimum and maximum allele size; Tm° C, annealing temperature. *Redefined primers

Class	Name	GenBank	Forward primer	Reverse primer	Repeat	Tm° C	C	Ind	*S*	S_d	*A*	H_o	H_e	*E*	D_B/D_W
1	**ALU005**	AF271427	TATGTGATTAGGGCTTGC	CTTGGCGTAGTTTAGTGC	Dinucleotide	58	45	341	179–263	84	32	0·84	0·89	2·67	1·64
1	**Ssa171**	U43693	TGATGGCAGGGTAAGAGG	GGTCAAAACCTCGCTGTG	Tetranucleotide	56	45	367	225–297	72	24	0·80	0·83	2·11	1·63
1	**BHMS283**	AF256695	CCATTTATCACTCGACCC	GCATATACAGTGCCGTCG	Dinucleotide	58	45	330	255–373	118	39	0·80	0·87	2·47	1·62
1	**BHMS235**	AF256846	AGCGAGCTTTCTTTCCAG	AGCTGTCTATTCACGACTC	Dinucleotide	52	45	191	150–213	63	23	0·91	0·80	2·40	1·61
1	**BHMS377**	AF256707	TGGCTACAACAGGGATAC	AGTCTCTTACATGGAGGC	Dinucleotide	54	45	341	130–200	70	32	0·90	0·89	2·70	1·60
1	**BHMS331**	AF256744	CAGCACCAGAACATAACC	AGCCATCAACACTCCCTG	Dinucleotide	54	42	341	158–244	86	40	0·93	0·92	2·88	1·59
1	**Ssa65**	AF019184	TGTTGTGGCTCGTGACAG	GAACACAGGGTAGAGTGG	Dinucleotide	56	45	367	213–260	47	21	0·88	0·86	2·27	1·56
1	**BHMS181**	AF256674	AAAGACACGGAGCAAGGC	AAGACAGGAGTCTGGGTG	Dinucleotide	56	45	367	97–158	61	19	0·89	0·83	2·13	1·55
1	**BHMS278**	AF256693	**ATGTAAACATTCACAGGCACA***	**TGAATGAGGGGTGACCTAGC***	Dinucleotide	55	42	341	185–238	53	23	0·88	0·86	2·31	1·55
1	**SsaD144**	AF525203	TTGTGAAGGGGCTGACTAAC	TCAATTGTTGGGTGCACATAG	Tetranucleotide	58	45	367	135–253	118	41	0·93	0·93	2·99	1·53
2	**Ssa9**	AF019197	ACAATTACCAGAGCCAGG	TCTGACGACAAATTACCAC	Dinucleotide	54	42	367	203–261	58	27	0·89	0·88	2·56	1·53
2	**BHMS365**	AF256703	GGAATTGTGTGTAGATGGG	ACTTGGCAAGGAGCAGAC	Dinucleotide	58	42	294	244–301	57	26	0·87	0·86	2·38	1·51
2	**Ssa197**	U43694	TGAGTAGGGAGGCTTGTG	TGACATAACTCTTCTATGGC	Tetranucleotide	56	45	367	173–284	111	29	0·85	0·84	2·24	1·51
2	**BHMS328**	AF256731	GATGGGTGCTATTGACTC	CCACACAATCACCGTTGC	Dinucleotide	54	42	323	204–301	97	40	0·85	0·93	3·07	1·51
2	**BHMS429**	AF256719	CCCCTGTCAAACGTCTTC	AGCACACTGGATTCAAGG	Dinucleotide	50	45	312	203–254	51	22	0·82	0·84	2·14	1·48
2	**BHMS230**	AF256788	CACCTGAAGATCAGCATC	GCTGTGGATTTAGGAGAG	Dinucleotide	54	42	367	198–283	85	34	0·93	0·93	3·03	1·48
2	**BHMS212**	AF256739	AAACTAGGGGTGCATAGG	AATGTATTGACCGTAGTAGC	Dinucleotide	58	42	367	143–220	77	18	0·83	0·83	2·06	1·48
2	**SSsp2216**	AY081811	GGCCCAGACAGATAAACAAACACGC	GCCAACAGCAGCATCTACACCCAG	Tetranucleotide	60	45	367	224–302	78	18	0·81	0·85	2·22	1·48
2	**BHMS386**	AF256712	**AGGTGAAGGAAAGGGTGTTG***	GACTAAAAAGCGTCTGGC	Dinucleotide	54	42	367	233–282	49	18	0·79	0·84	2·09	1·48
2	**Ssa85**	U43692	ACTGCGGGACATTTGAGG	TCAGATTTCTTACACTCTCG	Dinucleotide	56	45	367	139–199	60	25	0·81	0·85	2·26	1·47
2	**BHMS259**	AF256687	CACTAATGCACAGTGTCAG	GCATAAATGGCATGTGTTC	Dinucleotide	54	42	294	150–218	68	28	0·85	0·85	2·30	1·47
2	**BHMS241**	AF256683	TGGTAGATGTCTTCTCCC	GATCTGGTGACCTGTTCC	Dinucleotide	56	45	341	150–212	62	29	0·87	0·87	2·43	1·47
2	**BHMS404**	AF256715	GCAACACAAACACATTTGC	TATGGAGAGGGTTGGTAG	Dinucleotide	54	42	272	146–245	99	28	0·86	0·87	2·45	1·46
2	**SSOSL85**	Z48596	AGACTAGGGTTTGACCAAG	ATTCAGTACCTTCACCACC	Dinucleotide	56	44	367	136–183	47	22	0·86	0·83	2·10	1·45
2	Ssa86	AF019191	ACTAGATGAAGCGTGTGC	CTGAATGCTACTGATGCC	Dinucleotide	54	42	367	129–144	15	7	0·37	0·36	0·74	1·45
2	Ssa289		CTTTACAAATAGACAGACT	TCATACAGTCACTATCATC	Dinucleotide	54	45	367	125–141	16	5	0·63	0·60	1·10	1·45
2	SSOSL438	Z49134	TGACAACACACAACCAAGG	GTAAAATGGAAGCATCTGTG	Dinucleotide	58	45	367	112–146	34	10	0·75	0·73	1·54	1·44
2	**Ssa202**	U43695	AGGTAACTGGCTCAACTC	ATGTTGCGTGAGTGAGTG	Tetranucleotide	54	45	367	77–126	49	13	0·83	0·84	2·06	1·44
3	**BHMS217**	AF256786	GCTGTTCATTTCTGAGCAG	GACACACCGAATCAGTGC	Dinucleotide	52	45	316	271–302	31	12	0·88	0·80	1·83	1·44

TABLE I. Continued

Class	Name	GenBank	Forward primer	Reverse primer	Repeat	Tm° C	C Ind	*S*	S_d	*A*	H_o	H_e	*E*	D_B/D_W
3	**SsaD157**	AF525204	ATCGAAATGGAACTTTTGAATG	GCTTAGGGCTGAGAGAGGAATAC	Tetranucleotide	58	45 367	318–437	119	30	0·90	0·91	2·78	1·43
3	BHMS111	AF256661	TCCTCTATCATACGGCTG	ATCAGATGACCCAGTGGC	Dinucleotide	52	45 367	120–148	28	10	0·72	0·71	1·51	1·40
3	Ssa224	AF019168	ACAGACAGAACTGTGCATC	**TCGATTTTGGTTGACTGCAT***	Dinucleotide	56	45 367	203–223	20	8	0·69	0·70	1·42	1·39
3	BHMS155	AF256667	TGAAACTAGGATGCCTGG	TCTGACCCACACACAAGC	Dinucleotide	54	42 367	170–189	19	10	0·68	0·67	1·39	1·39
3	BHMS184B	AF257049	GCAGCATAGAGAGTAATGG	TGGAAAAGTCCCTCACCC	Dinucleotide	55	42 367	165–187	22	8	0·57	0·56	1·07	1·38
3	BHMS176	AF256672	GACTTGGAGACTCTTTGG	GAGAGGGAGATAGCATCG	Dinucleotide	56	42 367	139–157	18	8	0·70	0·70	1·34	1·37
3	BHMS176A	AF257058	AGTGCGTCTGTGGCTTTG	GGGCAGAGAGAAGAGATAG	Dinucleotide	56	42 367	177–194	17	5	0·48	0·49	0·84	1·36
3	**BHMS304/1**	AF256698	CAGAACCGTGATCTGAAG	**TCCTCTGTGACACTGCATCC***	Dinucleotide	54	42 367	159–185	26	8	0·41	0·35	0·78	1·34

showed a significant deviation from HWE ($P < 0{\cdot}01$) in only one or two populations, and one marker (*BHMS328*) in three populations, suggesting that null alleles might be segregating in these populations. Departures from neutrality were detected with FDIST2 (Beaumont & Nichols, 1996) for markers *BHMS235*, *BHMS230*, *BHMS328* and *SsaD157* ($P < 0{\cdot}01$), which had the smallest F_{ST} values.

Previous studies showed little genetic divergence between *S. salar* populations, a result that can be explained by the likely recent colonizations 8000–10 000 years before present (Crossman & Mcallister, 1986; Ståhl, 1987; Davidson *et al.*, 1989; Verspoor, 1997, King *et al.*, 2001). Given the low level of differentiation among the four populations examined ($F_{ST} \approx 0{\cdot}05$), the value of the selected markers was checked using factorial correspondence analysis (FCA) and a Bayesian approach for detecting population structure (STRUCTURE; Pritchard *et al.*, 2000), which performs well at low levels of population differentiation (Latch *et al.*, 2006). In addition, the efficiency of the markers for assignment was tested.

Differentiation between the four populations was visualized by FCA in GENETIX 4.05.2 (Belkhir *et al.*, 1996) with different numbers of markers (Fig. 1). Differentiation between populations was estimated with Euclidian distances between individuals and was calculated with the first four components of the analysis with sets of 1–37 markers. For a given set, this provided the mean within-population (D_W) and mean between-population (D_B) distances between individuals. The ratio D_B/D_W indicated the efficiency of the set of markers to cluster individuals from the same population. The joint efficiency of markers is shown in Fig. 2 for sets of 1–37 markers, where markers are sorted according to decreasing efficiencies. This revealed the three classes of 10, 18 and 9 markers with large, medium and low efficiency, respectively, and suggested a minimum of 25–28 markers to reach a nearly maximal differentiation. The unsupervised Bayesian approach implemented in STRUCTURE revealed four clusters among 367 individuals. Using 10, 20, 28 or 37 markers, the most likely number of clusters was always four, with high proportions of identity between the STRUCTURE clusters and the rivers (90, 93, 94 and 94%, respectively).

The power of markers to assign individuals to their populations was checked using allelic frequencies in WHICHLOCI (Banks *et al.*, 2003) and a maximum likelihood algorithm in. WHICHLOCI employs an empirical method for determining which combination of loci most likely provides a predefined population assignment power for individuals as well as statistical bounds on the performance of any particular group of loci. The method, run with 200 000 iterations, indicated that a minimal set of 28 markers was needed to assign any individual to its population with 99% confidence (bold in Table I).

In parallel, assignment power was tested after randomly splitting the data set into two equal parts. The first one was used to derive population-specific allele frequencies, and then assignment probabilities to the four populations were estimated for the remaining individuals. The proportions of correct assignment were derived with 100 000 simulations for sets of 10, 20, 28 and 37 markers (Fig. 2). These results indicated strong assignment efficiency with a plateau at 98% for 28 markers or more, but a slightly lower value than that estimated with WHICHLOCI.

The present set of 37 microsatellite markers showed high genetic quality and can be used with confidence to study the genetic diversity and structure of

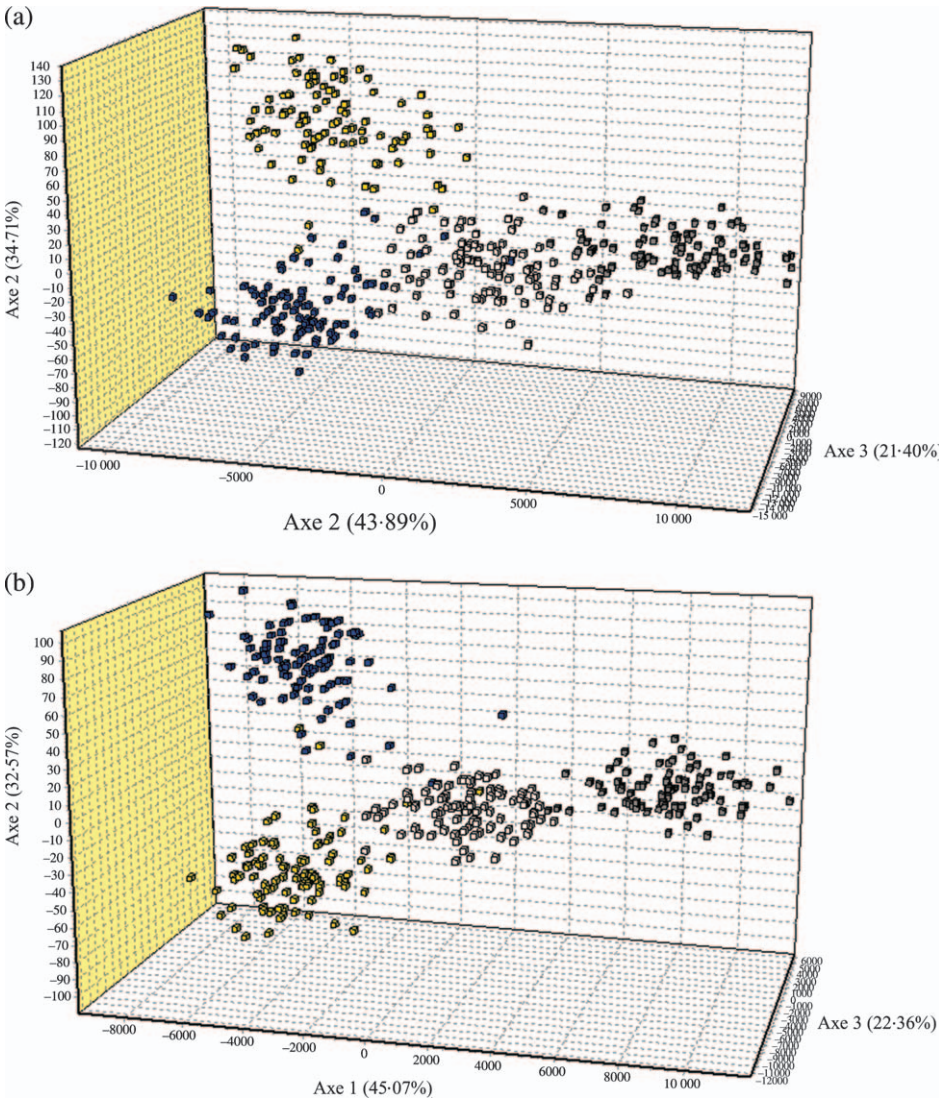


FIG. 1. Factorial correspondence analysis (FCA) in three dimensions of four *Salmo salar* populations: L'Oir (yellow), Scorff (blue), Spey (white) and Shin (grey) with 10 (a), 20 (b), 28 (c) and 37 (d) microsatellite markers (sub-sets of markers as defined in Fig. 2).

S. salar populations. Nevertheless, two markers (*BHMS179A* and *BHMS304/1*) may be less informative, and one (*BHMS328*) should be used with caution. Although the present results are based on samples from only four populations in two European regions (France and Scotland), the large range of alleles detected in the sample suggests that these markers may be informative for other populations. The choice of optimal sub-sets of markers for individual classification or for other questions may depend on the populations under study,

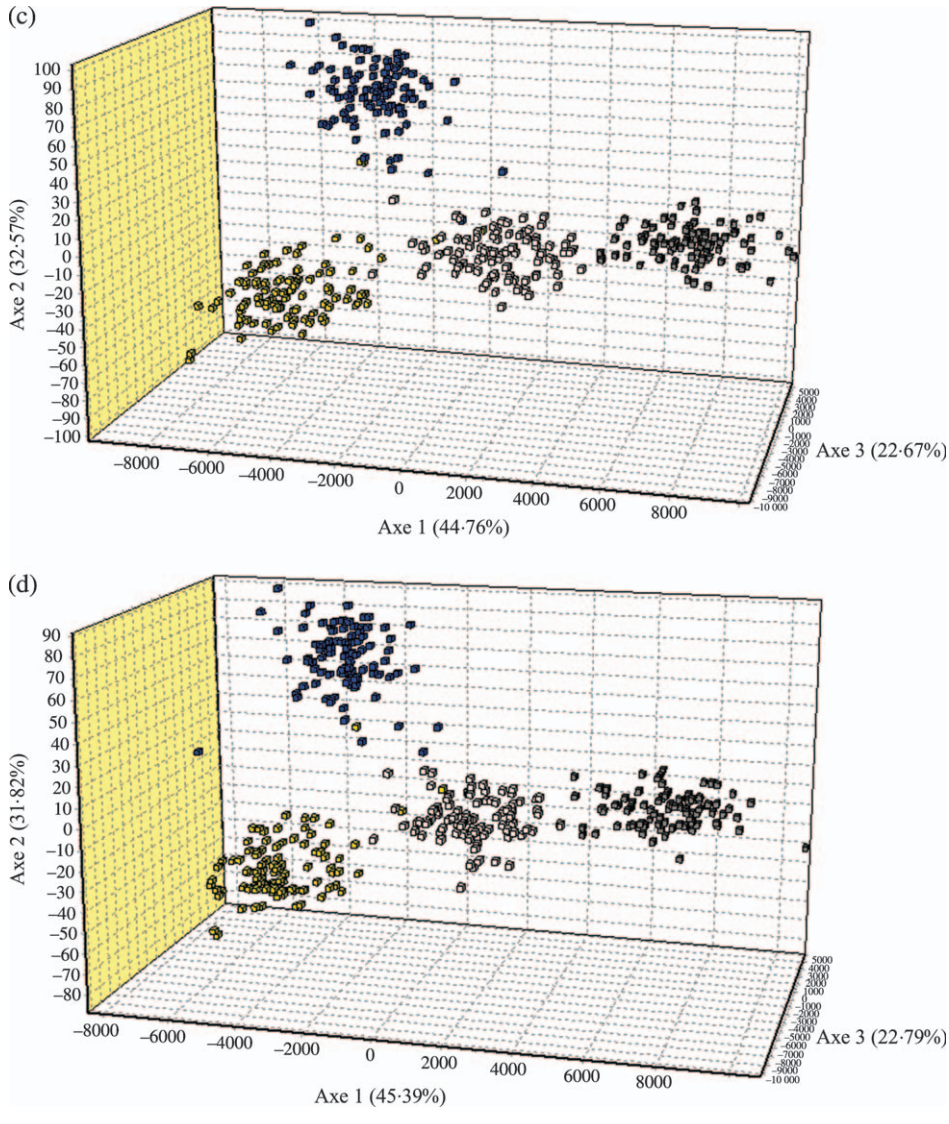


FIG. 1. Continued.

but the number of available markers and their high information content will be useful for many applications.

The authors wish to thank J.-L. Baglinière, N. Jeannot, F. Marchand and D. Azam (INRA Rennes, France), E. Prévost (INRA, St Pée sur Nivelle, France), J. Butler (James Cook University, Townsville, Australia), S. Burns, B. Laughton, J. Woods, J. Reid, D. Ferguson, R. Whyte and S. Grant (Spey Fishery Board Research Office, Scotland, U.K.), D. Knox (Fisheries Research Services, Perthshire, Scotland, U.K.) and I.A.G. McMyn (Kyle of Sutherland District Salmon Fishery Board, Scotland,

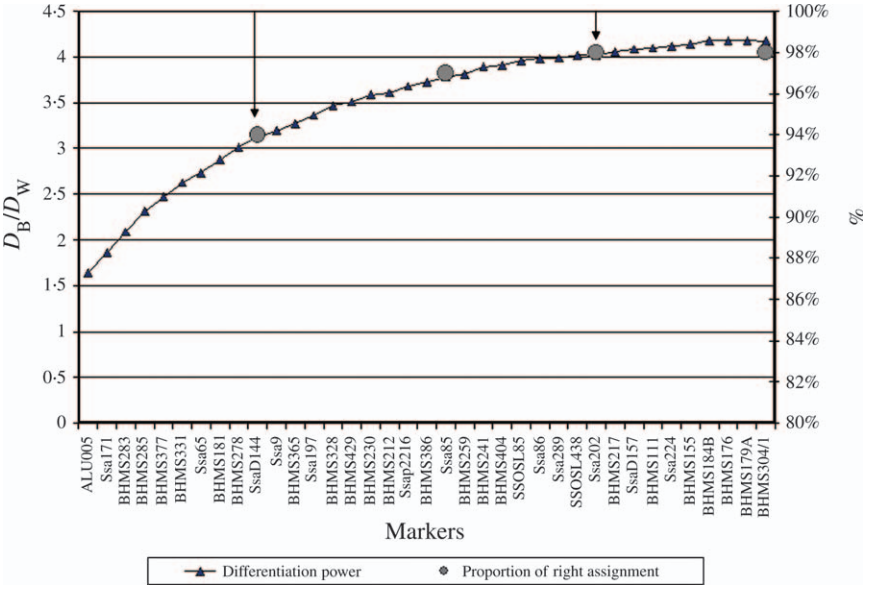


FIG. 2. Joint efficiency of markers for differentiation and assignation. Left scale: differentiation power (D_B/D_W) of sets of 1–37 markers in *Salmo salar*. Solid vertical lines delimit the three classes of markers. Right scale: proportion of right assignment for sets of 10, 20, 28 and 37 markers (likelihood method, using different sub-sets of individuals and repeated random sampling).

U.K.) for providing support in the collection of scale and fin samples. Samples were genotyped at the Toulouse Genopole Platform (http://www.genotoul.fr/).

References

Banks, M. A., Eichert, W & Olsen, J. B. (2003). Which genetic loci have greater population assignment power? *Bioinformatics* **19,** 1436–1438.

Barker, J. S. F., Bradley, D. G., Fries, R., Hill, W. G., Nei, M. & Wayne, R. K. (1993). *An integrated global programme to establish the genetic relationships among the breeds of each domestic animal species*. Rome: FAO Animal Production and Health Paper, Report of a working group.

Beaumont, M. A. & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B* **263,** 1619–1626.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. & Bonhomme, F. (1996). *GENETIX, logiciel sous WindowsTM pour la génétique des populations. Laboratoire Génome, Populations, Interactions CNRS UMR 5000*. Montpellier: Université de Montpellier II.

Crossman, E. J. & Mcallister, D. E. (1986). Zoogeography of freshwater fishes of the Hudson Bay Drainage, Ungava Bay and the Arctic Archipelago. In *Zoogeography of North American Freshwater Fishes* (Hocutt, C. H. & Wiley, E. O., eds), pp. 53–104. New York, NY: Wiley.

Davidson, W. S., Birt, T. P. & Green, J. M. (1989). A review of genetic variation in Atlantic salmon, *Salmo salar* L. and its importance for stock identification, enhancement programmes and aquaculture. *Journal of Fish Biology* **34,** 547–560.

Dieringer, D. & Schlötterer, C. (2003). Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes* **3,** 167–169.

Estoup, A. & Angers, B. (1998). Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. In: *Advances in Molecular Ecology* (G. Carvalho, ed.), pp. 55–86. Amsterdam: IOS Press.

Estoup, A., Largiadèr, C.-R., Perrot, E. & Chourrout, D. (1996). One-tube rapid DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. *Molecular Marine Biology and Biotechnology* **5,** 295–298.

ICES. (2002). Report of the Working Group on North Atlantic Salmon. Copenhagen: International Council for the Exploration of the Sea. *CM 2002/ACFM:* **14**.

King, T. L., Kalinowski, S. T., Schill, W. B., Spidle, A. P. & Lubinski, B. A. (2001). Population structure of Atlantic salmon (*Salmo salar* L.): a range-wide perspective from microsatellite DNA variation. *Molecular Ecology* **10,** 807–821.

King, T. L., Eackles, M. S. & Letcher, B. H. (2005). Microsatellite DNA markers for the study of Atlantic salmon (*Salmo salar*) kinship, population structure, and mixed-fishery analyses. *Molecular Ecology Notes* **5,** 130–132.

Knockaert, C. (2006). *Salmonidés d'aquaculture. De la production à la consommation*.Versailles: Quae Editions.

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*). *Molecular Ecology* **10,** 2525–2539.

Latch, E. K., Dharmarajan, G., Glaubitz, J. C. & Rhodes, O. E. Jr (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics* **7,** 295–302.

Parrish, D. L., Behnke, R. J., Gephard, S. R., McCormick, S. D. & Reeves, G. H. (1998). Why aren't there more Atlantic salmon (*Salmo salar*)? *Canadian Journal of Fisheries and Aquatic Sciences* **55** (Suppl. 1), 281–287.

Pritchard, J. K., Stephens, M. & Donnelly, P. J. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959.

Raymond, M. & Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86,** 248–249.

SanCristobal, M., Chevalet, C., Haley, C. S., Joosten, R., Rattink, A. P., Harlizius, B., Groenen, M. A. M., Amigues, Y., Boscher, M.-Y., Russell, G., Law, A., Davoli, R., Russo, V., Désautés, C., Alderson, L., Fimland, E., Bagga, M., Delgado, J. V., Vegapla, J. L., Martinez, A. M., Ramos, M., Glodek, P., Meyer, J. N., Gandini, G. C., Matassino, D., Plastow, G. S., Siggens, K., Laval, G., Archibald, A. L., Milan, D., Hammond, K. & Cardellino, R. (2006). Genetic diversity within and between European pig breeds using microsatellite markers. *Animal Genetics* **37,** 189–198.

Schuelke, M. (2000). An economic method for the fluorescent labelling of PCR fragments. *Nature Biotechnology* **18,** 233–234.

Ståhl, G. (1987). Genetic population structure of Atlantic salmon. In *Population Genetics and Fishery Management* (Ryman, N. & Utter, F., eds), pp. 121–140. Seattle, WA: University of Washington Press.

Tiira, K., Laurila, A., Peuhkuri, N., Piironen, J., Ranta, E. & Primmer, C. R. (2003). Aggressiveness is associated with genetic diversity in landlocked salmon (*Salmo salar*). *Molecular Ecology* **12,** 2399–2407.

Vasemägi, A., Gross, R., Paaver, T., Kangur, M., Nilsson, J. & Eriksson, L. O. (2001). Identification of the origin of an Atlantic salmon (*Salmo salar* L.) population in a recently recolonized river in the Baltic Sea. *Molecular Ecology* **10,** 2877–2882.

Verspoor, E. (1997). Genetic diversity among Atlantic salmon (*Salmo salar* L.) populations. *ICES Journal of Marine Science* **54,** 965–973.

Withler, R. E., Supernault, K. J. & Miller, K. M. (2005). Genetic variation within and among domesticated Atlantic salmon broodstocks in British Columbia, Canada. *Animal Genetics* **36,** 43–50.

DONNÉES

COMPLÉMENTAIRES

Ce paragraphe du manuscrit présente les résultats complémentaires du premier article de cette thèse intitulé « A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon (*Salmo salar*) populations ».

1.1. Erreur de génotypages

L'ensemble des génotypages a été effectué en prenant des témoins négatifs pour estimer l'erreur de génotypage. Celle-ci s'est révélée faible (2.16%).

1.2. Tétraploïdie

Au cours des lectures des 13 579 profils génotypiques, j'ai pu observer que certains profils présentaient un nombre d'allèles supérieurs à 2. Avec 3 ou 4 allèles, ces profils tri- ou tétra-alléliques pourraient provenir d'une simple contamination. Cependant, lorsque cette polyploïdie se retrouve sur un grand nombre d'individus pour un locus donné, on peut supposer que, dans cette partie d'ADN amplifié, la polyploïdisation ancestrale (duplication du génome, voir chapitre I) qui s'est produite chez les salmonidés, est toujours observable. Lorsqu'on regarde les profils des marqueurs Alu005 (Figure III-1.a) et SSA0048NVH (Figure III-1.b), on peut distinguer deux zones d'amplifications, dont l'une a une intensité plus faible. Il semblerait que nous ayons amplifié deux zones pour ces locus. L'intensité plus faible pourrait s'expliquer par des conditions d'hybridation moins strictes et/ou des mutations au niveau de l'amorce. Nous avons pu également identifier une polyploïdie sur plusieurs locus chez un même individu (cas de 4 individus de la rivière Spey et 1 individu de la rivière Scorff). Il pourrait s'agir d'individus d'élevages (ou d'hybrides) dans lesquels des modifications génétiques ont été pratiquées pour obtenir une plus grande croissance (l'énergie perdue pour les gonades est allouée à la croissance) en leur conférant un profil triploïde, les rendant ainsi stériles. Cependant, les résultats obtenus sur nos profils ne nous permettent pas de conclure sur ces hypothèses. Ces observations pourraient être seulement le fruit de contamination ou d'artefacts, lors de l'amplification. Un séquençage de ces profils est nécessaire pour valider ou invalider ces hypothèses.

a) *Marqueur Alu005*

b) *Marqueur SSA0048/1NVH*

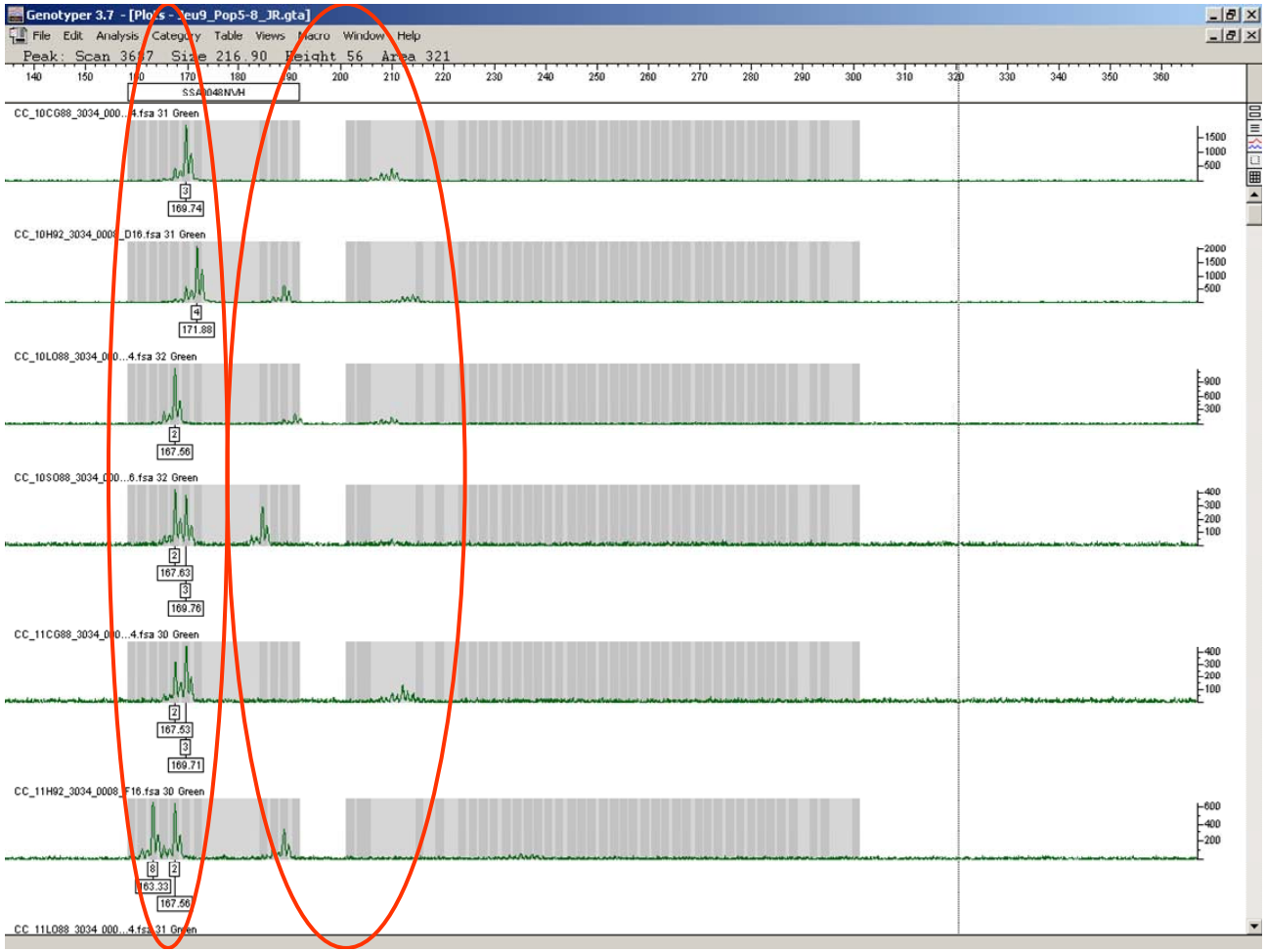


Figure III-1. *Présence de plusieurs allèles pour le marqueur Alu005 (a) et SSA0048/1NVH (BHMS304/1) (b) avec une distinction de deux zones d'amplification pour chacun.*

1.3. Hardy-Weinberg

Lorsque nous continuons notre analyse sur le panel de 37 marqueurs microsatellites, nous pouvons voir des différences par marqueurs et par populations. Comme mentionné dans le premier article de cette thèse, très peu de marqueurs ont été trouvés en déséquilibre Hardy-Weinberg (HW) (P-value<5%) par le logiciel GENEPOP 3.4 (Raymond et Rousset 1995) sur l'ensemble des 37 marqueurs (Table III-1).

Population	Marqueur à l'écart HW	P-value	Total
Oir 2005	SSA0086NVH	0.0071	1
Scorff 2005	SSA0037NVH	0.0081	5
	SSA0057NVH	0.0160	
	SSA0086NVH	0.0376	
	SSOSL85	0.0405	
	SSSP2216	0.0231	
Spey 2005	ALU005	0.0472	5
	SSA0062NVH	0.0322	
	SSA0086NVH	0.0067	
	SSA0152NVH	0.0190	
	SSA85	0.0161	
Shin 2005	SSA0045NVH	0.0057	2
	SSA0086NVH	0.0099	
Oir 1988	ALU005	0.0143	6
	SSA0071NVH	0.0216	
	SSA224	0.0183	
	SSA85	0.0183	
	SSAD157	0.0363	
	SSOSL438	0.0328	
Scorff 1988	SSA0045NVH	0.0002	5
	SSA0086NVH	0.0034	
	SSA0101NVH	0.0295	
	SSA289	0.0313	
	SSSP2216	0.0037	
Spey 1988	SSA0045NVH	0.0025	5
	SSA0086NVH	0.0021	
	SSA0146NVH	0.0247	
	SSA0152NVH	0.0057	
	SSAD157	0.0071	
Shin 1992	SSA0062NVH	0.0072	2
	SSA0086NVH	0.0014	

Table III-1. *Marqueurs qui ne sont pas à l'équilibre Hardy-Weinberg (p<5%) par le logiciel GENEPOP. Les couleurs servent seulement à identifier les mêmes marqueurs dans différentes populations.*

Parmi les marqueurs en déséquilibre HW, le marqueur SSA0086NVH (BHMS328) est en déséquilibre au sein de trois populations, en 2005 et en 1988. Même si le logiciel MICRO-CHECKER laisse sous-entendre que ceci est dû à la présence d'allèles nuls, comme nous le verrons ci-dessous, la lecture délicate des profils (Figure III-2) nous oblige à rester vigilant sur nos conclusions.

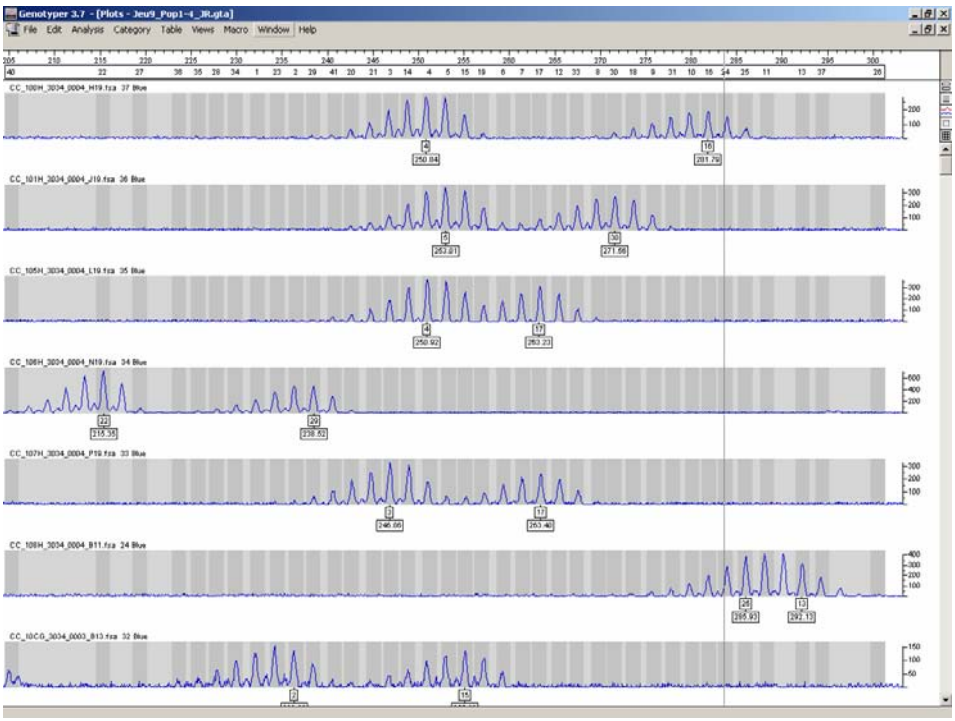


Figure III-2. *Profil de lecture du marqueur SSA0086NVH (BHMS328) par le logiciel Genotyper 3.7 NT (Applied Biosystems).*

Les résultats sur les calculs des allèles nuls, par le logiciel MICRO-CHECKER (Van Oosterhout et al. 2004), laissent apparaître un équilibre HW majoritaire sur l'ensemble de nos populations. Seuls, quelques marqueurs semblent posséder des allèles nuls dans certaines populations:

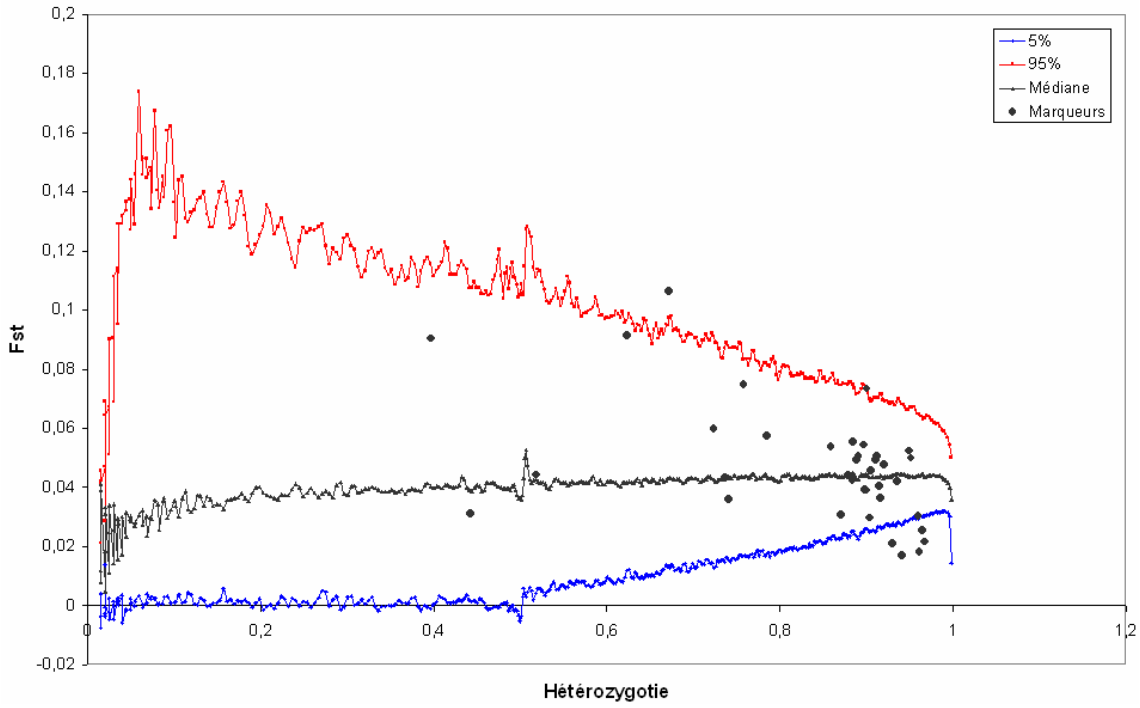
- Oir 2005 (SSA0149NVH) ; Oir 1988 (SSA0037NVH, SSA224).
- Scorff 2005 (SSA0086NVH, SSA202) ; Scorff 1988 (SSA0045NVH, SSA0086NVH, SSA0101NVH, SSsp2216).
- Spey 2005 (Aucun) ; Spey 1988 (Alu005, SSA0045NVH, SSA0065NVH, SSA0086NVH, SSAD157).
- Shin 2005 (SSA0045NVH); Shin 1992 (SSA0062NVH, SSA0086NVH).

1.4. Marqueurs sous sélection

Les locus qui montrent un niveau de différenciation génétique, anormalement bas ou haut, sont souvent supposés être soumis à la sélection naturelle (Beaumont et Nichols, 1996). La différenciation peut être quantifiée en utilisant le F-statistique (*Fst*) de Weir et Cockerham (1993). Pour une gamme de structure et d'histoire démographique des populations, la distribution du *Fst* est fortement liée à l'hétérozygotie à un locus. La méthode FDIST2 (Beaumont et Nichols, 1996) permet l'identification de locus supposés sous sélection.

Comme il est mentionné au sein du premier article, lorsqu'on se place autour de la médiane *Fst* entre 1 et 99%, 4 marqueurs présentent des niveaux de *Fst* trop faibles (Figure III-3b). Lorsqu'on se place à 95%, deux marqueurs supplémentaires (SSA289 et SSA171) se retrouvent avec des valeurs de *Fst* trop élevées (Figure III-3a). Néanmoins, à la lecture des autres analyses, il semblerait que seul le marqueur BHMS328 (SSA0086NVH) pourrait être considéré comme un marqueur non neutre, ce qui nécessiterait de l'utiliser avec précaution**.**

(a)



(b)

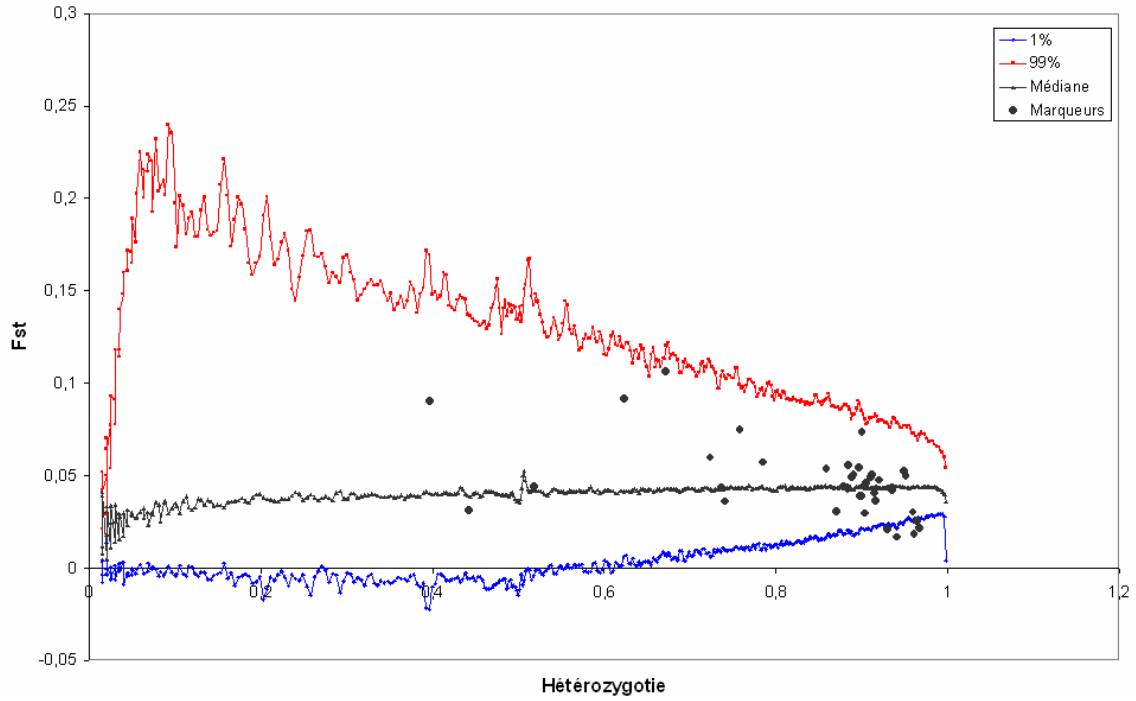


Figure III-3. *Valeurs de Fst attendues en fonction de l'hétérozygotie sur 37 marqueurs microsatellites chez 367 individus de Saumon atlantique (Salmo salar). Les quantiles à 95% (a) et 99% (b) sont délimités par rapport au Fst médian attendu sous un modèle en île.*

1.5. Précisions sur le marqueur SSA202

Les amorces utilisées pour le marqueur SSA202 sont celles d'O'Reilly et al (1996), qui donnent une taille de séquence entre 77-126pb. Les articles mentionnant des tailles supérieures à 200 n'utilisent pas les amorces définies par ces auteurs mais celles modifiées par Withler et al (2005) (Figure III-4).

CACCAACCCTGGTTTTATATATAGCTTGGAATATCTAGAATATGGCTTAGACACAAATCAGGGAACTTGA
CCATTTCTCTGTCCTGTGTGTCCAGGGTCGTGTGGAGGTGGAGTTTGGTAAAGAGGCTCTGAGGTTTGAG
AACGGNCGTTCGCTACTACGGGGTGCCTGCCATCATGCCCACAGGTAACTGGCTCAACTCACACACTCAC
TCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTTCACT
CACTCACTCACTCACGCAACATTAACACATGAATATAGAGAAACAGCCTGTGCACACATTACAA

Figure III-4. *Séquence Fasta du marqueur U43695 (SSA202) avec les amorces modifiées selon Withler et al (2005).*

Partie 2.

ANALYSE DES DONNÉES

ARTICLE 2

An examination of genetic diversity and effective population size in Atlantic salmon

NATACHA NIKOLIC
JAMES BUTLER
JEAN-LUC BAGLINIERE
ROBERT LAUGHTON
IAIN Mc MYNE
CLAUDE CHEVALET

Genetics Research (Soumis en avril 2009)

Résumé

La taille efficace d'une population est un paramètre important en conservation et en gestion des espèces, car elle reflète le niveau de diversité génétique, de consanguinité et de fixation des allèles délétères. Bien que plusieurs études aient démontré l'efficacité des modèles utilisant la généalogie, l'évaluation de leurs limites et de leurs potentialités par des études comparatives, à partir de données réelles, sont trop peu nombreuses.

Dans cette étude, nous avons estimé la sensibilité d'estimateurs basés sur les modèles de coalescence en utilisant des données réelles très différentes, notamment en termes d'effectif réel et efficace. Il s'agit de quatre populations européennes de Saumon atlantique sauvage et anadrome (*Salmo salar L.*) qui sont suivies et intégrées dans des programmes de conservation: l'Oir et le Scorff, dans le Nord-Ouest de la France, ainsi que la Spey et la Shin dans le Nord-Est de l'Écosse. Nous avons débuté notre étude par des analyses classiques en génétique des populations afin de mieux connaître ces populations. Pour cela, nous avons pratiqué des génotypages sur 37 marqueurs ADN microsatellites très polymorphes. Nous avons ensuite estimé leur taille efficace au moyen de trois méthodes de coalescence qui utilisent, pour deux d'entre elles, un échantillon et, pour la troisième, deux échantillons. Les premières analyses suggèrent que les populations proviennent d'un ancêtre commun, démontrant qu'elles possèdent une grande variabilité génétique, un faible flux génique et ayant fait l'objet d'un goulot d'étranglement récent. Pour comprendre la forte diversité génétique au sein des petites populations, nous avons développé un nouveau modèle de prédiction de diversité génétique actuelle, en fonction de l'histoire démographique de la population. Ce modèle suggère que cette forte variabilité ne peut être dû au faible flux génique observé mais parce que ces populations ont diminuée très récemment.

Pour ce qui concerne les analyses comparatives sur les tailles efficaces, celles-ci confirment la sensibilité des modèles vis-à-vis du nombre de marqueurs utilisés, de leur polymorphisme et des paramètres de l'histoire de vie des populations. Nous avons pu constater que les introductions artificielles avaient un impact sur ces estimateurs qui négligent la migration. Pour finir, nous discutons dans cet article de l'application de ces estimateurs au sein des futurs programmes de conservation du Saumon.

An examination of genetic diversity and effective population size in Atlantic salmon populations.

Natacha Nikolic [a], James R.A. Butler [b], Jean-Luc Baglinière [c], Robert Laughton [d], Iain A.G. McMyn[e] and Claude Chevalet[a]

[a]Laboratoire de Génétique Cellulaire (UMR 444), INRA-ENVT, BP 52627, 31326 Castanet Tolosan Cedex, France.
[b]CSIRO Sustainable Ecosystems, James Cook University, PO Box 12139, Earlville BC, Cairns, QLD 4870, Australia.
[c]INRA-Agrocampus, 65 rue de St Brieuc CS 84215 35042 Rennes, France.
[d]Spey Fishery Board and Spey Research Trust, 1 Nether Borlum Cottage, Knockando, Aberlour, Morayshire, AB38 7SD, U.K.
[e]Kyle of Sutherland District Salmon Fishery Board, c/o Bell Ingram Estate Office Bonar Bridge Sutherland IV24 3EA U.K

Running title:
Genetic diversity and effective size in Atlantic salmon

This paper presents the study of two important parameters in the conservation and management of endangered species, genetic diversity and effective population size (*No*), in wild Atlantic salmon.
We developed a new model predicting present diversity as a function of past demographic history. This model suggested that the high genetic diversity observed in wild anadromous European Atlantic salmon (*Salmo salar* L.) populations with low effective sizes could be explained by a recent bottleneck rather than by the only effect of gene flow.
Comparative studies on real data that gauge the relative accuracy of *No* methods are few and a better understanding of limitations and potentials of *No* estimators is needed. Previous studies demonstrated the efficiency of coalescence models. Using nine subsets from 37 microsatellite DNA markers and four salmon populations of different sizes, we compared three coalescence estimators based on single and dual samples. Comparing *No* estimates confirmed the efficiency of increasing the number and variability of microsatellite markers. This efficiency was more accentuated on small populations except for one method. Analysis with low number of neutral markers presenting uneven distributions of allelic frequencies overestimated short-term *No*. In addition we revealed the incidence of Bayesian priors and of artificial stock enhancement using native and non-native origin. Estimates of *No* are proposed for the four populations, and their applications discussed for conservation and management issues.

Keywords: Coalescence, genetic diversity, effective size, microsatellite markers, *Salmo salar*.

1. Introduction

The concept of effective population size (*No*) was introduced by Wright (1931, 1938), and can be defined as the number of breeding individuals in an idealized population that would show the same amount of variation of allele frequencies under random genetic drift. This parameter is usually smaller than the absolute population size. It is a key parameter in conservation and management because it affects the degree to which a population can respond to selection. *No* influences the rate of loss of genetic diversity, the rate of fixation of deleterious alleles and the efficiency of natural selection at maintaining beneficial alleles (Berthier et al. 2002). If *No* declines too far, the loss of genetic variation resulting from genetic drift may put species or populations at risk of extinction by losing the raw material on which selection can operate.

Demographic data from mating systems or genetic data can be used to infer *No*. Unfortunately demographic data are generally difficult to collect in many wild populations and life-history information is often unavailable with sufficient precision to make a good estimate of *No* (Frankham 1995). Genetic methods are more widely used. They are increasingly being developed by applying polymorphic molecular markers to resolve taxonomic problems and describe evolutionary and demographic history of species and populations. Currently, many different methods are available to infer *No* from genetic data in one or multiple samples. The most widely used estimator consists of measuring the variance of allele frequencies between generations (Tallmon et al. 2004). However, several studies noted that it is often biased towards high values (Luikart et al. 1999; Wang 2001; Berthier et al. 2002). Phylogenetic methods are more efficient than nonphylogenetic methods (Felsenstein 1992), primarily due to additional information provided by the tree structure. Genealogical modelling has greatly facilitated the estimation of demographic and mutational parameters using length variants microsatellite data (Chakraborty and Kimmel 1999; Feldman et al. 1999; King et al. 2000 by Storz and Beaumont 2002). In the past, most studies of genetic variation had used protein electrophoresis (Guyomard 1994). However protein polymorphism has a limited potential for discrimination (Norris et al. 1999).

In this paper we focus on highly variable microsatellite loci due to their power to detect genetic structures within and among populations. Furthermore their assumed neutrality avoids the effects of selection, and allows the standard coalescence model to be used to estimate effective population sizes from the distribution of allelic frequencies. Effects of migration are integrated in some models but make the estimation more complicated. In the framework of the coalescence approach, a tree linking the alleles up to their common ancestor describes the relationships among alleles. A coalescence event appears each time two lineages in the tree join into a common ancestor, and the intervals between such events have a distribution that depends on *No* (Kuhner et al. 1995). The usual approach for estimating this parameter is to compare statistics calculated from empirical datasets with a distribution generated by Monte Carlo simulations of the coalescent process (Balding and Wilson 1998; Storz and Beaumont 2002).

Here we compare estimates of *No* based on the coalescence model, and the performances of such methods based either on a single sample (MSVAR and DiyABC) or on two samples (TM3). Hence, we also compare a likelihood approach (TM3 and MSVAR) and Approximate Bayesian Computation (DiyABC). Comparative studies on real data that gauge the relative performances of *No* methods are few. Salmonids are an ideal species for assessing population structure's influence on *No* estimations. They show a propensity for structuring into

genetically distinct populations due to the combined effects of their rearing habitat and their homing behaviour (Stabell 1984). Here we use four wild anadromous (i.e. adults migrate from the sea to breed in freshwater) European Atlantic salmon populations from north-west France (Rivers Oir and Scorff) and north-east Scotland (Rivers Spey and Shin). Atlantic salmon are subject to many pressures in Europe, including physical barriers to migration, exploitation by net and rod fisheries, pollution, the introduction of non-native salmon stocks, physical degradation of spawning and nursery habitat, and increased marine mortality. During the last 30 years, the decline of wild salmon on both sides of the North Atlantic (Parrish et al. 1998; Jonsson and Jonsson 2004) has affected populations to differing degrees (Hawkins 2000). The four populations studied are pressured by different factors and are therefore subject to varying conservation and management strategies. Because their characteristics are well understood (Baglinière and Champigneulle 1986; Baglinière et al. 2005; Butler 2004; Butler et al. 2008), and they have large differences in abundance, they provide a useful opportunity to design tools for estimating *No*.

For each population, fish from two samples have been genotyped with 37 microsatellite markers chosen for their high genetic quality (Nikolic et al. 2009). Based on a general analysis of their present genetic variability and structure, and on different subsets of markers, this paper provides an overview on the performance and sensitivity of three different estimators of *No* applied to declining populations of different sizes.

2. Materials and methods

2.1 Current status and management

Atlantic salmon provide highly valued ecosystem services. They support rod and net fisheries (e.g. Butler et al. 2009), and provide source stock for an aquaculture industry that produces over a hundred million Atlantic salmon, and whose biomass already exceeds that of wild populations (Gross 1998). However, wild Atlantic salmon are considered an endangered species. As a consequence of contracting range, a decline in abundance and the modification of demographic structure of adult populations (Anonymous 2001, 2003; Caron and Fontaine 2003), the species has been placed on the Red List of threatened species in Europe (Porcher and Baglinière 2001). Recent studies (e.g. McGinnity et al. 2003; Finnegan and Stevens 2008) have shown that the introduction of non-native salmon stocks can result in the alteration of genetic structure of the recipient populations, posing another potential threat to population fitness.

Atlantic salmon conservation and management is complicated by the species' life history. Juvenile salmon spend 1-4 years in freshwater depending on the growth conditions and the latitudinal position of the river (Baglinière 1976). They then migrate to the ocean and return as adults to spawn after 1 year ('grilse') or multiple years ('Multi-sea winter' or MSW) (Klemetsen et al. 2003). The Atlantic salmon is a good example of a highly migratory species with complex spatial and temporal patterns demonstrating significant local adaptations (Taylor 1991), homing behaviour (Hansen and Johnsson 1994; Saglio 1994) and reproductive strategies (Fleming 1996).

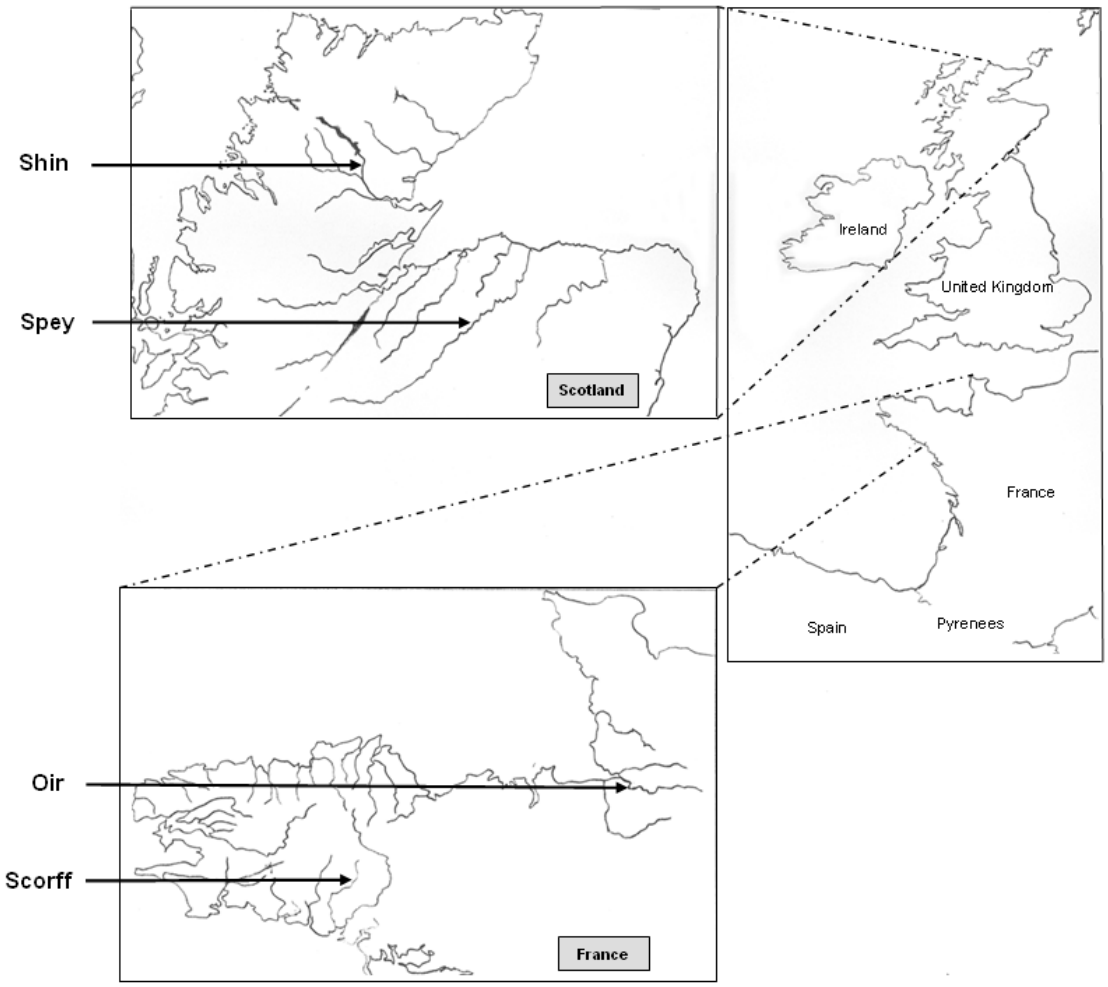


Figure 1. Geographic locations of the four wild Atlantic salmon populations: Rivers Oir and Scorff in north-west France and the Rivers Shin and Spey in north-east Scotland.

Table 1 shows the differences in habitat, population size, fishery pressure and conservation and management regimes for the Oir, Scorff, Spey and Shin. The Scorff, Shin and Spey are coastal catchments while the Oir is an estuarine tributary of the river Sélune and the most productive area for salmon within the Sélune catchment. Salmon in the Spey are little affected by artificial barriers to migration whereas the Oir, Scorff and Shin are more affected. While water quality in the Oir and Scorff is impacted by human pressures (mainly agricultural practices and urban run-off in the estuaries), quality in the Spey and Shin is not acutely affected by anthropogenic activities. The Spey and Shin juvenile populations predominately migrate to sea after 2-4 years, and returning adults include both grilse and MSW fish. In the Oir and Scorff most juveniles migrate after 1 year, and most returning adults are grilse. Exploitation rates by net and rod fisheries are low (10-20%) for the Shin and Spey, but high (30-50%) for the Oir and Scorff. According to the International Union for the Conservation of Nature's (IUCN 1994) classification of conservation status, the Spey and Shin populations are of 'minor concern' while the Oir and Scorff populations are 'vulnerable'.

River	Oir	Scorff	Shin	Spey
Mean annual water temperature fluctuates (°C)	3-19	5-20	1-17	1-20
Drought year	1976. 2003	1976. 2003	1976. 2003	1976. 2003
Catchment area (km²)	87	480	494.6	3000
Streamflow (m³/s)	0.86	5	15.2	65
River length (km)	19.5	75	27.2	172
Slope (‰)	11	3.6	NA	22
Soil type	Schiste and granite	Schiste and granite	Schiste and gneiss	Schiste and gneiss
Recent pH	7.1-7.8	7	6-7	6-9
Recent Nitrate concentrations (mg/l)	High : 30	18.4	12.5	27
Number of major artificial barriers	2	16	2	3
Adult salmon population structure 2005	Grilse (90%)	Grilse (90%)	MSW and grilse	MSW and grilse
Adult salmon population structure 1988 (1992 for Shin)	Grilse	MSW and grilse	MSW and grilse	MSW and grilse
Approximate adult population size 2005	130	1000	3000	<60 000*
Approximate adult population size 1988 (1992 for Shin)	230-260	NA	2000 - 4000	<60 000*
Conservation and management policies	1976 Plan Saumon	1976 Plan Saumon	1986 Salmon Act	1986 Salmon Act, 1999 EU Habitats Directive Special Area of Conservation
Conservation status (IUCN 1994)	Vulnerable	Vulnerable	Minor concern	Minor concern
Coastal net fisheries	Present	Present	Present but declining	Present but negligible effect
River rod fisheries	March-July (extension possible until October)	March-July (extension possible until October)	11 January-30 September	11 February-30 September
Estimated total exploitation rates	30% - 50%	30% - 50%	10-20%	10-20%
Non-native stock enhancement programs	Introduction of Sélune smolts , 1990s	Non-native juveniles of unknown Scottish source 1973-1979	Some fish farm escapees	Some fish farm escapees
Introduction of native salmon stocks	None	None	Native juveniles introduced to mitigate hydroelectricity dam impacts since 1950s.	Native juveniles introduced since 1970s.

Table1. Biotic and abiotic characteristics of the four studied salmon populations and their catchments.
* Spey samples were taken from an upper catchment sub-stock of unknown size, estimated to be less than the minimum total adult population
NA = not available

There is a history of stock enhancement in all of the rivers. Hatchery-reared juvenile salmon of unknown Scottish origin have been planted in the Scorff during 1973-1979 (Baglinière 1979). In the Oir, some returning adults might have originated from hatchery-reared smolts sourced from native Sélune broodstock in the 1990s. In the Spey, native wild broodstock have been used to source ova and juveniles for enhancing production in areas above impassable obstacles since the 1970s. In the Shin, a hatchery program was established to mitigate the effects of the installation of hydro-electricity dams in the 1950s, based on native broodstock. Escaped farmed salmon occasionally enter the Spey and Shin in small numbers.

The Spey supports one of the largest Atlantic salmon populations in Scotland, with at least 60 000 adults entering the river annually (Butler 2004). Because of the relatively pristine nature of the river's habitat, and the status of the salmon population, the Spey was designated a Special Area of Conservation under the EU Habitats Directive (Council Directive 92.43/EEC) in 1999, with Atlantic salmon as a qualifying species. The objectives of SACs are to avoid deterioration of the habitats of qualifying species or significant disturbance to those species, ensuring that the integrity of the site is maintained and it achieves favourable conservation status of the qualifying features (Anonymous 2000).

2.2 Samples and DNA Extraction

A total of 367 wild adult anadromous Atlantic salmon were sampled during the spawning migration in 2005 and 1988 for the Scorff, Oir and Spey, and in 2005 and 1992 for the Shin. For the Scorff, Oir and Shin it was assumed that the samples were representative of the annual adult spawning populations, which are small (100-4000 fish; Tab. 1). However, in the Spey samples were taken from the upper catchment, where spring-running fish are known to originate (Laughton 1991). Larger populations in rivers such as the Spey are known to contain genetically-distinct population units ('sub-stocks') which differ in the timing of their return migration (Stewart et al. 2002; Jordan et al. 2005). Consequently it was assumed that the Spey samples were taken from a sub-stock of spring-running fish of unknown size, but of less than 60 000. In all rivers the individuals came from the same cohort (Tab. 2) to avoid biases in effective population sizes (Jorde and Ryman 1996).

Sample sizes from each river ranged between 89-96 individuals. Pectoral or caudal fin clips were conserved in 95% ethanol, and scale samples were placed and dried in paper envelopes. All samples were kept at ambient temperature. Genomic DNA from fin and scale samples was extracted by boiling samples in 230 µl solution (proteinase K, TE buffer (tris/EDTA) and chelex) at 55°C for several hours, with a final period of 105°C for 15 minutes (Estoup et al. 1996; S. Launey *personal communication*). After one night at 4°C, the supernatant was diluted in 200 µl of chelex and then stored at -20°C.

2.3 Genotyping

In the present study, a set of 37 salmon microsatellite loci from the Salmon Genome project (http://www.salmongenome.no/cgi-bin/sgp.cgi) previously identified by Nikolic et al. (2009) were screened in all the individuals with the M13 labelling method (Schuelke 2000). Polymerase Chain Reaction (PCR) for all 367 individuals was handled in 10 µl within a 384 plate (TECAN 200): 1.5 mM MgCl2, 200 µM dNTPs, 0.1 µM forward primer, 0.15 µM reverse primer, 0.15 µM of M13-Fluo, 25-50 ng DNA and 0.5 U Taq DNA polymerase. Precisions on the primers and amplification conditions for each marker are given in Nikolic et al. (2009). A 2 µl volume of PCR product was added to 8 µl of deionized formamide and the

internal size standard GENESCAN-400HD Rox (Applied Biosystems). Individual genotypes were obtained using ABI 3730 multi-capillary sequencer. Fluorescent DNA fragments size data were labelled by Genescan Analysis Software v3.7 (Applied Biosystems) to assign individuals by Genotyper 3.7 NT software (Applied Biosystems). From 10 random replicates we evaluated genotyping error as relatively common (2.16%). From the initial set of 37 microsatellite markers, subsets of 28, 20, 10 and 5 markers have been selected according to their highest (H+ subsets) or lowest (H- subsets) observed heterozygosity, as estimated from the 367 individuals (Nikolic et al. 2009).

2.4 Genetic diversity and structure

The genetic structure of the populations was assessed from the complete genotypic dataset involving the 37 markers. Genetic diversity parameters were estimated using GENETIX software version 4.05.2 (Belkhir et al. 1998) and GenAlEx 6 software (Peakall and Smouse 2006). Mean number of alleles per locus, actual heterozygosity and unbiased expected heterozygosity were calculated using the Hardy-Weinberg equilibrium (*Nall, Hobs, Hnb and Hexp* respectively). The inbreeding coefficient F_{IS} per population was estimated with 10 000 bootstraps by GENETIX software. Genetic differentiation was estimated with R_{ST} (Slatkin 1995) by GenAlEx software and with F_{ST} according to Weir and Cockerham (1984) by Fstat 2.9.4 software (Goudet 1995). Statistical significance was calculated from 10 000 permutations. Genetic distances (Nei 1978) were derived from allele frequencies. Ninety-five percent confidence intervals of the mean $F_{ST,}$ F_{IS} and F_{IT} estimates were obtained by bootstrapping (1000 replicates) over loci by GENETIX software.

Partition of genetic variance by Euclidean genetic distances among and within populations was calculated according to AMOVA (Analysis of Molecular Variance) resulting in R_{ST} estimations by 9999 permutations. The repartition of individual populations was graphically represented using Factorial Corresponding Analysis (FCA) with GENETIX software. Exact tests of departure from Hardy-Weinberg equilibrium (HWE) within samples were conducted using GENEPOP v3.4 software (http://genepop.curtin.edu.au/index.html, Raymond and Rousset 1995).

Average effective numbers of migrants per generation (*Nm*) were derived using the four salmon populations sampled in 2005 and 1988/1992 by applying the private allele method (Barton and Slatkin 1986) using GENEPOP v3.4 (Raymond and Rousset 1995), and from F_{ST} according to the relationship $Nm=(1- F_{ST})/(4* F_{ST})$ using GENETIX v 4.03. To have a better idea of the migration rate per population, we used IM (Nielsen and Wakeley 2001) between the populations of the same continent with 500 000 burning steps and 5000 000 records period. Genetic assignment was performed using a Bayesian method (Rannala and Mountain 1997) as in GENECLASS 2.0 (Piry et al. 2004), and using the STRUCTURE software (Pritchard et al. 2000) with 20 000 iterations and a burn-in of 5000. All individuals were given a probability of being a resident of each river.

2.5 Genetic diversity under variable population size

We developed a coalescent model to estimate the final genetic diversity when population size has undergone changes in the past called DemoDivMS. Based on the stepwise mutation model, the method allows the present diversity at a microsatellite marker locus to be predicted as a function of present and past population size and of the mutation rate. Analytical calculations provide the expected frequency of pairs of alleles that are alike (i.e. the expected

homozygosity at the locus) and the frequencies of pairs of alleles with any given difference of the numbers of the microsatellite motif (Chevalet and Nikolic, in preparation). The program is available at https://qgp.jouy.inra.fr/.

For the smaller populations (Oir and Scorff), we simulated the evolution of small populations (200-1000 individuals), derived from a larger population (10 000–50 000 effective size) 2000 to 4000 generations ago. We built a scenario in three steps assuming known ancestral and present effective sizes. Between the origin and the present time, we considered the occurrence of a bottleneck and checked various values of the effective sizes before and during the bottleneck, and of the times when population size changed. We assumed the Single Step Mutation model and a mutation rate of 3.10-4 and 9.10-4.

2.6 Estimation of effective population sizes

Present (*No*) and past effective population sizes (*Na*) were evaluated with different available approaches. General features concerning the evolutionary scenarios were derived using the complete set of 37 markers.
Specific estimations of effective population size were run with this complete markers set in order to evaluate the dependency of results on the available genetic information.

2.6.1 Evolutionary scenarios

Past evolution of populations was analysed using four algorithms and the 37 available markers using LAMARC (version 2.1, http://evolution.gs.washington.edu/lamarc, Kuhner 2006), MSVAR (Beaumont 1999), BOTTLENECK (Piry et al. 1999) and DiyABC (Cornuet et al. 2008). The average mutation rate over loci (*u*), ancestral time (*Tf*) and ancestral effective population size (*Na*) were estimated using MSVAR, running 2x10^7 steps per MCMC chain and DiyABC with 500 000 simulations. LAMARC was run only with the set of 37 markers to derive the growth rate (*g*) and provide the global evolution of populations (growing if *g* is positive, shrinking if *g* is negative). The *g* was determined by running 10 initial chains of 1000 steps each, discarding the first 500 and for 2 final chains of 10 000 steps discarding the first 1000. The recent effective size decline was tested using BOTTLENECK with the Two Phase Step Mutation model with 10 000 iterations. Evolutionary scenarios were compared using DiyABC software to derive the most likely outcomes.

2.6.2 Short-term No using the temporal method

In each population a temporal method was used to estimate the harmonic mean of *No* between samples. The method was based on short-term allelic frequency changes between sampling periods (Oir, Scorff and Spey 1988 versus 2005; Shin 1992 versus 2005). Four generations were assumed to have elapsed between the 1988 and 2005 samples, and three generations between the 1992 and 2005 samples. The TM3 temporal method (Berthier et al. 2002) was chosen for its higher efficiency compared to the classical F-statistic estimator (Nei and Tajima 1981; Waples 1989), because it shows a narrower credible interval and greater accuracy when genetic drift is strong (Berthier et al. 2002). The method was run using the 8 subsets of markers and all the 37 markers with 500 000 iterations. Bayesian priors on the maximum size were based on demographic population data.

2.6.3 Long-term No

Two methods, DiyABC and MSVAR were used to assess the long-term effective size from the distribution of alleles. For each population, both samples (1988/1992 and 2005) were analysed separately. DiyABC is an Approximate Bayesian Computation (ABC) which simulates data sets from priors, and then only data sets that are closest to the observed set are retained. The parameter values used to simulate these selected data provide an approximate posterior distribution by local linear regression. A second difference is the simulation of coalescence. Traditional it has been assumed that *No* is large enough to discount the probability that two or more coalescence events occurred in the same generation. However, population size can be very small and multiple coalescence events can occur in the same generation. The alternative is to reconstruct the lineages one generation at a time. DiyABC swaps between these two algorithms according to *No*. The last algorithm is taken when the effective size is very small or when the first algorithm overestimates the number of lineages (Cornuet et al. 2008).

MSVAR version 1.3 was run assuming a linear trend of population size, and setting as prior a starting value compatible with demographic data. We assumed linear evolution because the average changes in population over long periods are more likely to be linear than exponential (Beaumont 1999). As for TM3, a Bayesian prior was set on the maximum size for DiyABC. Also, a *No* of 50 000 (which could represent a maximum ancestral size) was set for all four populations in order to evaluate the influence of prior information on posterior distributions.

3. Results

3.1 Genetic diversity and structure

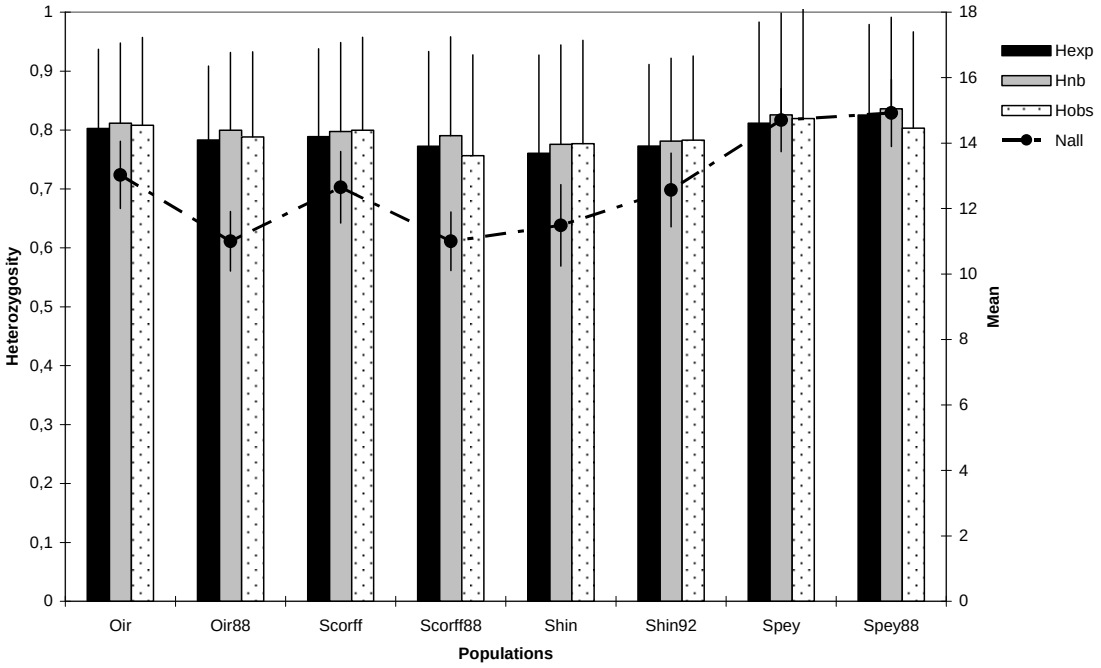


Figure 2. Genetic diversity of the four salmon populations. Expected heterozygosity under equilibrium conditions (*Hexp*), unbiased expected heterozygosity (*Hnb*), observed heterozygosity (*Hobs*) and mean numbers of alleles per locus (*Nall*) for each population sampling in 2005 (Oir, Scorff, Shin and Spey) and previously (Oir88, Scorff88, Spey88 and Shin92) (bars represent ± standard deviation).

Average numbers of alleles per locus (*Nall*) and observed heterozygosity (*Hobs*) across the four Atlantic salmon populations for the two sampling periods (1988/1992 and 2005), ranged from 11.0 (Oir in 1988 and Scorff in 1988) to 14.9 (Spey in 1988), and from 0.76 (Scorff) to 0.82 (Spey) respectively (Fig. 2). *Nall* was stable for Spey, increased for the Oir and Scorff and decreased for the Shin (Fig. 2). Expected heterozygosity under equilibrium conditions (*Hexp*) and its unbiased estimate (*Hnb*) were very close to *Hobs* in all samples. *Hobs* levels across the 37 microsatellites markers were very high with a maximum at 0.93 (SsaD144, BHMS331 and BHMS230) (Nikolic et al. 2009).

Few loci showed a significant deviation from Hardy-Weinberg Equilibrium (HWE) (Tab. 2) with no significant difference between 1988/92 and 2005. These populations seem in Hardy-Weinberg equilibrium.

Inbreeding coefficient (F_{IS}) according to Weir and Cockerham (1984) quantified the difference between *Hobs* and *Hexp* and evaluated the reduction of heterozygosity due to nonpanmictic reproduction. F_{IS} was very low in all populations (Tab. 2) with the highest values for the Shin in 2005 and the Oir in 1988. The genetic differentiations (R_{ST} and F_{ST}) between samples were significant ($p<0.05$) only for the Scorff with the highest values (0.022 and 0.009), indicating that this population underwent a significant genetic change during this period (Appendix A). The lowest values between samples were found for the Shin (0 and 0.001) (Appendix A). Similar trends were observed with *Nei* distance. We have not presented the pairwise R_{ST} in Appendix A because they were of similar magnitude to the F_{ST}. An AMOVA (based on Euclidean R_{ST} distances) across the four populations for the two sampling times revealed that the largest proportion of variation (90%) was found within populations. AMOVA between samples from the same populations showed the highest variance (2%) for the Scorff and lowest (0%) for the Shin, while the Oir and Spey were 1%.

The pairwise *Nm* derived from the private allele method of Barton and Slatkin (1986) were lower than the pairwise from F_{ST} of Weir and Cockerham (1984) (Appendix B) with the highest (8.07) between Spey 1988 and Scorff 1988. The migration rates by IM were at 0.0030 and 0.0038 for Oir, 0.0047 and 0.038 for Scorff, 0.0026 and 0.0025 for Shin and 0.005 and 0.0067 for Spey respectively in 2005 and past samples. Four individuals were assigned outside their population of origin with GENECLASS 2.0: two from the Oir in 2005 were assigned to the Scorff 2005 population, one individual from the Oir in 1988 was assigned to the Spey 1988 population, and one individual from the Scorff in 1988 was assigned to the Spey 1988 population. These four migrants were evident in the Factorial Correspondence Analysis (Fig. 3). All identified migrants were males.

In spite of the low F_{ST} values, the clear structure shown in Fig. 3 was confirmed when using the unsupervised Bayesian approach implemented in STRUCTURE. The set of 367 individuals was clearly split into four clusters. indicating the presence of mixed stock introduction in the Oir 2005 sample and non-native introduction in the 1988 Scorff sample, which was absent in the 2005 sample.

Rivers	**Oir**	**Scorff**	**Shin**	**Spey**	**Oir**	**Scorff**	**Shin**	**Spey**
Names (Fig. 5)	**Oir88**	**Scorff88**	**Shin92**	**Spey88**	**Oir**	**Scorff**	**Shin**	**Spey**
Sampling year	1988	1988	1992	1988	2005	2005	2005	2005
Cohort year	1985	1985	1989	1985	2003	2003	2003	2003
Adults sampled	47	45	48	48	48	48	41	42
FIS 95% interval	0.014 [-0.022-0.024]	0.047 [-0.006-0.058]	-0.005 [-0.038-0.006]	0.038 [-0.003-0.049]	0.003 [-0.033-0.017]	0.001 [-0.035-0.015]	0.005 [-0.033-0.016]	0.001 [-0.030-0.006]
Number of Locus not in HWE	6	5	2	5	5	5	2	4
Mean FIS 95% interval	0.021 [0.005-0.037]				0.001 [-0.012-0.014]			
Mean FST 95% interval	0.055 [0.048-0.063]				0.049 [0.043-0.055]			
Mean FIT 95% interval	0.075 [0.059-0.091]				0.050 [0.038-0.062]			
Nm	3.58				3.43			
Na (a) 95% interval	49015 [8201-296625]	54883 [9230-328417]	47941 [8000-279413]	54780 [9099-330348]	50106 [8510-293430]	64152 [10595-386467]	54682 [9273-321886]	64416 [9912-372382]
Na (b) 95% interval	18186 [4926-35466]	29643 [910-39664]	18776 [798-42136]	14500 [2404-40061]	29995 [6547-35247]	27235 [826-40713]	29995 [942-39172]	13371 [1963-42608]
Tf (a) 95% interval	10038 [1580-65963]	11475 [1852-73502]	7945 [1275-50047]	19374 [2387-196931]	6243 [1014-39214]	10985 [1588-79286]	7717 [1245-49084]	14245 [1982-121566]
Tf (b) 95% interval	12876 [1076-29448]				13652 [1252-31704]			
Growth rate (*g*)	-307	-127	-247	-251	-284	-234	-209	-236
No (a) 95% interval	501 [63-3767]	831 [113-5899]	598 [81-4384]	12503 [1940-80556]	383 [51-2968]	689 [81-5549]	304 [36-2468]	7344 [1076-50364]
No (b) 95% interval	45 [1-485]	1165 [100-1798]	1840 [140-3758]	9596 [958-18052]	100 [12-661]	1174 [102-1791]	1842 [144-3764]	9417 [963-18049]
Populations	**Oir88/2005**	**Scorff88/2005**	**Shin92/2005**	**Spey88/2005**				
No (c) 95% interval	196 [135-283]	212 [152-322]	3424 [1237-4833]	10082 [548-19045]				

Table 2. Demographic and genetic parameters of the four studied salmon populations**.**
(*FIS. FST. FIT*) Wright's F-statistics.
(*Nm*) Number of migrants per generation with the private allele method of Barton and Slatkin (1986).
(*Na*) Estimation of median ancestral population size with the interval at 95%.
(*Tf*) Estimation of time (in years) since population started to decline with the interval at 95%.
(*No*) Estimation of median current population size with the interval at 95%.
(a) MSVAR method; *(b)* DiyABC method; *(c)* TM3 method**.**

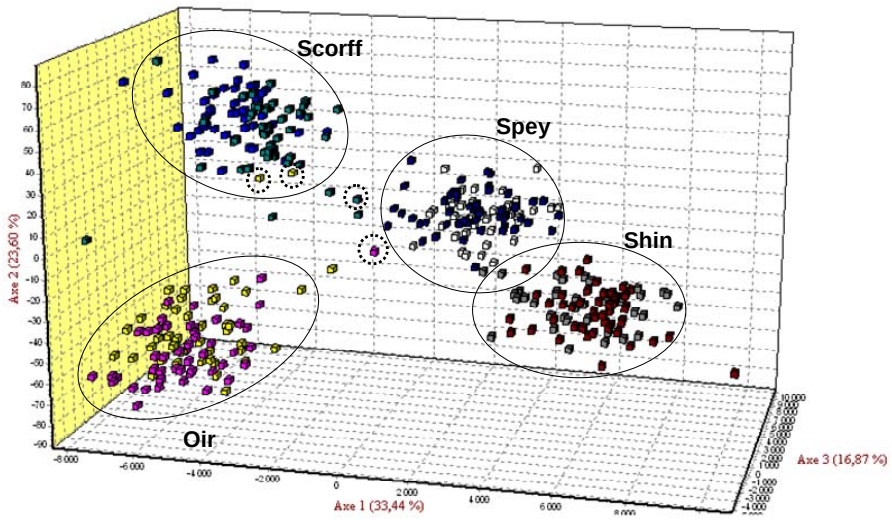


Figure 3. Factorial Correspondence Analysis of 367 wild salmon from the four populations (GENETIX software version 4.05.2, Belkhir et al. 1998). The doped circles represent the migrants found by GENECLASS 2.0 (Piry et al. 2004).

3.3 Genetic diversity under variable population size

Using our calculation of the expected evolution of diversity under variable population size (section 2.5), the high diversity observed in the smaller populations (79-81% for Oir and 76-80% for Scorff) can be qualitatively explained by a recent bottleneck, 25-100 generations ago, assuming an effective size of 2500-5000 before the bottleneck. Even if the results on differentiation and assignment showed that the migration is not strong, we checked its impact on smaller populations from IM values (0.004 and 0.005). Accounting for possible migration was simulated by increasing the mutation rate and suggested an older bottleneck. It also indicated that the high observed genetic variability (>76%) could not be maintained by gene flow in the smallest population (N=200). This suggested that this population derived from a larger one that underwent a recent bottleneck.

3.4 Estimators of effective population sizes

3.4.1 Evolutionary scenarios

Average mutation rates for the 37 microsatellites were estimated at 3.10^{-4} by MSVAR and slightly higher by DiyABC at 9.10^{-4}. The first result is concordant with previous studies on *Salmo salar* by O'Reilly et al. (1998). A negative posterior distribution of *log10(r)* values (with *r* equal *No/Na*) was revealed by MSVAR 0.4 and a negative growth rate (*g*) for overall populations by Lamarc software, which is consistent with a decline in effective sizes (Tab. 2). The various tests of heterozygosity deficit proposed in BOTTLENECK (Sign test and Wilcoxon test, TPM model) suggested a significant departure from constant population size in all populations.

Given an assumed generation time of 3-4 years, the estimated time since decline (*Tf*) ranged from 8000 to 20 000 years ago according to MSVAR 1.3, and around 13 000 years ago according to DiyABC. This was consistent for the four populations and for both sampling times (Tab. 2). The ancestral population size (*Na*) values calculated by MSVAR 1.2 were approximately the same for all populations (approximately 50 000 individuals) suggesting a common ancestor. DiyABC revealed a smaller ancestral population size of 13 000-30 000 (Tab. 2) and two equally likely scenarios in which the populations were separated from a common ancestor by a star or cascade process from the southern (Scorff) to the northern population (Shin).

3.4.2 Markers' number and heterozygosity variation

An increase in the markers' number and polymorphism led to a decrease in the variance of posterior distribution of *No* estimates in almost all methods. While clearly shown for all populations with the MSVAR method, this phenomenon was less visible using TM3 and DiyABC methods for larger populations, and was absent for the Spey, the largest (C and D, Fig. 5 and 6).

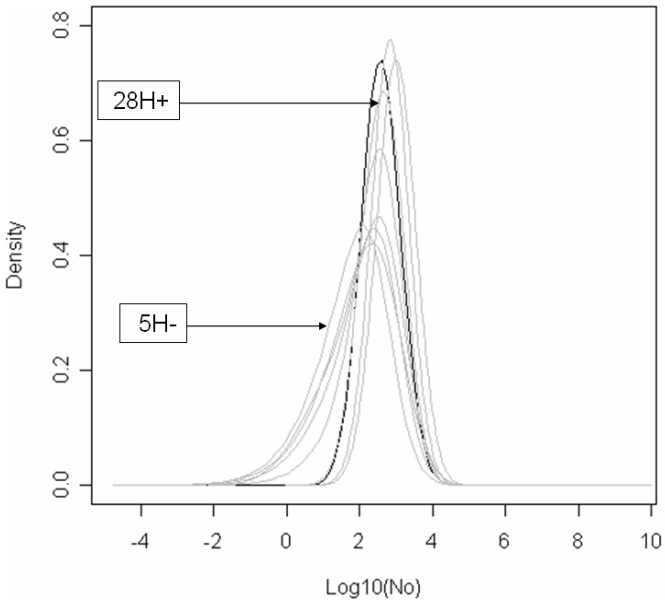


Figure 4. Effects of numbers and polymorphism of markers on the posterior distribution of *Log10(No)*, with *No* the effective population size, for river Oir (2005 sampling), according to MSVAR. Curves in grey refer to the eight subsets of markers, according to their lower (-) or higher (+) heterozygosity (H), (5H-, 5H+, 10H-, 10H+, 20H-, 20H+, 28H-, 28H+). The curve in black refers to the full set of 37 markers. Similar results were obtained with other samples and rivers.

Using MSVAR (Fig. 4) the same trend was observed for the four rivers, showing an increase of mean values and a decrease of variances when the number and polymorphism of markers was increased. In contrast, the DiyABC (Fig. 6) and TM3 (Fig. 5) methods provided unexpected profiles with the subsets of five and ten loci (5mH-, 5mH+, 10mH-, 10mH+) for the Oir and Scorff. With these marker subsets both methods generated large estimates of population size for the Oir (Fig. 5-A and Fig. 6-A). TM3 also generated large estimates for the Scorff population size with the lowest informative subset (5H-) (Fig. 5-B).

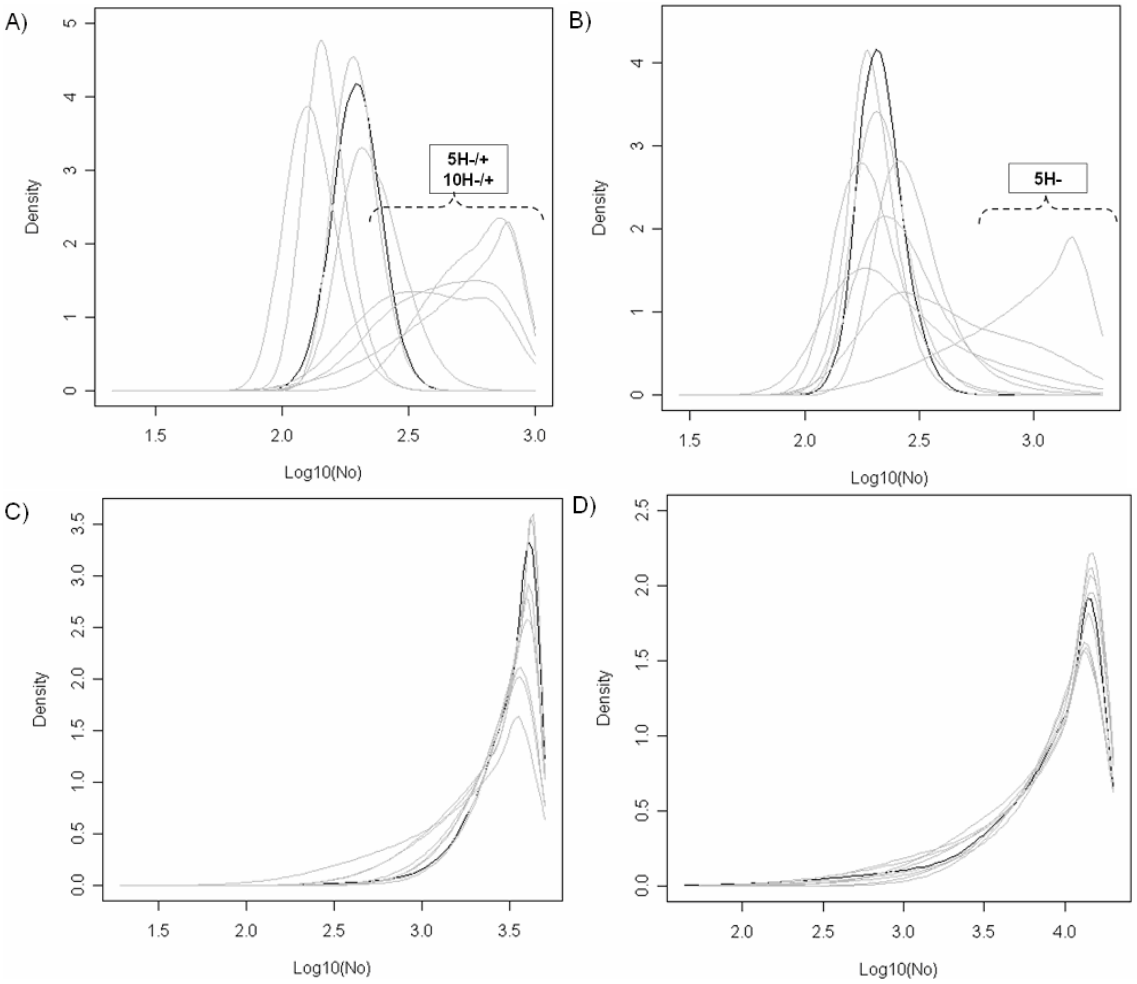
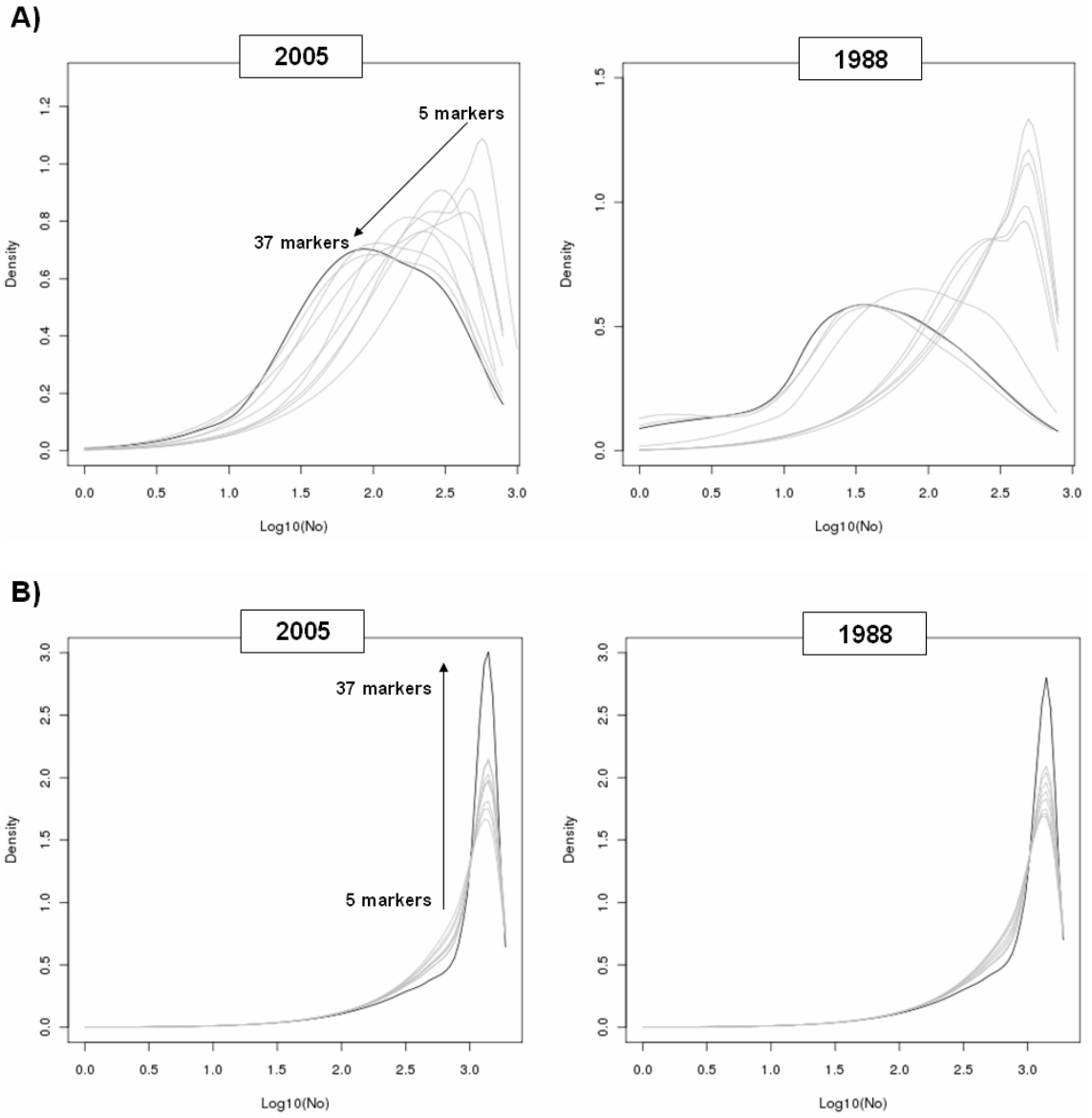


Figure 5. Posterior distribution of *Log10(No)* according to TM3, for the populations Oir (A), Scorff (B), Shin (C) and Spey (D). Curves in grey refer to the eight subsets of markers (5H-, 5H+, 10H-, 10H+, 20H-, 20H+, 28H-, 28H+) and the curve in black to the full set of 37 markers.

3.4.3 Estimation when setting priors on the mean No

Short-term *No* estimates using TM3 were more accurate for smaller populations (≤1000) (A and B, Fig. 7) than for larger populations (≥3000) (C and D, Fig. 7). Long-term *No* estimates using both MSVAR and DiyABC were very consistent between sampling years, but their variances and 95% Confidence Intervals remained high for all samples. For example, the Spey estimate derived by MVAR was 1076-50 364. Overall, *No* for the four populations in 2005 were estimated to be 383 (Oir), 689 (Scorff), 304 (Shin) and 7344 (Spey) using MVAR, and 100 (Oir), 1174 (Scorff), 1842 (Shin) and 9417 (Spey) using DiyABC (Tab. 2).

No common trend was found regarding the effect of stock enhancement. TM3 seemed to underestimate the Scorff population size (B, Fig. 7) and MSVAR to overestimate the Shin population size (C, Fig. 7).


A)
2005
1988
5 markers
37 markers
Density
Log10(No)
B)
2005
1988
37 markers
5 markers
Density
Log10(No)

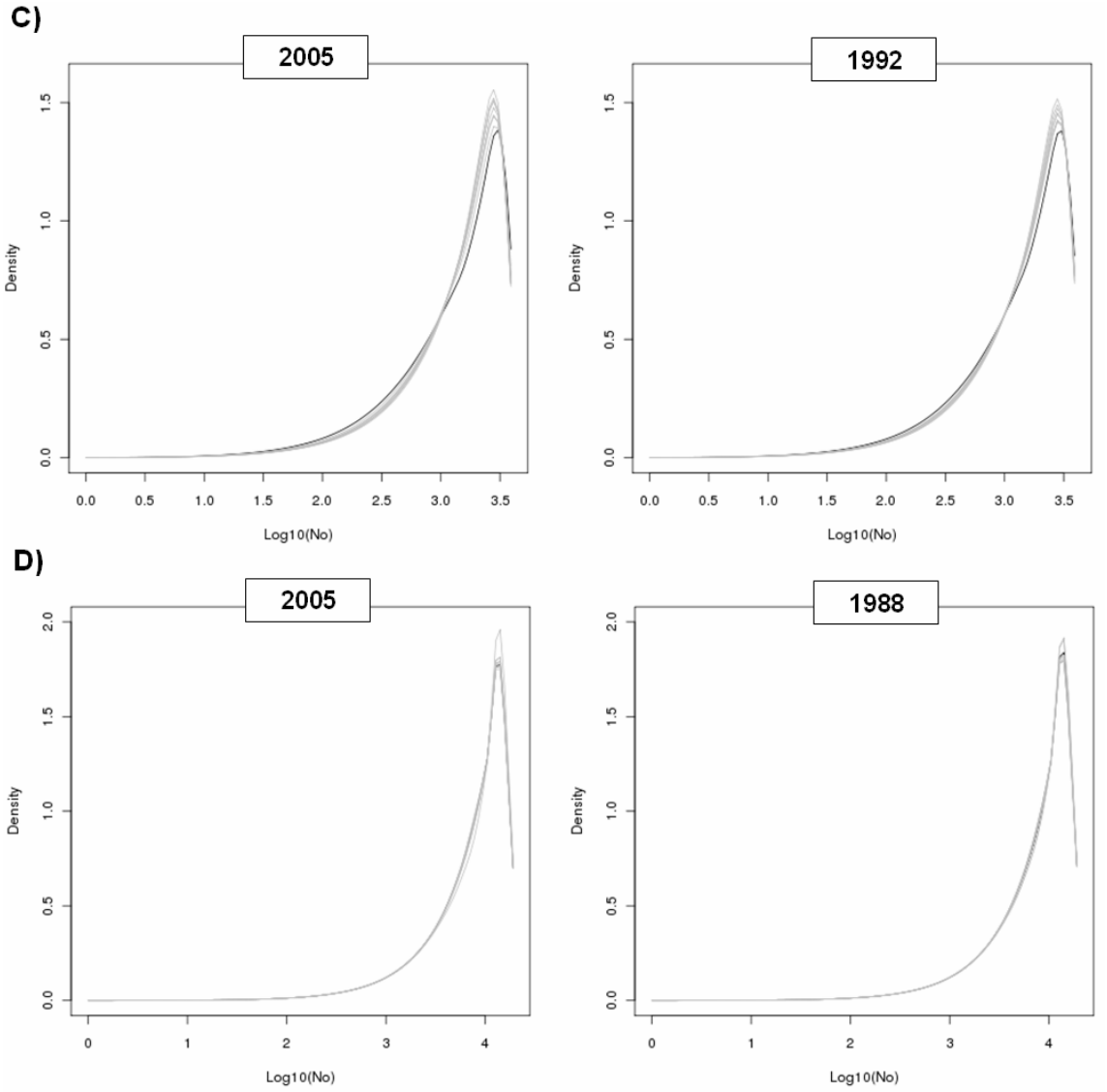


Figure 6. Posterior distributions of *Log10(No)* according to the DiyABC method for the populations Oir (A), Scorff (B), Shin (C) and Spey (D) in 2005 and 1988/1992. Curves in grey refer to the eight subsets of markers (5H-, 5H+, 10H-, 10H+, 20H-, 20H+, 28H-, 28H+) and the curve in black to the full set of 37 markers.

3.4.4 Estimation when setting priors on the maximum No

Changing the prior maximum size to 50 000 had an effect on point estimates of *No* so that for all populations when running DiyABC, mean values were increased and the upper 95% limits of Confidence Intervals approached this maximum. On the contrary no effect was seen with TM3 for the smaller populations (Oir and Scorff). Setting the starting value at 50 000 using MSVAR had no appreciable effect on point estimates of *No* in any population.

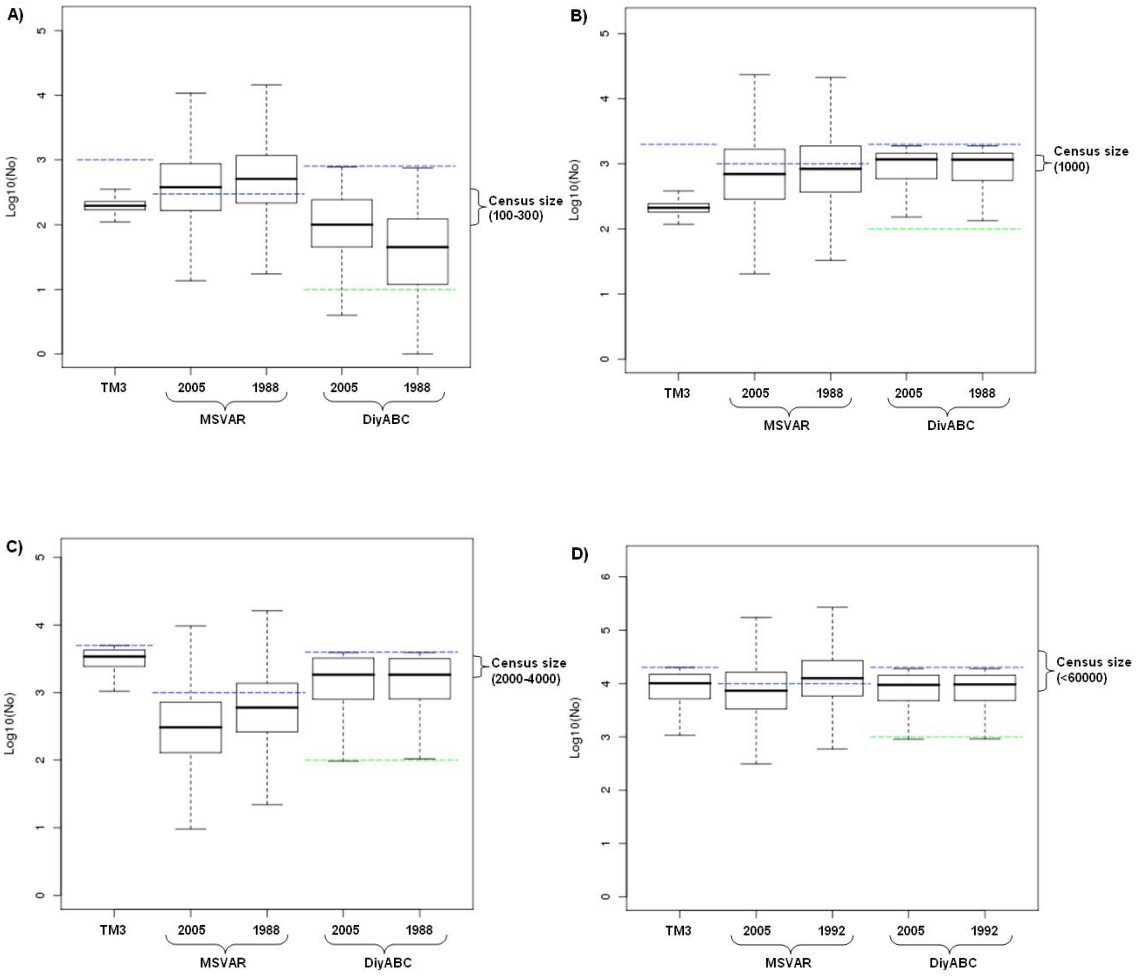


Figure 7. Summary statistics of effective population size estimators (*Log10* scale) from the different methods (TM3, MSVAR, DiyABC) for the four populations: Oir (A), Scorff (B), Shin (C) and Spey (D).
Census size is mentioned on the right y-axis.
Median (horizontal black line), variance (box) and CI (confidence interval) at 95% (dotted black line on both sides of boxes) are given for each method.
Priors are plotted as follows:
Blue dotted lines: highest *No* value for TM3 and DiyABC, starting value for MSVAR.
Green dotted lines: lowest *No* value for DiyABC.

4. Discussion

Measurements of effective population size (*No*) are of importance in conservation and management because they give an overview on the evolution of genetic diversity. Effective size determines the rate at which genetic diversity is lost in the population by genetic drift (Franklin 1980). The most genetically diverse populations are assumed to be 'fitter' (Ligoxygakis 2001). A genetically viable population possesses the evolutionary legacy of the species and the genetic variability on which future evolutionary potential depends (Dobson et al. 1998). Overfishing, blockage of migratory routes by hydroelectric dams and destruction of spawning habitat have severely depleted many wild salmon stocks (Waples 1990). Humans have altered natural ecosystems for many thousands of years, but the magnitude and rate of

these changes have increased dramatically since the industrial revolution. Over recent decades, Atlantic salmon populations have declined or have been extirpated in many parts of its ancestral range (Parrish et al. 1998; Knockaert 2006) and measurements of *No* are few. Genetic estimates of effective size have never been conducted on the Oir, Scorff or Spey populations. Consuegra et al. (2005) investigated *No* in the Shin, but based their estimate on four markers and did not provide estimates. These populations have management regimes and legislation in place to minimize the risks of further declines, but so far effective population size has not been implemented as a tool in their management strategies. Furthermore, the diversity of the rivers' habitats makes them interesting case studies of the effects of contrasting environments on their genetic structure (Whitlock and McCauley 1999; Leberg 2005; Wang 2005).

Before discussing the results obtained about the four salmon populations and their implication for conservation and management, it is useful to recall some features of these populations.
Atlantic salmon populations are highly structured with strong genetic differentiation between regions and among rivers (Ståhl 1987; Nielsen, Hansen and Loeschcke 1999; King et al 2001), with weaker differentiation within river systems (Nielsen et al. 1999; Garant et al. 2000; Primmer et al. 2006; Vaha et al. 2007; Vaha et al. 2008). This differentiation is partly explained by high fidelity to natal rivers (Stabell 1984). Genetic differences observed at the 37 markers reflect the effects of underlying evolutionary forces, such as migration and drift. Within populations the genetic differentiation between samples was significantly higher for the Scorff and lower for the Shin. These differences are explained in the Scorff by the introduction of non-native juveniles of unknown Scottish origin in the 1980s. If the Spey was the source, this last point could explain the highest values of numbers of migrants (*Nm*) between the Spey and Scorff in 1988. The evidence of reduced genetic variability in the Shin and of low difference between samples is probably from the result of long-term artificial stock enhancement program using fish of native (Shin) origin employed to mitigate the blocking of freshwater habitat by hydroelectricity dams in the 1950s. This native stock enhancement may explain the low genetic distance between samples. An effect of stock enhancement in the Oir is suggested by the STRUCTURE analysis. Significant differentiation among the Scorff samples, but not among the Oir samples can be explained by a brief and genetically distant gene flow into the Scorff from a Scottish source, whereas gene flow into the Oir has been more continuous and from a genetically similar French source, the Sélune. For the Spey, gene flow can be negligible given its large *No*. The clear differentiation between rivers and the high rate of assignment of individuals to their river of origin confirm that the populations maintained their originality.
The four populations also seem subject to the recent global decrease in wild salmon stocks. The genetic analysis revealed a negative growth rate and a recent bottleneck for all populations. Whereas allelic richness seemed to decrease in the Shin, increase in the Oir and Scorff and be stable in the Spey, genetic analysis on heterozygosity demonstrated a high and stable variability over time with no inbreeding for all samples. These results are concordant with previous studies that revealed high genetic diversity on wild populations of Atlantic salmon from Ireland and Norway with *Nall* and *Hobs* of 17.8 and 0.70, respectively with 17 microsatellites (Norris et al. 1999) and from Scotland for the Shin population (sampled in 1987-1990) with a *Hobs* of 0.71 with 12 microsatellites (King et al. 2001) and a bit lower with 4 microsatellites (Consuegra et al. 2005). The rate of loss of genetic diversity via genetic drift is higher in small populations and, in the absence of migration, this rate is expected to rise as the effective size decreases (Frankham et al., 2002). Regarding the high genetic diversity (75-81%) detected in smaller populations (Oir and Scorff), we used our model to evaluate the period when the bottleneck occurred. Our results suggest a common ancestor

with a large size (10 000–50 000 effective individuals) dating back to the last glaciations, representing 2000 to 4000 generations. We assumed a scenario with an ancestral large population at great genetic diversity in equilibrium that split into several populations. Trials with our method (using a mutation rate of 3.10^{-4} to 9.10^{-4} and the Single Step Mutation model for microsatellites) suggested that observed small populations (200-1000 effective individuals) could show the high observed heterozygosity if they derived from larger populations of 2500-5000 that experienced a bottleneck 25-100 generations ago (i.e. 100-400 years). Another scenario could be a gradual reduction of effective population size since divergence. Adding migration suggested an older bottleneck. In the majority of salmonid populations, small populations are expected to have higher heterozygosity if open rather than isolated (Palstra and Ruzzante 2008). Even if the migration contributes to diversity, our qualitative approach confirms that the gene flow in population Oir could not maintain the high observed diversity.

Given the importance of identifying *No* for populations in decline, we compared recent coalescent methods to assess their sensitivity to the number of polymorphic markers, to the priors used in Bayesian approaches and to the structure of populations. Using different subsets of markers, we showed that methods providing estimates of *No* may be more or less efficient. The efficiency of estimators is improved when the number of markers and their variability are increased. We chose markers which were independent, but violation of the independence among markers' loci should not affect the accuracy of estimation of *No* (Wang and Whitlock 2003). The Credible Intervals (CI's) and variance decrease by increasing the number and variability of markers with the three methods. However this behaviour depended on populations, except for MSVAR. MSVAR estimations improved accuracy by increasing the number of polymorphic markers for all studied populations (Figure 2). The difference of accuracy between sets of markers (from low number and variability markers to high number and variability markers) was less accentuated for larger populations with TM3 and DiyABC (Figure 5 and 6). These methods presented more complex behaviour. They overestimate *No* in smaller populations (Oir and Scorff) when using few markers. It appears that the used marker sets corresponded to uneven distributions of allelic frequencies. The sets corresponded to five and ten markers at low (H-) and high (H+) heterozygosity. The sets at low heterozygosity possess alleles at high frequencies around 60-80%. The sets at high heterozygosity (H+) did not show alleles with high frequency (>50%), but show uneven distributions between samples from the same population. The absence of alleles at intermediate frequency and high frequency of a single allele may lead to an overestimation of *No* (Waples 1989). MSVAR did not seem to be impacted by the uneven distribution of allele frequencies in the smaller populations.

The incidence of Bayesian priors was weak using MSVAR, but quite strong using DiyABC for all populations, and TM3 for the largest populations (Spey and Shin). Hence, at least for these practical methods, Bayesian priors should be applied using reliable field data to avoid the derivation of inaccurate predictions.

Results from the set of 37 microsatellite markers demonstrated different sensitivity to the size of populations. The CI's and variance were greater for the larger populations (Shin and Spey) using TM3 (short-term *No*), large and homogeneous on all populations using MSVAR (long-term *No*) and heterogeneous using DiyABC (long-term *No*). The long-term *No* estimates were strongly correlated between samples from the same population but CI's were larger for MSVAR than DiyABC except for the smallest population (Oir). Additional analysis using the classical temporal approach based on F-statistics (see Equation (9) in Waples 1989) from

NeEstimator v 1.3 (Peel et al. 2004) demonstrated a high uncertainty in larger populations. The imprecision of temporal methods in large populations seems to occur with classical and coalescent models due to weak genetic drift. Berthier et al. (2002) compared TM3 with the classical F-statistical-based *No* estimator and showed narrower CI's and greater accuracy with TM3 when genetic drift was strong, but there was no improvement with weak genetic drift. Large CIs and variance observed for the three methods can be explained by the ratio of sample size (*S*) to *No*. Too small a sample size may lead to insufficient extraction of genetic information in large populations. As pointed out by Nei and Tajima (1981), precision increases with the ratio *S/No*. Waples (1989) demonstrated that a low ratio would translate into larger CIs, which means that populations with small *No* are most effectively studied. Furthermore, the effects of genetic drift in large populations might be swamped by sampling error using the temporal method, and if $tS/No > \sqrt{2}$, with *t* the sampling interval, increasing the number of loci (or alleles) has a greater effect on precision than increasing *t* or *S* (Waples 1989). This means that given the effective size to be estimated, minimum numbers of generations and sample sizes are required to achieve reasonable precision, unless a large number of loci can be surveyed.

In most studies, coalescent-based long-term estimates of *No* were of the order of two to ten times higher than short-term temporal *No* (Fraser et al. 2007); in this study this was not the case for the larger populations (Shin and Spey). The long-term methods (MSVAR and DiyABC) assumed that selection and migration were unimportant in changing population allelic frequencies relative to genetic drift. For the short-term *No* estimates (TM3) all systematic forces (mutation, selection and migration) are assumed to be absent. The assumption of no mutation is reasonable because our sampling periods of 3-4 generations are sufficiently short to discount the effects of mutations. It may be reasonable to neglect the effects of selection because selection on most markers is unlikely to be strong enough to cause substantial changes in their frequencies (Wang and Whitlock 2003). Moreover, the loci used in our study have been tested to detect potential selection (Nikolic et al. 2009) and can be considered a reliable panel. In the context of this study, even if stock enhancement programs (based on non-native or native source stocks in Scorff and Shin, respectively) were associated with low values of *Nm*, they affected the genetic differentiation between samples, with stronger *Fst*, *Rst* and *Nei* distances in the Scorff and lower in the Shin. Since temporal estimators (TM3) interpreted allele frequency changes as due to genetic drift, they led to decreased estimated effective size in the Scorff, yielding a lower value than the long-term *No*. The effective size estimates derived from short-term methods on the Shin had a median higher than with the long-term methods because native reintroductions helped reduce the genetic differences between samples. Wang and Whitlock (2003) revealed using simulations that *No* will be underestimated in the short-term and overestimated in the long-term if migration is ignored. Regarding ours results, we confirmed that migration modified effective size estimators but in different ways depending on the kind of migration. A native migration, or from a genetically close source, will lead to overestimations of *No* with short-term methods; and inversely with non-native migration, or from a genetically distant source. Even with small values of *Nm*, the temporal methods seemed more sensitive than expected to weak gene flow. In cases when artificial or natural gene flow is suspected, we suggest that both short-term and long-term estimates of population size should be evaluated.

In spite of these challenges, the *No* estimates presented here are the first to be proposed for the French and the larger Scottish (Spey) studied populations, and provide potentially useful tool for their management. Efforts to establish lower population size thresholds have suggested *No* =500 to maintain sufficient evolutionary potential (Franklin 1980, Lande 1988,

Franklin and Frankham 1998, Lynch and Lande 1998). Estimating *No* is thus an important tool in the assessment of genetic variability (Mace and Lande 1991). For the Oir, *No* estimates of 383 and 100 compare with a 2005 estimated adult population of 130, suggesting that the river's IUCN (1994) conservation status of 'vulnerable' is still justified. The same is true for the Scorff where *No* estimates are 689 and 1174, compared to a 2005 adult population of 1000. For the Shin, the 2005 adult population of 3000 exceeds the *No* estimates of 304 and 1842. Similarly on the Spey *No* estimates of 7344 and 9417 are probably exceeded by the 2005 population, although the sub-stock sampled is of unknown size but likely to be <60 000 adults. In both cases the IUCN (1994) status of 'minor concern' remains valid. However, the genetic impact of stock enhancement using native fish should now be taken into account. While the effects are most evident for the Shin, future impacts for the Spey and its SAC (Special Areas of Conservation) objectives may have to be considered. For the Oir and Scorff the impact of 30-50% exploitation rates would seem to be of greatest concern if adult populations are to be restored above our *No* estimates.

'The discrepancies between *No* estimators across the salmon populations contrasted here raises the question of which methods provide best estimates of *No*. All coalescent methods are potentially useful, but they may be biased because of their assumptions, particularly regarding migration. A short-term coalescent-based estimate, such as TM3, seems better suited to anadromous salmon populations than long-term estimates if the population is not large, if it does not undergo identified gene flow and if a high number of highly polymorphic markers are available. CI's become wider as *No* values become larger. In the case of the long-term coalescence estimator, with a minimum sample of 50 individuals as recommended by Waples (1989), increasing the number of sampled individuals (*S*) may marginally improve the results because the variance of the estimator of θ *(4No*μ) decays slowly, at a rate proportional to *1/log(S)* (Deonier et al. 2005). Hence, for larger populations such as the Shin (up to 4000 individuals) and Spey (up to 60 000 individuals) the 37 markers may not be sufficient to accurately estimate *No*, and it would make little difference to increase sample sizes. For the short-term estimator, increasing sample sizes may be useful for larger populations. Using simulations, Ovenden et al. (2007) estimated that 2000 sampled individuals are required to get a reliable estimate of an effective size of 8000 (Palstra and Ruzzante 2008). Another way to decrease CI's for short-term estimates (TM3) in the Shin and Spey populations would be to increase the number of generations between samples. Nevertheless, the amount of information is no longer simply proportional to sampling interval *t*, and too large a *t* may decrease the power of the method (Wang and Whitlock 2003). Usually, a short *t* is recommended (two to four generations) but if the migration rate is low, *t* could be larger to increase the estimation precision (Wang and Whitlock 2003). Finally, the temporal method (TM3) analysis is not equally efficient for all population sizes and other methods of estimating effective population size are necessary if *No* is very large such as in the Shin and Spey, and if the influence of migration cannot be ignored. Regardless, the importance of selecting the appropriate tool for estimating *No* is important for the conservation of wild salmon populations.

Conclusion

Despite their small size and declining status, French populations, still show high levels of genetic diversity, similar to those found in larger populations. Our results and coalescent simulations, where populations are assumed to derive from a common ancestor, suggest that a recent bottleneck and not only gene flow can explain the high genetic diversity found in all studied populations. Concerning the sensitivity of methods, the results raise the importance of number and variability of neutral markers, of the Bayesian priors and of the structure of

populations. Large populations require higher numbers of markers for long-term coalescence estimators and larger sample sizes for short-term coalescence estimators. Even in the case of low migration, establishing a conservation program should rely on both short-term and long-term estimates of population size.

Perspectives

Because variability of effective population size is a main factor determining the risk of extinction (Waples 2002) fluctuating population size is an important consideration for evolution and conservation. Some models have been recently developed (Drummont 2005; Heled and Drummond 2008) to calculate the fluctuation of *No* from the most recent ancestor. Although restricted for the moment to sequence data (mitochondrial or viral DNA) and large intervals of time, extension to microsatellite markers could help provide information on more recent history.

Acknowledgements

The authors wish to thank Jeannot N., Marchand F., Evanno G. and Azam D. (INRA Rennes, France); Prévost E. (INRA St Pée sur Nivelle, France); Burns S., Woods J., Reid J., Ferguson D. and Grant S. (Spey Fishery Board Research Office, Scotland); Knox D. and Verspoor E. (Fisheries Research Services Perthshire, U.K) and Stoddart J. (Kyle of Sutherland District Salmon Fishery Board, U.K) for providing support collection of scale and fin samples. Genotyping were done at the Toulouse Genopole Platform (http://www.genotoul.fr/) with the help of Fève K. and Riquet J. that we fully thank.

References

Anonymous, (2000) Revised Guidance Updating Scottish Office Circular No. 6/1995. Nature Conservation: Implementation in Scotland of EC Directives on the Conservation of Natural Habitats and of Wild Flora and Fauna and the Conservation of Wild Birds (The Habitats and Bird Directives). Scottish Executive: Edinburgh, pp96.

Anonymous, (2001) Report of the working group on the Atlantic salmon. ICES CM2001/ACFM 15, pp199.

Anonymous, (2003) Report of the working group on the Atlantic salmon. ICES CM2001/ACFM, 19, pp297.

Baglinière JL, (1976) Etude des populations de Saumon atlantique (Salmo salar L. 1766) en Bretagne-Basse-Normandie. 1 - Caractéristiques des smolts de la Rivière Ellé. Annales d'Hydrobiologie 7: 141-158.

Baglinière JL, (1979) Production de juvéniles de Saumon atlantique (Salmo salar L.) sur quatre affluents du Scorff, rivière de Bretagne-Sud. Annales de Limnologie 3: 347-366.

Baglinière JL, Champigneulle A, (1986) Populations estimates of juvenile Atlantic salmon (Salmo salar) as indices of smolt production in the River Scorff, Brittany. Journal of Fish Biology 29: 467-482.

Baglinière JL, Marchand F, Vauclin V, (2005) Interannual changes in recruitment of Atlantic salmon (Salmo salar) population in the River Oir (Lower Normandy, France): relationship with spawners and in-stream habitat, Journal of Marine Science 62: 695-707.

Barton NH, Slatkin M, (1986) A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity 56: 409-415.

Beaumont MA, (1999) Detecting population expansion and decline using microsatellites. Genetics 153: 2013-2029.

Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F, (1998) GENETIX, Logiciel sous WindowsTM pour la Génétique des Populations, Laboratoire Génome et Populations, CNRS UPR 9060. Université de Montpellier II, Montpellier, France.

Berthier P, Beaumont MA, Cornuet JM, Luikart G, (2002) Likelihood-based estimate of the effective population size using temporal changes in allele frequency: a genealogical approach. Genetics 160: 741–751.

Butler JRA, (2004) Moray Firth Seal Management Plan; a pilot project for managing interactions between seals and salmon in Scotland. Spey Fishery Board: Aberlour, Moray-shire, pp64.

Butler JRA, Middlemas SJ, McKelvey SA, McMyn I, Leyshon B, Walker I, Thompson PM, Boyd IL, Duck C, Armstrong JD, Graham IM, Baxter JM, (2008) The Moray Firth Seal Management Plan: an adaptive framework for balancing the conservation of seals, salmon, fisheries and wildlife tourism in the UK. Aquatic Conservation: Marine and Freshwater Ecosystems 18: 1025-1038.

Butler JRA, Radford A, Riddington G, Laughton R, (2009) Evaluating an ecosystem service provided by Atlantic salmon and other fish species in the River Spey, Scotland: the economic impact of recreational rod fisheries. Fisheries Research 96: 259–266.

Caron F, Fontaine PM, (2003) L'état des stocks de Saumon atlantique au Québec en 2002. Société de la faune et des parcs du Québec. Direction sur la Faune, pp48.

Chakraborty R, Kimmel M, (1999) Statistics of microsatellite loci: estimation of mutation rate and pattern of population expansion. pp. 139-150 in D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford Univ. Press, Oxford, U. K.

Consuegra S, Verspoor E, Knox D, De Leaniz CG (2005) Asymmetric gene flow and the evolutionary maintenance of genetic diversity in small, peripheral Atlantic salmon populations. Conservation Genetics 6: 823–842.

Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ, Guillemaud T, Estoup A, (2008) Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation, Bioinformatics 24(23): 2713-2719.

Deonier RC, Waterman MS, Tavaré S, (2005) Computational genome analysis. An introduction. Mathematics and Statistics. Springer New York: 403-404.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG, (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology Evolution 22: 1185-1192.

Estoup A, Largiadèr C-R, Perrot E, Chourrout D, (1996) One-tube rapid DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. Molecular Marine Biology and Biotechnology 5: 295-298.

Feldman MW, Kumm J, Pritchard J, (1999) Mutation and migration in models of microsatellite evolution. pp98-115 in D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford Univ. Press, Oxford, U. K.

Felsenstein J, (1992) Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. 59: 139-147.

Finnegan AK, Stevens JR, (2008) Assessing the long-term genetic impact of historical stocking events on contemporary populations of Atlantic salmon, Salmo salar. Fisheries Management and Ecology 15(4): 315-326.

Frankham R, Ballou JD, Briscoe DA, (2002) Introduction to Conservation Genetics. Cambridge University Press, Cambridge, UK.

Frankham R, (1995) Conservation genetics. Annu. Rev. Genet. 29: 305-327.

Franklin IR, (1980) Evolutionary change in small populations In: Conservation Biology: An Evolutionary–Ecological Perspective (eds. Soule´ ME, Wilcox BA), pp. 135–149. Sinauer Associates, Sunderland, MA.

Franklin IR, Frankham R, (1998) How large must populations be to retain evolutionary potential? Animal Conservation, 1: 69–73.

Fraser DJ, Hansen MM, Østergaard S, Tessier N, Legault M, Bernatchez L, (2007) Comparative estimation of effective population sizes and temporal gene flow in two contrasting population systems. Molecular Ecology 16: 3866-3889.

Garant D, Dodson JJ, Bernatchez L, (2000) Ecological determinants and temporal stability of the within- river population structure in Atlantic salmon (Salmo salar L.). Molecular Ecology 9: 615–628.

Goudet J, (1995) Fstat version 1.2: a computer program to calculate Fstatistics. Journal of Heredity 86(6): 485-486.

Guyomard R, (1994) La diversité génétique des populations de Saumon atlantique. In "Le Saumon atlantique : Biologie et gestion de la ressource", J. C. Gueguen et P. Prouzet (Eds), IFREMER, Brest, pp141-151.

Gross MR, (1998) One species with two biologies: Atlantic salmon (Salmo salar) in the wild and in aquaculture. Canadian Journal of Fisheries and Aquatic Science 55: 131-144.

Hansen LP, Jonsson B, (1994). Homing of Atlantic salmon: effects of juvenile learning on transplanted post-spawners. Animal Behavior 47: 220-222.

Hawkins AD, (2000) Problems facing salmon in the sea – summing up. The ocean Life of Atlantic Salmon. Environmental and Biological Factors Influencing Survival (ed D. H. Mills), pp211-222. Fishing News Books, Oxford.

Heled J, Drummond AJ, (2008) Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology 8: 289.

IUCN (International Union for Conservation of Nature), (1994) IUCN red list categories. IUCN, Species Survival Commission, land, Switzerland.

Jonsson B, Jonsson N, (2004) Factors affecting marine production of Atlantic salmon (Salmo salar). Canadian Journal of Fisheries and Aquatic Sciences 61: 2369–2383.

Jordan WC, Cross TF, Crozier WW, Ferguson A, Galvin P, Hurrell RH, McGinnity P, Martin SAM, Moffett IJJ, Price DJ, Youngson AF, Verspoor E, (2005) Allozyme variation in Atlantic salmon from the British Isles: associations with geography and the environment. Journal of Zoology 65 (Supplement A): 146-168.

Jorde P, Ryman N, (1996) Demographic genetics of brown trout (Salmo trutta) and estimation of effective population size from temporal change in allele frequencies. Genetics 143: 1369-1381.

King TL, Kalinowski ST, Schill WB, Spidle AP, Lubinski BA, (2001) Population structure of Atlantic salmon (Salmo salar L.): a range-wide perspective from microsatellite DNA variation. Molecular Ecology 10: 807–821.

King JP, Kimmel M, Chakraborty R, (2000) A power analysis of microsatellite-based statistics for inferring past population growth. Mol. Biol. Evol. 17: 1859-1868.

Klemetsen A, Amundsen PA, Dempson JB et al., (2003) Atlantic salmon Salmo salar L., brown trout Salmo trutta L. and Arctic charr Salvelinus alpinus (L.): a review of aspects of their life histories. Ecology of Freshwater Fish 12: 1–59.

Knockaert C, (2006) Salmonidés d'aquaculture. De la production à la consommation. Quae Editions, Versailles, France, pp327.

Kuhner MK, (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters, Bioinformatics applications Notes 22(6): 768–770.

Kuhner MK, Yamato J, Felsenstein J, (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140: 1421-1430.

Lande R, (1988) Genetics and demography in biological conservation. Science, 241: 1455–1460.

Laughton R, (1991) The movements of adult Atlantic salmon (Salmo salar L.) in the River Spey, as determined by radio telemetry during 1988-1989. Scottish Fisheries Research Report 50.

Leberg P, (2005) Genetic approaches for estimating the effective size of populations. Journal of Wildlife Management 69: 1385–1399.

Ligoxygakis P, (2001) Genetic diversity and conservation efforts, Trends in Genetics, Elsevier, 17(8): 442.

Luikart GH, Cornuet JM, Allendorf FW, (1999) Temporal changes in alleles frequencies provide estimators of population bottleneck size. Conserv. Biol. 13: 523-530.

Lynch M, Lande R, (1998) The critical effective size for a genetically secure population. Animal Conservation, 1: 70–72.

Mace GM, Lande R, (1991) Assessing extinction threats – toward a reevaluation of IUCN threatened species categories. Conservation Biology, 5: 148–157.

McGinnity P, Prodohl P, Ferguson A, Hynes R, O'Maoileidigh N, Baker N, Cotter C, O'Hea B, Cooke D, Rogan G, Taggart JB, Cross T, (2003) Fitness reduction and potential extinction of wild populations of Atlantic salmon Salmo salar as a result of interactions with escaped farm salmon. Proceedings of the Royal Society of London B 270: 2443–2450.

Nielsen R, Wakeley J, (2001) Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158: 885-96.

Nikolic N, Chevalet C, (2009) Distribution of coalescence times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology (in preparation).

Nikolic N, Fève K, Chevalet C, Høyheim B, Riquet J, (2009) A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon Salmo salar populations. Journal of Fish Biology 74: 458-466.

Norris AT, Bradley DG, Cunningham EP, (1999) Microsatellite genetic variation between and within farmed Atlanctic salmon (Salmo salar) populations. Aquaculture 180: 247-264.

Nei M, (1972) Genetic distances between populations. American Naturalist 106: 283–92.

Nei M, and F Tajima, (1981) Genetic drift and estimation of effective population size. Genetics 98: 625-640.

Nielsen EE, Hansen MM, Loeschcke V, (1999) Genetic variation in time and space: Microsatellite analysis of extinct and extant populations of Atlantic salmon. Evolution 53: 261–268.

O'Reilly PT, Herbinger C, Wright JM, (1998) Analysis of parentage determination in Atlantic salmon (Salmo salar) using microsatellites. Animal genetics 29: 363-370.

Ovenden JR, Peel D, Street R et al. (2007) The genetic effective and adult census size of an Australian population of tiger prawns (Penaeus esculentus). Molecular Ecology 16: 127–138.

Palstra FP, Ruzzante DE, (2008) Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? Molecular Ecology, 17: 3428-3447.

Parrish DL, Behnke J, Gephard SR, McCormick SD, Reeves GH, (1998) Why aren't there more Atlantic salmon (Salmo salar)? Can. J. Fish. Aquat. Sci. 55(Suppl. 1): 281-287.

Porcher JP, Baglinière JL, (2001) Le Saumon atlantique. In "Atlas des poisons d'eau douce de France", Keith P. et J. Allardi (coords.). Patrimoines Naturels 47: 240-243.

Peel D, Ovenden JR, Peel SL, (2004) NeEstimator: software for estimating effective population size, Version 1.3. Queensland Government, Department of Primary Industries and Fisheries.

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

Piry S, Luikart G, Cornuet J-M, (1999) Bottleneck: a program for detecting recent effective population size reductions from allele data frequencies. Journal of Heredity 90: 502–503.

Primmer CR, Veselov AJ, Zubchenko A, Poututkin A, Bakhmet I, Koskinen MT, (2006) Isolation by distance within a river system: genetic population structuring of Atlantic salmon, Salmo salar, in tributaries of the Varzuga River in northwest Russia. Molecular Ecology 15(3): 653-66.

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

Rannala B, Mountain JL, (1997) Detecting immigration by using multilocus genotypes. Proceedings of the National Academic Sciences USA 94: 9197–9221.

Raymond M, Rousset F, (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Journal of Heredity 86: 248–249.

Saglio P, (1994) Le retour aux sites de frai ou "homing" : les mécanismes chimiosensoriels. in : "Le Saumon atlantique". J.C. Gueguen and P. Prouzet. Ifremer, pp101-122.

Schuelke M, (2000) An economic method for the fluorescent labeling of PCR fragments, Nature Biotechnology 18(2): 233 – 234.

Slatkin M, (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 1463.

Stabell OB, (1984) Homing and olfaction in Salmonids — a critical review with special reference to the Atlantic salmon. Biological Reviews of the Cambridge Philosophical Society 59: 333–388.

Ståhl G, (1987) Genetic population structure of Atlantic salmon. In: N. Ryman & F. M. Utter (Eds.), Population Genetics and Fisheries Management. Seattle: University of Washington Press. pp121–140.

Stewart DC, Smith GW, Youngson AF, (2002) Tributary-specific variation in timing of return of adult Atlantic salmon (Salmo salar) to freshwater has a genetic component. Canadian Journal of Fisheries and Aquatic Science 59: 276-281.

Storz JF, Beaumont MA, (2002) Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. Evolution 56(1): 154-166.

Tallmon DA, Luikart G, Beaumont MA, (2004) Comparative evaluation of a new effective population size estimator based on approximate bayesian computation. Genetics 167: 977-988.

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

Vaha JP, Erkinaro J, Niemela E, Primmer CR, (2007) Life-history and habitat features influence the within-river genetic structure of Atlantic salmon. Molecular Ecology 16: 2638–2654.

Vaha JP, Erkinaro J, Niemela E,, Primmer CR, (2008) Temporally stable genetic structure and low migration in an Atlantic salmon population complex: implications for conservation and management. Evolutionary Applications 1: 137–154.

Wang JL, (2001) A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genetical Research 78: 243-257.

Wang J, Whitlock M.C, (2003) Estimating effective population size and migration rates from genetic samples over space and time. Genetics 163: 429-446.

Wang JL, (2005) Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 360: 1395–1409.

Waples RS, (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121: 379-391.

Waples RS, (1990) Conservation genetics of Pacific salmon. III. Estimating effective population size. Journal of Heredity 81: 277–289.

Waples RS, (2002) Effective size of fluctuating salmon populations. Genetics 161: 783-791.

Weir BS, Cockerham CC, (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–70.

Whitlock MC, McCauley DE, (1999) Indirect measures of gene flow and migration: Fst = 1/(4Nm + 1). Heredity 82: 117–125.

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

Wright S, (1938). Size of population and breeding structure in relation to evolution. Science 87: 430-431.

Supplementary files

Appendix A. Pairwise F_{ST} of Weir and Cockerham (1984) by Fstat software on the superior half matrix (* Statistical significance, p<0.05, from permutations (10 000) test). Nei distance (1978) by GENETIX software on the inferior half matrix.

Sampling	**Years**	**2005**	**1988**	**2005**	**1988**	**2005**	**1992**	**2005**	**1988**
Years	**Populations**	**Oir**	**Oir**	**Scorff**	**Scorff**	**Shin**	**Shin**	**Spey**	**Spey**
2005	**Oir**	0	0.007	0.037*	0.041*	0.068*	0.070*	0.038*	0.037*
1988	**Oir**	0.019	0	0.051*	0.056	0.084*	0.089*	0.050*	0.047*
2005	**Scorff**	0.170	0.219	0	0.009*	0.075*	0.076*	0.045	0.037*
1988	**Scorff**	0.178	0.233	0.031	0	0.075*	0.074*	0.041*	0.032
2005	**Shin**	0.331	0.414	0.368	0.350	0	0.001	0.032*	0.039*
1992	**Shin**	0.333	0.433	0.367	0.331	0.005	0	0.037*	0.042*
2005	**Spey**	0.193	0.245	0.217	0.1772	0.146	0.153	0	0.004
1988	**Spey**	0.189	0.235	0.185	0.135	0.177	0.178	0.023	0

Appendix B. Pairwise numbers of migrant (*Nm*) from F_{ST} of Weir and Cockerham (1984) by GENETIX software on the superior half matrix and number of migrants using private alleles (Barton and Slatkin 1986) by GENEPOP software on the inferior half matrix.

Sampling	**Years**	**2005**	**1988**	**2005**	**1988**	**2005**	**1992**	**2005**	**1988**
Years	**Populations**	**Oir**	**Oir**	**Scorff**	**Scorff**	**Shin**	**Shin**	**Spey**	**Spey**
2005	**Oir**	0		6.51	6.10	3.38	3.39	6.43	6.64
1988	**Oir**		0	4.93	4.68	2.67	2.62	5.08	5.26
2005	**Scorff**	2.68	1.83	0		2.96	2.99	5.38	6.29
1988	**Scorff**	2.46	1.67		0	3.01	3.23	6.47	8.07
2005	**Shin**	1.97	1.25	1.81	1.20	0		7.08	6.01
1992	**Shin**	2.21	1.40	1.93	1.92		0	6.71	5.87
2005	**Spey**	2.79	2.17	2.81	2.09	2.91	3.62	0	
1988	**Spey**	2.83	2.58	3.11	3.10	4.00	3.79		0

DONNÉES

COMPLÉMENTAIRES

Ce paragraphe décrit les résultats complémentaires de l'étude présentée au second article intitulé : « *An examination and genetic diversity of effective population size in Atlantic salmon* » ainsi que les documents relatifs au nouveau modèle que l'on peut retrouver sur le site *https://qgp.jouy.inra.fr/*. Ces résultats complémentaires sont issus d'analyses en génétique des populations sur quatre populations européennes de Saumon atlantique sauvage et anadrome, au moyen de logiciels couramment utilisés et mentionnés dans le Chapitre II de ce mémoire.

2.1. Diversité génétique

La diversité génétique peut être représentée par la richesse allélique ou par le taux d'hétérozygotie. Comme on peut le voir sur les Figures III-5, 6 et 7, cette diversité est homogène par marqueurs sur l'ensemble des quatre populations. Un marqueur, à grande diversité génétique pour une population, le sera également pour les trois autres.

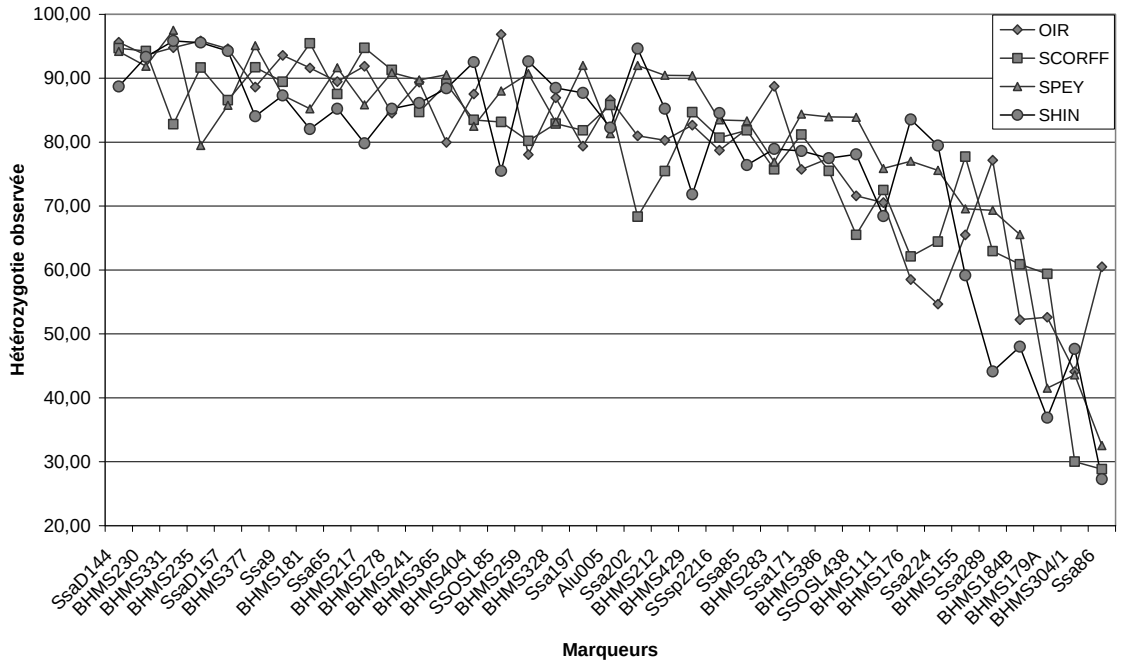


Figure III-5. *Hétérozygotie moyenne observée sur les deux échantillons par population avec un classement en abscisse de l'hétérozygotie observée moyenne sur l'ensemble des populations.*

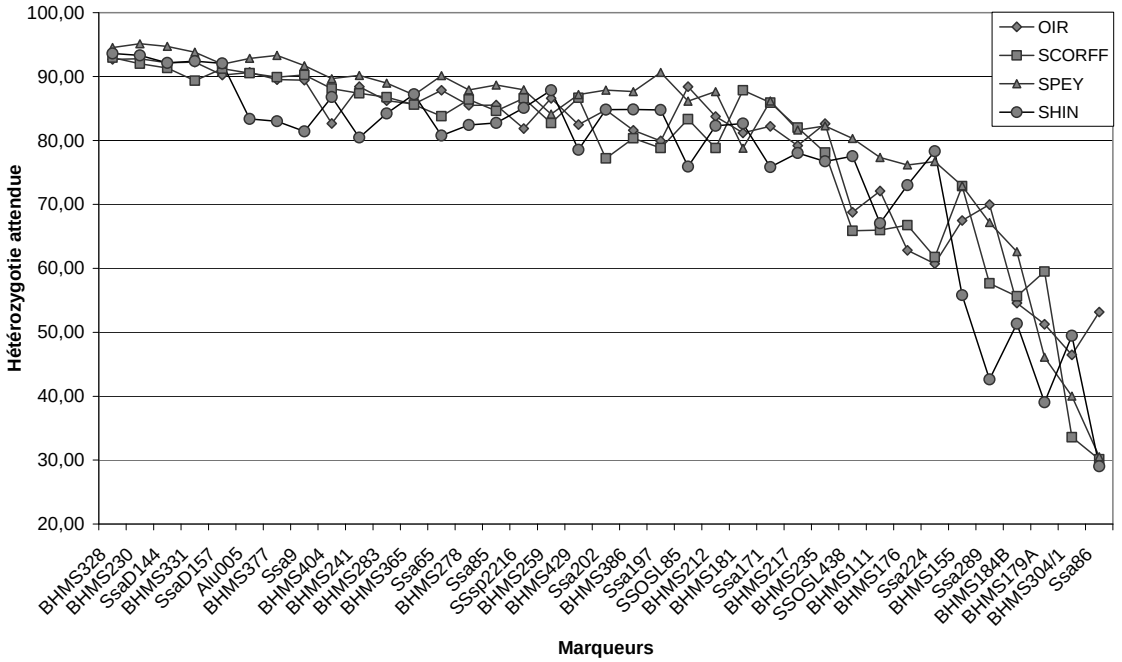


Figure III-6. *Hétérozygotie moyenne attendue sur les deux échantillons par population avec un classement en abscisse de l'hétérozygotie attendue moyenne sur l'ensemble des populations.*

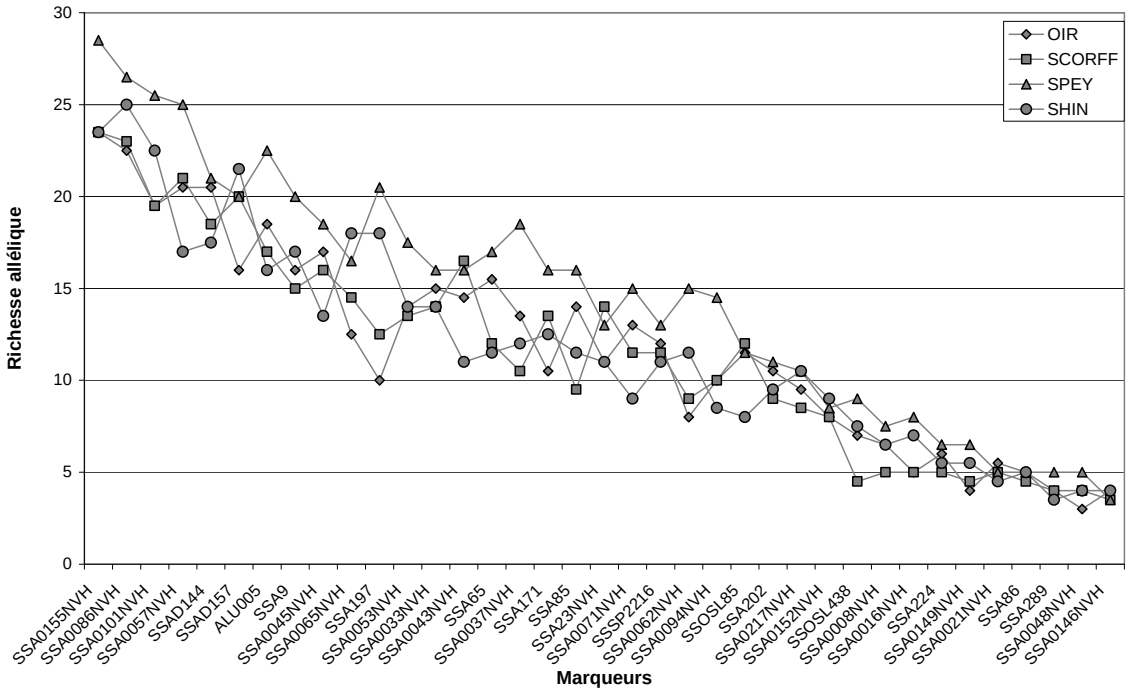


Figure III-7. *Richesse allélique moyenne sur les deux échantillons par population avec un classement en abscisse de la richesse allélique moyenne sur l'ensemble des populations.*

2.2. Distances génétiques

Les estimations de la distance génétique sur les données réelles entre les quatre populations et les deux échantillonnages sont plus élevées avec les calculs de Nei 1972 (a) que les calculs de Nei 1978 (b). Ces différences proviennent de la considération du biais d'échantillonnage par la distance de Nei 1978. Néanmoins, les deux estimations (Table III-2) révèlent des divergences élevées entre les populations de 2005 (21,4-42,2% (a) et 14,6-36,8% (b)) et de 1988 (20,8-49,2% (a) et 13,5-43,3% (b)). Ces divergences sont nettement plus fortes que celles observées entre échantillons de la même population. Cependant ces valeurs qui se situent autour de 5-9% (a) et 0.5-3% (b) ne sont pas négligeables au vu de l'écart en nombre de générations (3-4) des deux échantillons. La population Shin (échantillonnée une génération avant les autres populations) montre une divergence plus faible entre ses échantillons.

	L'Oir	**Scorff**	**Spey**	**Shin**	**L'Oir**	**Scorff**	**Spey**	**Shin**
	2005	**2005**	**2005**	**2005**	**1988**	**1988**	**1988**	**1992**
L'Oir 2005	0	0,214	0,254	0,387	0,081	0,242	0,242	0,374
Scorff 2005	0,170	0	0,276	0,422	0,280	0,093	0,236	0,406
Spey 2005	0,193	0,217	0	0,219	0,323	0,257	0,093	0,211
Shin 2005	0,331	0,368	0,146	0	0,487	0,425	0,241	0,057
L'Oir 1988	0,019	0,219	0,245	0,414	0	0,314	0,305	0,492
Scorff 1988	0,178	0,031	0,177	0,350	0,233	0	0,208	0,391
Spey 1988	0,189	0,185	0,023	0,177	0,235	0,135	0	0,227
Shin 1992	0,333	0,367	0,153	0,005	0,433	0,331	0,178	0

Table III-2. *Distance génétique entre les 4 populations (L'Oir, Scorff, Spey et Shin) échantillonnés en 2005 et 3-4 générations antérieures. La distance de Nei 1972 (matrice supérieure) (calcule les distances génétiques entre populations sans correction de biais) et la distance de Nei 1978 (matrice inférieure) (calcule les distances génétiques en introduisant une correction pour le biais d'échantillonnage d'individus).*

Les résultats des distances génétiques de Nei (1978) entre échantillons ont également été visualisés sous la forme d'un dendrogramme (UPGMA) (Figure II-8) par le package Phylip en regardant la solidité des nœuds sous 1 000 tirages avec remises (bootstrap).

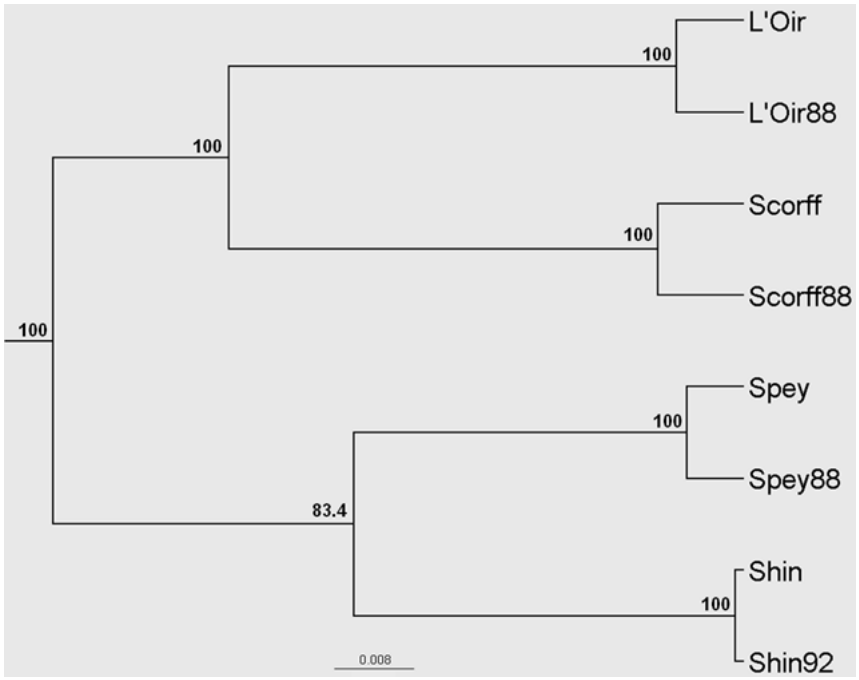


Figure III-8. *Dendrogramme UPGMA (Unweighted Pair Group Method) basé sur les distances de Nei 1978 pour les huit échantillons : 2005 (Oir, Scorff, Spey, Shin), 1988 (Oir88, Scorff88, Spey88) et 1992 (Shin92). Les nombres sur les branches représentent les pourcentages avec 1 000 bootstraps.*

L'isolation par distance a été évaluée par le test de Mantel (GENALEX 6, Figure II-9). La distance de Nei standard (1978) a été superposée à la distance géographique (distance marine la plus faible entre l'embouchure des rivières) pour tester si les migrants potentiels s'étaient déplacés de façon aléatoire ou, préférentiellement, à proximité de leur rivière d'origine. Ce test a révélé une corrélation significative ($R^2 = 0.50$, $p < 0,01$) pour tous les échantillons.

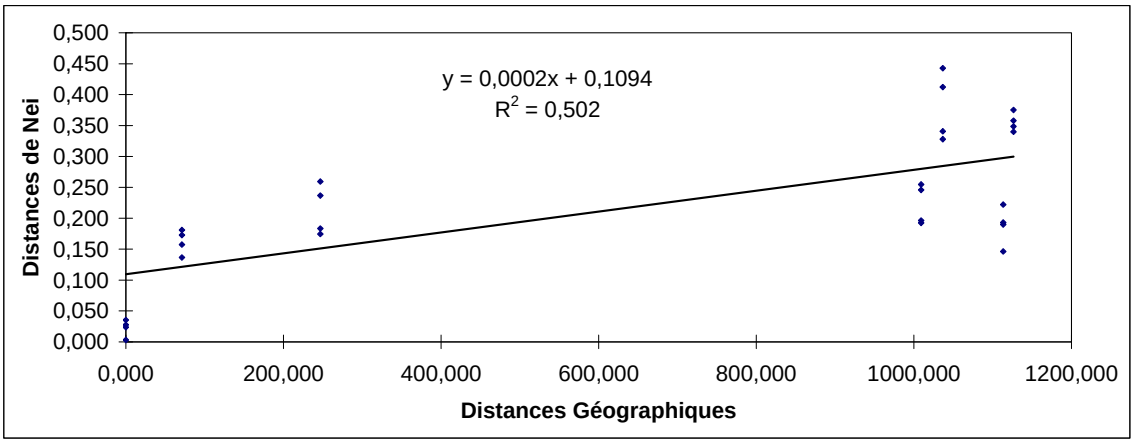


Figure III-9. *Tests de Mantel entre les distances géographiques et génétiques (Nei, 1978) sur les 8 échantillons.*

2.3. F-statistiques

Pour ces analyses de F-statistiques, j'ai préféré enlever le marqueur SSA00217NVH qui avait beaucoup de données manquantes sur les échantillons de 1988 dans la population Oir et Scorff. Les différences sont très légères mais semblent plus cohérentes notamment au niveau des calculs de consanguinité (F_{IS}) par le logiciel GENETIX software version 4.05.2 (Belkhir et al. 1998).

Populations	a) Réel	(IC 95%)	b) Réel	(IC 95%)
Oir 2005	0,00252	(-0,03242 - 0,01572)	0,00275	(-0,03329 - 0,01665)
Scorff 2005	0,00030	(-0,03567 - 0,01420)	0,00075	(-0,03538 - 0,01468)
Spey 2005	0,00880	(-0,03654 - 0,01311)	0,00062	(-0,02974 - 0,00557)
Shin 2005	-0,00361	(-0,04575 - 0,01204)	0,00533	(-0,03319 - 0,01635)
Oir 1988	0,01377	(-0,03080 - 0,01963)	0,01426	(-0,02216 - 0,02367)
Scorff 1988	0,04538	(-0,00451 - 0,05424)	0,04699	(0,00600 - 0,05850)
Spey 1988	0,03747	(0,00179 - 0,04676)	0,03824	(0,00253 - 0,04863)
Shin 1992	-0,00421	(-0,03729 - 0,00684)	-0,00466	(-0,03811 - 0,00629)

Table III-3. *Résultats sur la totalité des locus (a) ou sans SSA00217NVH (b) de F_{IS} après 10 000 tirages avec remise des individus, pour chaque population par le logiciel Genetix.*

Pour ce qui concerne les estimations de F_{ST} et des distances de Nei (1978), avec ou sans ce marqueur, les valeurs sont similaires.

2.4. Analyse factorielle des correspondances

Lorsqu'on réalise une Analyse Factorielle des Correspondances (AFC), à partir des fréquences allèliques sur les 37 locus microsatellites, nous pouvons distinguer nettement les quatre populations (Oir, Scorff, Spey et Shin). Une séparation nette est visible entre les deux populations françaises (Oir et Scorff) par les axes 1 et 3, les françaises et les écossaises (Spey et Shin) par l'axe 1, et entre les deux écossaises, par l'axe 2. D'après les pourcentages d'inertie et la rupture de pente de la représentation graphique des valeurs propres (Figure III-10 et Table III-4), les trois premiers axes donnent une représentation pertinente de l'analyse de ces populations. Concernant les échantillons 2005 et passés (1992 pour Shin et 1988 pour les 3 autres) ceux-ci forment une même masse, sans réelle séparation.

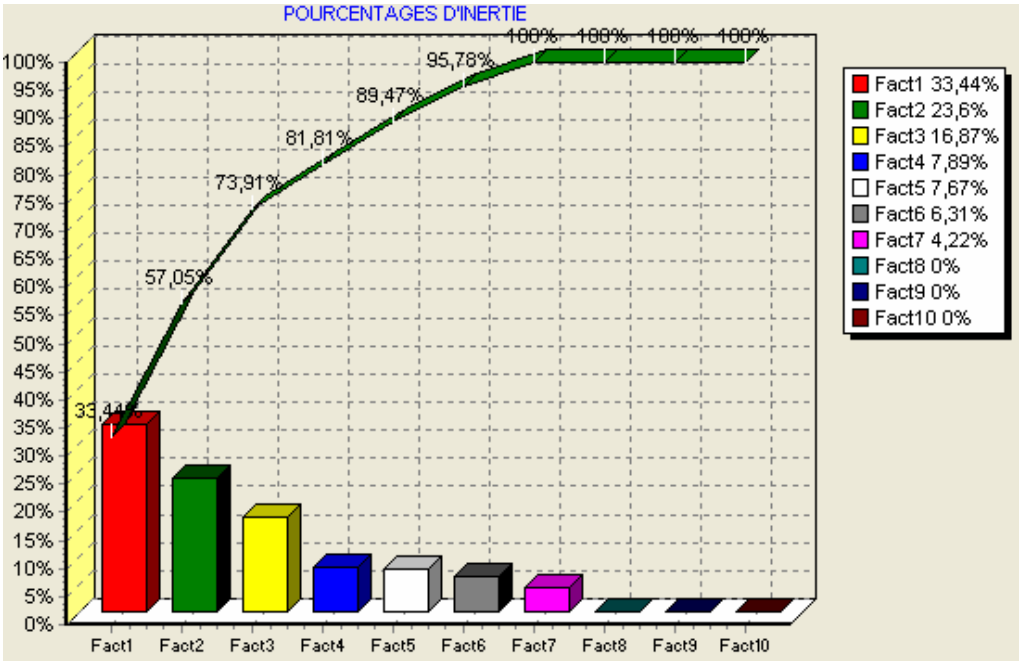


***Figure III-10**. Valeurs propres de l'AFC.*

Facteur	Valeurs propres	Pourcentage d'inertie	Pourcentage cumulé
1	0,22244	33,44%	33,44%
2	0,15698	23,60%	57,04%
3	0,11219	16,87%	73,91%
4	0,05249	7,89%	81,80%

***Table III-4.** Valeurs propres, pourcentages d'inertie et pourcentages cumulés de l'AFC.*

2.5. Structure

Les analyses avec le logiciel Structure (Pritchard *et al.* 2000) sur l'ensemble des échantillons (8) indiquent un début de plateau des valeurs moyennes des logarithmes népériens de la probabilité des données conditionnées par le nombre de groupes (Ln P(D)) pour un nombre de groupes K=4 (Figure III-11).

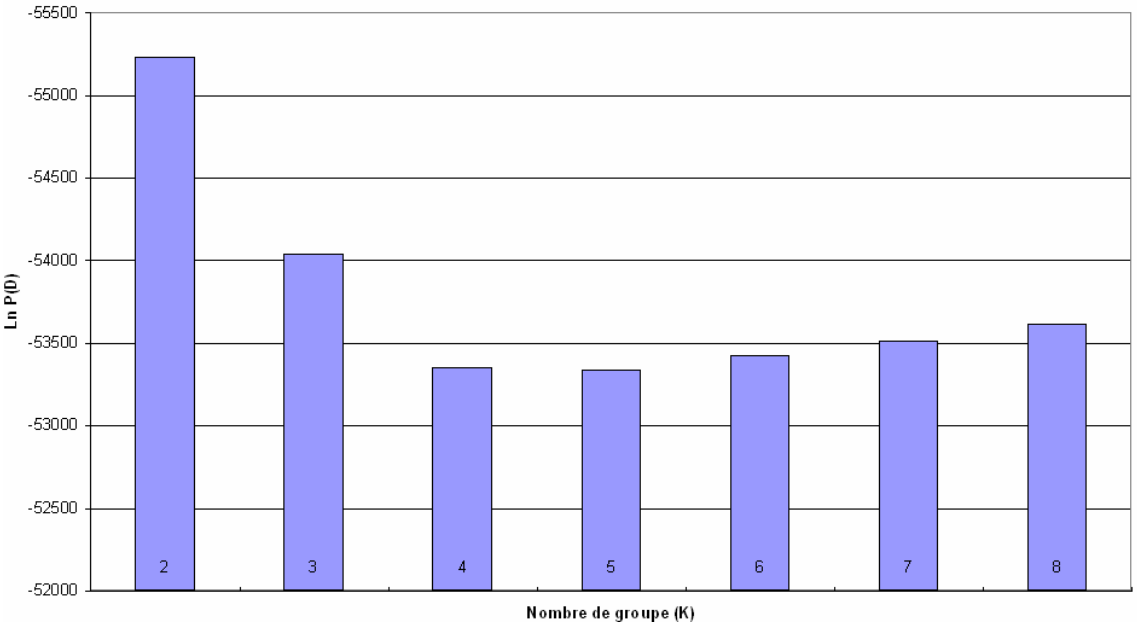


Figure III-11. *Évolution des Ln P(D) de K=2 à K=8.*

Nous avons pu observer que cette distinction des 4 populations se retrouvait, avec le logiciel STRUCTURE, de 37 à 10 marqueurs polymorphes. Le nombre de groupes le plus probable pour nos 8 échantillons est de 4 (Figure II-12) avec les échantillons de la même population (2005 et passés) par groupe.

Néanmoins, cette distinction s'affaiblit lorsqu'on diminue le nombre de marqueurs. Le pourcentage d'individus correctement assignés à sa population d'origine varie de 94% avec une analyse sous 37 et 28 marqueurs, 93% avec 20 marqueurs et 90% avec 10 marqueurs.

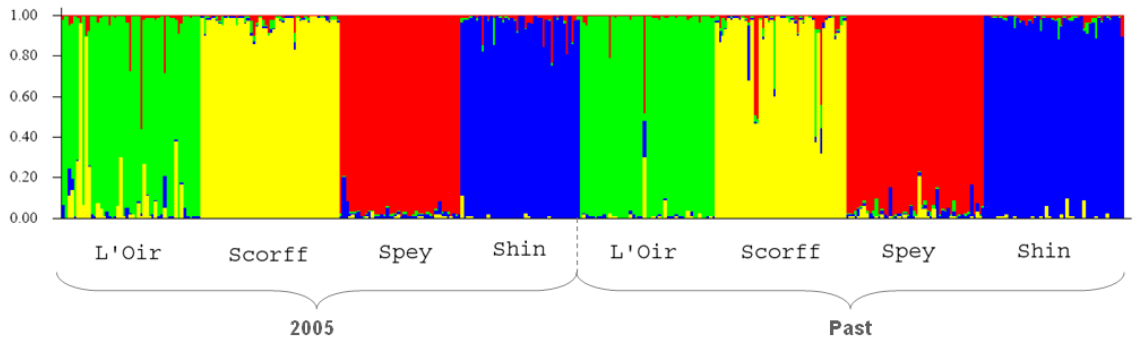


Figure III-12. *Regroupement des individus sous le logiciel Structure.*

2.6. Taille efficace

Les travaux de ce mémoire se sont principalement focalisés sur les estimateurs de taille efficace pour comprendre leurs sensibilités en fonction des marqueurs génétiques et des structures de populations. Ces analyses sont rapportées dans l'Article II avec l'ensemble des résultats et des interprétations. Les modèles généalogiques, tel que les modèles sous coalescence, sont de plus en plus utilisés pour estimer la taille efficace car ils facilitent la détermination des paramètres démographiques et de mutations en employant la variation de longueurs des séquences (Chakraborty and Kimmel 1999; Feldman et al. 1999; King et al. 2000 by Storz and Beaumont 2002). Nous avons, ainsi, choisi de comparer des méthodes sous coalescence avec des données réelles obtenues par génotypages microsatellites de populations de Saumon atlantique sauvage très différentes. L'ensemble de ces résultats est regroupé dans l'Article II de manière très approfondie. Ces résultats ont révélé que la précision des modèles dépendait du nombre et de la variabilité des marqueurs, de la structure des populations et de l'*a priori* sur la taille efficace, puisque ce sont des approches bayésiennes, avec des comportements atypiques pour certains. La méthode MSVAR semble plus homogène. Certaines observations sur les méthodes DIY ABC et TM3, non mentionnées dans l'article, sont présentées dans ces deux dernières parties qui suivent.

2.6.1. DIY ABC

Ce modèle de coalescence, développé récemment par Cornuet et al (2008), possède plusieurs particularités. Tout d'abord, il applique une approximation bayésienne (ABC) (voir Beaumont et Balding 2002). En second lieu, le calcul des tailles des populations peut se faire sous un scénario évolutif étant le plus probable parmi l'ensemble des scénarios testés. C'est-à-dire que le modèle propose de tester, parmi les scénarios possibles donnés par l'utilisateur, quel est l'agencement le plus probable des populations provenant d'un ancêtre commun. Dans ces scénarios, les nœuds peuvent être fixés et plusieurs échantillons de la même population peuvent être pris en compte sur la même branche. Dans notre étude, trois scénarios principaux ont été testés (Figure III-13) avec, pour chacun d'eux, toutes les combinaisons des populations possibles les unes par rapport aux autres : en étoile (scénario A), divergent (scénario B) et en cascade (scénario C).

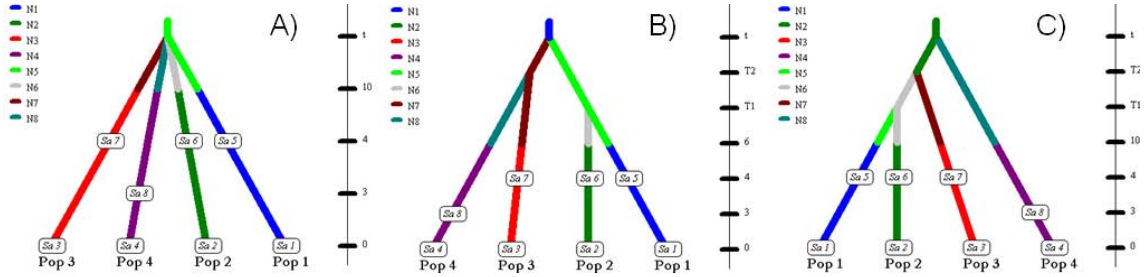


Figure III-13*. Scénarios évolutifs des quatre populations (Pop) échantillonnées deux fois (Sa) depuis un ancêtre commun. Le nombre de générations entre les échantillons de la population Shin (Pop4) est égal à 3 contre 4 générations pour les autres.*

Comme nous pouvons le voir à partir de ces figures, DIY ABC va estimer une taille ancestrale (*Na*) commune à toutes les populations et un temps ancestral (*Tf*) commun allant du moment de l'échantillonnage (*o*) au temps *t*.

Parmi les trois scénarios testés, deux scénarios sont apparus les plus probables parmi toutes les combinaisons possibles de position des 4 populations. Il s'agit des scénarios B et C. Pour le B, les populations écossaises (Shin et Spey) seraient au nœud le plus ancien et pour le C la cascade se ferait, en remontant vers les plus anciennes, des deux françaises (Oir et Scorff) vers Shin, et Spey. Comme l'arbre UPGMA donnait la même représentation que le scénario B, j'ai choisi de faire tourner les calculs d'estimation de taille efficace avec le scénario B.

Pour chacun des échantillons, les courbes de distribution $4Ne\mu$ (θ) *a posteriori* étaient obtenues en fonction des paramètres (*Ne* et μ) fixés *a priori***.** En comparant les courbes de distribution *a posteriori* et *a priori*, avec des *a priori*, proches ou non, de la réalité, on peut voir des différences par population suivant ces *a priori*. La population Oir, Scorff et Shin présente une courbe verte, *a posteriori*, qui ne suit pas la distribution *a priori* lorsque l'*a priori* est large (0-50 000 individus, supposés en taille efficace (*Ne*)) (voir l'Oir, Figure III-14b). Alors que pour la population Spey, les distributions *a posteriori* et *a priori* se suivent, quel que soit l'*a priori* fixé (Figure III-15). Ceci laisse à penser que l'information contenue dans le panel des 37 marqueurs microsatellites est insuffisante pour caractériser la population Spey par cette méthode. Des tests confrontant des rapports différents de taille de l'échantillon et de population (*S/N*) pourraient permettre d'anticiper l'impact sur la méthode et offrir à l'utilisateur une meilleure construction de ses données génétiques pour des analyses plus concordantes.

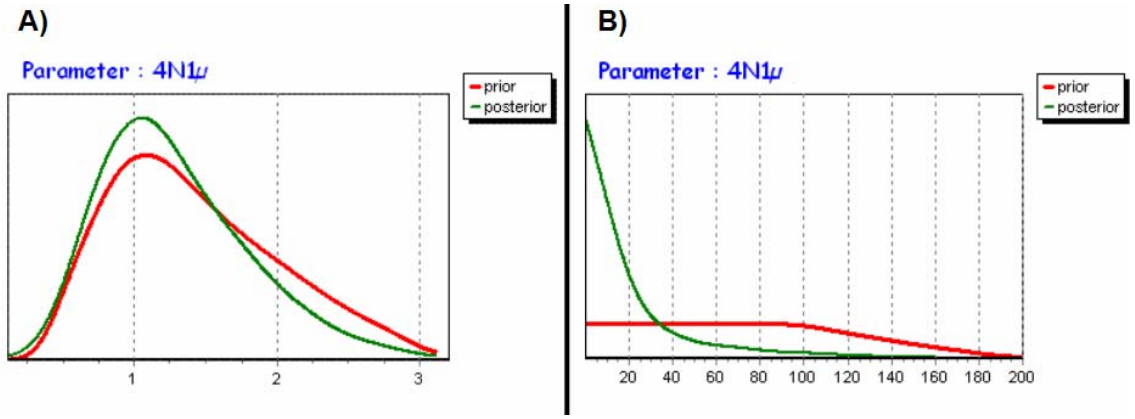


***Figure III-14**. Distributions pour la population Oir des Thêta (4N1μ) a priori (rouge) et a posteriori (verte) avec des a priori proches de la réalité [10-800] (A) et étendues [10-50000] (B). Les taux de mutations variant de 0,00005 à 0,05 pour les deux simulations avec un nombre d'itérations de 500 000.*

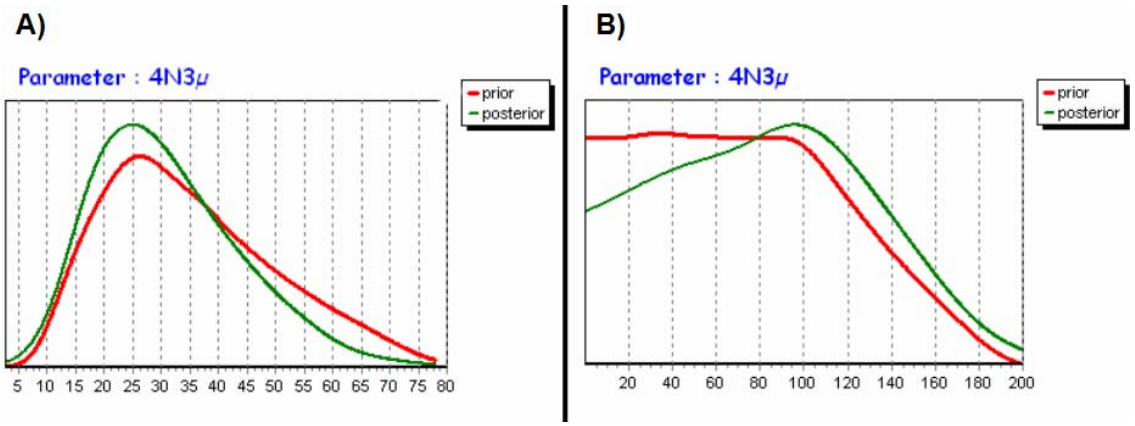


***Figure III-15**. Distributions pour la population Spey des Thêta (4N3μ) a priori (rouge) et a posteriori (verte) avec des a priori proches de la réalité [1000-20000] (A) et étendues [10-50000] (B). Les taux de mutations variant de 0,00005 à 0,05 pour les deux simulations avec un nombre d'itérations de 500 000.*

2.6.1. TM3

Rappelons brièvement, tout d'abord, le principe des méthodes basées sur les moments tels que celui de la méthode TM3. Ces méthodes utilisent deux échantillons de la même population pour calculer la taille globale de la population. Ces échantillons ne doivent pas être trop éloignés (quelques générations) de sorte à éliminer la force de mutation. Les différences génétiques observées entre les deux échantillons sont ainsi considérées comme uniquement liées à la force de dérive génétique. Plus ces différences sont grandes et plus la force agissant est considérée importante et, donc, l'estimation de la taille efficace se dirige vers une taille plus petite. La force de dérive génétique étant d'autant plus forte que la taille

efficace est petite. Ce modèle est un peu plus compliqué puisqu'il fait intervenir également la théorie de coalescence dans ces calculs (voir l'article de Berthier et al. 2002 pour plus de détail).

Waples (1989) avait démontré, par simulations, que les estimations de taille efficace pouvaient être biaisées (surestimation) si la distribution des allèles au sein des marqueurs était inégale. Tout d'abord, lorsqu'un allèle est surreprésenté (>50%), les différences entre échantillons sont réduites car elles se font sur un faible nombre d'allèles. Comme présenté dans l'article II, la population Oir et Scorff ont leur taille efficace surestimée, à partir d'un panel de marqueurs à faible nombre et faible hétérozygotie, car il contient un allèle surreprésenté (voir les exemples dans le panel 5H- (5 marqueurs les moins hétérozygotes) Figure III-16 et Figure III-17).

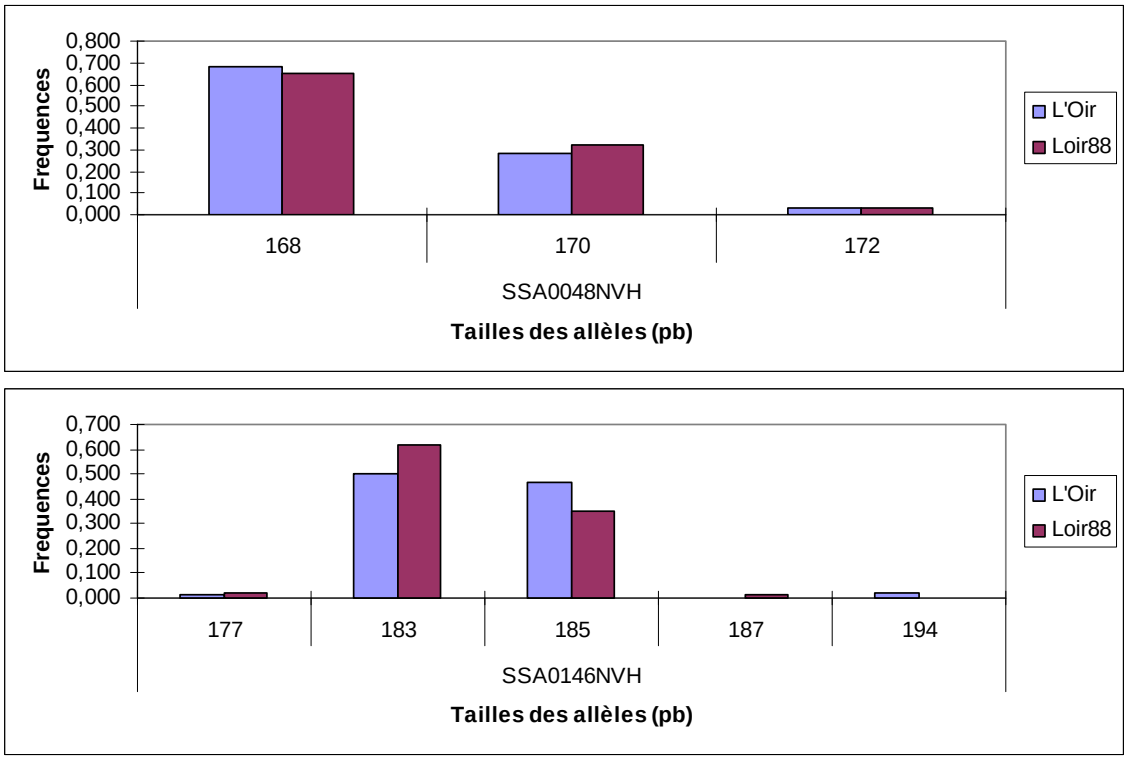


***Figure III-16**. Distributions alléliques des marqueurs SSA0048NVH et SSA0146NVH incluent dans le panel des 5 marqueurs les moins hétérozygotes (5H-) sur les échantillons de la population Oir en 2005 et 1988.*

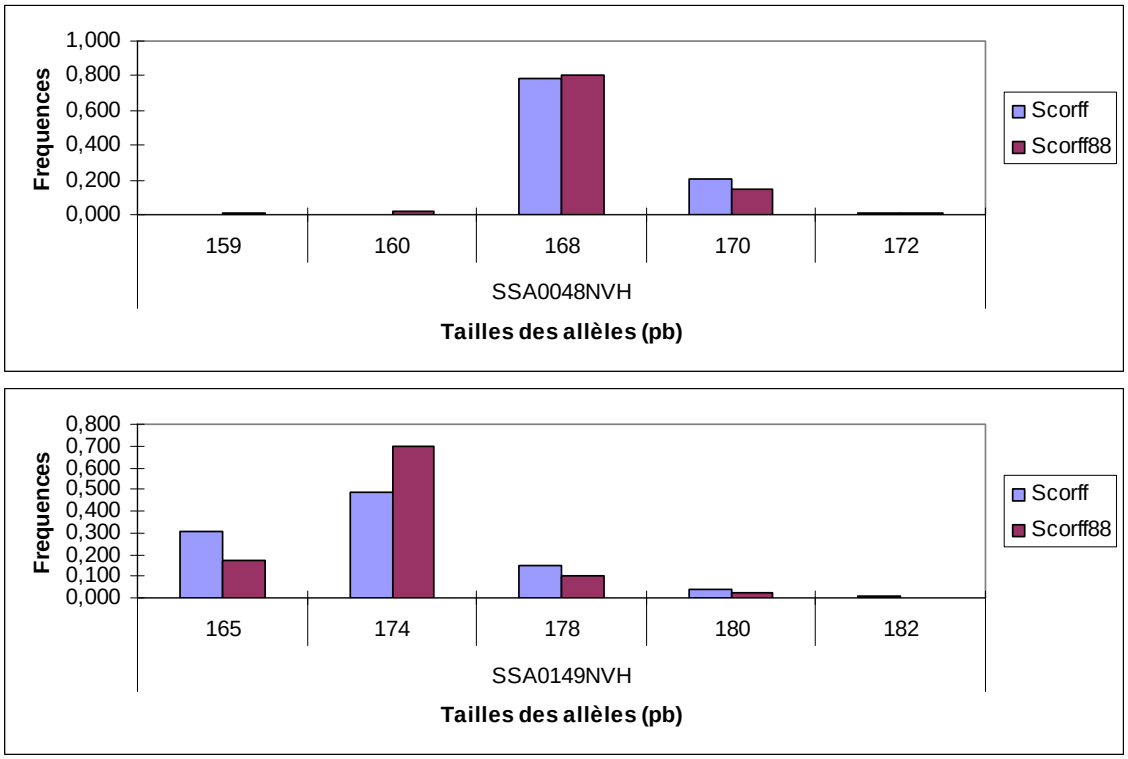


***Figure III-17**. Distributions alléliques des marqueurs SSA0048NVH et SSA0149NVH **inclus dans le panel** des 5 marqueurs les moins hétérozygotes (5H-) sur les échantillons de la population Scorff en 2005 et 1988.*

On pourrait cependant se poser la question : « pourquoi la population Oir est surestimée avec un panel, certes de peu de marqueurs (5 ou 10), mais à forte hétérozygotie (H+) ? ». Dans ce cas, il n'y a pas d'allèle majoritaire et l'exemple précédent n'est plus valable. La réponse se trouve en comparant les échantillons d'un même marqueur. La distribution allélique n'est pas homogène entre les deux échantillons (Figure III-18). Les allèles sont mal répartis et surestimeraient la taille car les différences se font sur un faible nombre d'allèles (exemple du marqueur SSA0217NVH).

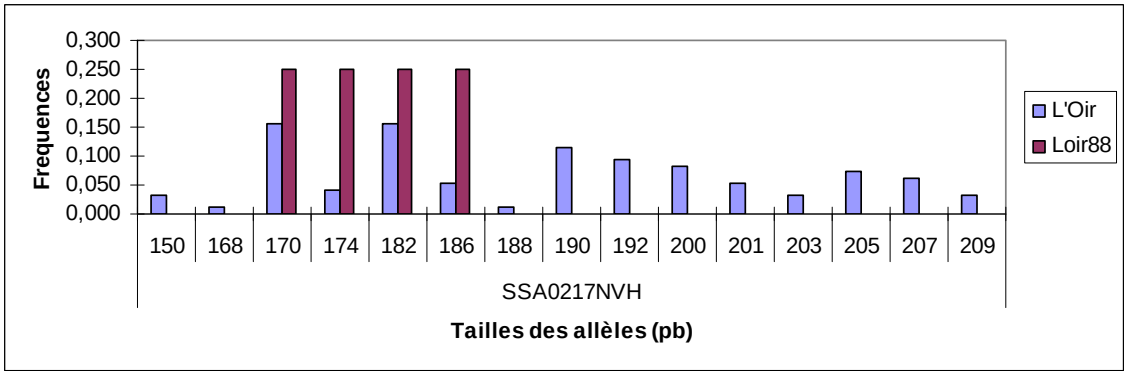


***Figure III-18**. Distributions alléliques du marqueur SSA0217NVH **inclues dans le panel** des 5 marqueurs les plus hétérozygotes (5H-) sur les échantillons de la population Oir en 2005 et 1988.*

2.7. Modèle DemoDivMS

DemoDivMS est un programme conçu pour prédire la diversité génétique à partir d'une population simulée sous microsatellites. La population peut être construite avec divers modèles de mutation et divers scénarios évolutifs faisant varier la taille efficace au cours des générations antérieures. Le programme est très rapide. Il est écrit en Fortran et utilise des routines numériques NAG. Le poids des fichiers ainsi créés est de l'ordre de 50 Ko.
Le programme se trouve sur le site de la plateforme de génétique quantitative de l'INRA, maintenu par David Robelin, Eddie Iannuccelli et Olivier Filangi (https://qgp.jouy.inra.fr).

2.7.1. Description du modèle

L'objectif du programme est de proposer un outil capable de décrire la diversité génétique actuelle, attendue à partir de données ADN de microsatellite, en fonction de l'histoire d'une population. L'utilisateur décrit les processus de mutation avec l'histoire démographique de la population, et le programme estime la diversité génétique actuelle en utilisant la théorie de la coalescence, développée avec tailles variables. Pour toute paire d'allèles échantillonnés dans la population, le programme calcule les probabilités *P0, P1, P2, ... Pd, ...* que les deux allèles montrent une différence *D* de *0, 1, 2, ... d ...* entre les nombres de répétitions du motif microsatellite des 2 allèles. Les mesures associées à la diversité sont : la moyenne et la variance de $|D|$ et les valeurs de Thêta (le paramètre $4N\mu$) correspondants à l'hétérozygotie attendue et à la distribution de *D*.

2.7.1.1. Processus de mutation

Les marqueurs microsatellites sont supposés évolués par saut (Stepwise Mutation Model) symétrique, de sorte que le nombre de motifs soit augmenté ou diminué de *k* avec la même probabilité. Ce programme propose trois modèles de mutation: « Single Step Model » qui fournit un processus de mutation par saut de +1 ou +1 ; « Geometric Model » les sauts se font avec la probabilité *(1-c) c^(k-1)* pour une distribution caractérisée par le paramètre *c* ($0<c<1$) qui est une raison de progression géométrique; et « User-Defined Model » avec des sauts k dont les probabilités sont fixées par l'utilisateur.

2.7.1.2. Démographie passée

L'histoire démographique est donnée comme une chaîne de *J* intervalles de temps [*t_i, t_ (i +1)*], dans lesquels les tailles efficaces *Ni* peuvent évoluer (augmenter ou diminuer) ou être stables. La taille est cependant supposée constante et égale à *N_J* avant un certain temps *t_J* dans le passé.

Figure III-19*. Représentation graphique des paramètres donnés par l'utilisateur au sein du modèle DemoDivMS, avec N pour la taille efficace et t pour le temps de coalescence, pour construire un scénario évolutif d'une population.*

2.7.1.3. Restrictions du modèle

a) Pas plus de 50 événements en temps *t_J* (*JMax*≤50).
b) Pas plus de 10 sauts par mutation dans le troisième modèle (*K*≤10).
c) Pas plus de 100 0000 générations entre *t_0* et *t_J*.

2.7.2. Exemple

Nous considérons une population actuelle de taille efficace égale à 1 000. Cette population aurait subi un goulot d'étranglement, il y a 100 générations, et un autre, il y a 2 000 générations, occasionnant une baisse de population de 10 000 individus à 5 000. Quelle est alors la diversité génétique actuelle de cette population sous un taux de mutation de 0.0009 et avec le modèle SMM ?

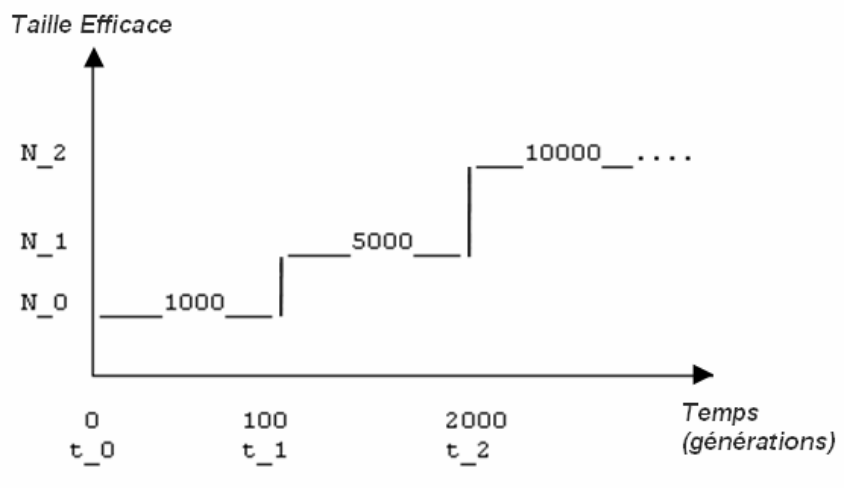


***Figure III-20**. Exemple graphique des paramètres donnés par l'utilisateur au sein du modèle DemoDivMS, avec N pour la taille efficace et t pour le temps de coalescence, pour construire un scénario d'une population ayant subi deux goulots d'étranglement il y a 100 générations et 2000 générations antérieures.*

Paramètres claviers :

>**1**

>**1000**

>**0.0009**

>**2**

>**5000 10000**

>**25 2000**

Fichier de sortie :

```
Microsatellite allelic diversity after demographic changes

Calculus of Pr(D = d / μ, N_0, N_1, ...),the probability that 2
microsatellite alleles show a difference of  d  repeat motifs according to
past demographic fluctuations

Single Step Mutation model, μ = 0.00090
Current effective size 1000.
Number of past demographic events = 2
Series of past effective sizes
   1000.   5000.  10000.
Series of times when size changed
      0.    100.   2000.
Largest value of D used = 39
Total probability of unconsidered events = 0.00008
Expected homozygosity Pr(D=0) = 0.17614
Probabilities Pr(D=1) to Pr(D=39)
0.09900 0.07014 0.05241 0.04035 0.03156 0.02485 0.01962 0.01551 0.01226
0.00969 0.00766 0.00605 0.00478 0.00378 0.00299 0.00236 0.00187 0.00148
0.00117 0.00092 0.00073 0.00058 0.00046 0.00036 0.00028 0.00023 0.00018
```

```
0.00014 0.00011 0.00009 0.00007 0.00005 0.00004 0.00003 0.00003 0.00002
0.00002 0.00001 0.00001
Mean and standard error of |D|: 3.74337   4.15841
Corresponding values of theta = 4Nμ under constant population size and SSM,
   from homozygosity :  15.61661
   from mean of |D|  :  28.51696
   from mean of D^2  :  31.30513
```

Résultats :

L'hétérozygotie de la population est élevée 84%.

L'ensemble des analyses que nous avons faites, grâce à ce modèle, ont permis de poser des conditions sur les possibilités d'observer des forts taux de diversité au sein de petites populations étudiées, l'Oir (200) et Scorff (1 000), ayant subi un goulot d'étranglement (2 000 - 4 000 générations passées) et provenant d'une population estimé à plus de 10 000 individus. Seule une taille intermédiaire variant de 2 500 à 5 000 individus effi**c**aces avec des taux de mutation de 0.0003-0.0009 peut expliquer ces forts taux d'hétérozygotie, et sous-entendant un autre goulot plus récent (25 - 100 générations) après cette période de taille intermédiaire.

Partie 3.

MODÈLE D'ESTIMATION DE LA TAILLE EFFICACE

Depuis les progrès en génétique moléculaire, les modèles utilisant l'information génétique pour estimer les paramètres génétiques des populations : les taux de migration, les taux de consanguinité, les taux de croissance et les tailles efficaces sont de plus en plus nombreux. Malgré les difficultés à estimer la taille efficace (Waples, 1989), elle reste l'un des paramètres les plus informatifs sur la viabilité d'une population. Ce qui en fait une variable clef en conservation puisqu'elle affecte la capacité d'une population à répondre à la sélection. Elle peut également renseigner sur les pertes de diversité génétique, les taux de fixation d'allèles délétères et l'efficacité de la sélection naturelle à maintenir les allèles avantageux (Berthier et al. 2002). Si *Ne* subit une diminution trop élevée, la perte de variabilité génétique résultant de la force de la dérive génétique devrait mettre l'espèce ou la population en risque d'extinction, en perdant le matériel nécessaire sur lequel opère la sélection.

3.1. Objectifs

Nous cherchons à évaluer la taille efficace d'une population à partir des observations obtenues sur la variabilité génétique observée en un certain nombre de marqueurs ADN microsatellites (marqueurs neutres et codominants). Cette variabilité est le reflet de l'interaction entre le processus de mutation qui génère de nouveaux allèles et le processus de dérive génétique qui tend à éliminer des allèles. La variabilité utilisée sera la différence des tailles entre allèles pour un marqueur donné, c'est-à-dire la différence du nombre de répétitions du motif microsatellite.

3.2. Procédures d'estimation

Nous recherchons plus spécifiquement l'ensemble des tailles efficaces depuis le moment de l'échantillonnage jusqu'à l'ancêtre commun. Afin d'estimer ces quantités inconnues (paramètres), on peut se placer dans le cadre statistique classique (recherche du maximum de vraisemblance) ou le cadre bayésien (mettre un *a priori* sur nos paramètres).

Notons Y l'ensemble des observations (les tailles des allèles observés dans l'échantillon, aux différents marqueurs ; dans notre modèle on verra que c'est plus spécifiquement la différence du nombre de répétitions entre deux allèles microsatellites), et θ l'ensemble des paramètres que l'on cherche à estimer. Il est à préciser que θ ne représente pas

seulement la taille efficace mais il inclut également le taux de mutation (*M*). Dans le modèle à taille constante, il se réduit en général au paramètre « Thêta » *4NeM*.

Dans le cadre bayésien, on va chercher à déterminer la distribution *a posteriori* de θ, soit : $P(\theta|Y)$. Cette probabilité *a posteriori* s'écrit, selon le théorème de Bayès :

$$P(\theta|Y) = P(Y|\theta)\, P_0(\theta)/P(Y)$$

Où $P_0(\theta)$ est la distribution *a priori* des paramètres qui traduit la connaissance préalable que l'on a des paramètres. Dans cette expression, $P(Y|\theta)$ est la vraisemblance des observations Y pour la valeur θ. Que l'on se place dans le cadre bayésien ou dans le cadre classique, le calcul de la vraisemblance est un point de passage obligé. Elle a été résolue dans notre modèle en passant par des transformées de Fourier. Pour estimer l'a postériori, on utilise un algorithme MCMC (Métropolis Hastings) qui permet d'intégrer numériquement l'intégrale de $P(Y)$.

3.2.1. Variables latentes

Le cadre Bayésien permet d'accéder à une estimation de la distribution *a posteriori* par le recours aux méthodes MCMC. On peut, en effet, introduire des variables latentes qui permettent de calculer facilement des vraisemblances conditionnées par ces variables, en générant ces variables et en intégrant sur tous les cas possibles pour obtenir *in fine* la distribution des paramètres d'intérêt θ. Cela revient à considérer ces variables latentes comme des paramètres supplémentaires du modèle, même si on ne cherche pas à les estimer. Calculer la taille efficace en génétique des populations peut se faire en passant par la théorie de la coalescence. Dans ce cas, les variables latentes correspondent aux arbres de coalescence *A*, représentant les généalogies possibles des gènes échantillonnés (Balding et Wilson, 1998). On écrit :

$$P(Y|\theta) = \Sigma_A\ P(Y|\theta\, A)\, P(A|\theta)$$

Les quantités $P(Y|A\theta)$ peuvent se calculer en détaillant la description des arbres. Mais la somme sur tous les arbres possibles ne peut pas se faire à cause de leur trop grand nombre. Les méthodes mises en œuvre simulent les arbres, en évitant ceux qui correspondent à des probabilités infimes. Même si, à la fin, on ne s'intéresse qu'à la distribution marginale des θ, des valeurs des paramètres d'intérêt θ et des paramètres « surnuméraires » *A* sont générés

simultanément, dans les chaînes de calcul, pour estimer la distribution de l'ensemble des θ et des *A*. Cependant ces méthodes (telles que LAMARC, MSVAR) s'avèrent très lourdes avec de nombreux marqueurs.

3.2.2. La vraisemblance

Pour limiter les temps de calcul, nous avons cherché à éviter d'introduire des simulations d'arbres. Pour cela, nous déterminons les paramètres d'intérêt par une approximation directe de la vraisemblance $P(Y|\theta)$ en se focalisant sur des modèles particuliers de mutation et de démographie.

3.2.2.1 Modèle de mutations

Nous avons fait le choix de travailler sur des marqueurs ADN microsatellites car leur neutralité permet d'éliminer l'effet de sélection dans le modèle et leur polymorphisme élevé permet d'accéder à une grande variabilité génétique. Ces marqueurs sont des séquences répétées en tandem qui permettent une estimation de leur taille en fonction de leur nombre de répétitions (ex : ATATATATAT = 5 répétitions de la paire Adénine Thymine). Le modèle le plus adapté pour ces marqueurs est le modèle dit SMM (Stepwise Mutation Modèle), qui est un modèle symétrique. Dans ce cas, les mutations s'accumulent au cours des générations par sauts en amont ou en aval. La probabilité (*m*) qu'une mutation apparaisse correspond à une modification du nombre de répétitions. Ces probabilités supposent que quelque soit la longueur (*i*) d'origine de l'allèle, la distribution des sauts, en cas de mutation, ne dépend pas de (*i*). La probabilité de mutation, à saut +1 ou -1, à chaque génération est égale à *u/2*. Et la probabilité qu'il n'y est pas de mutation est de 1-*u*.

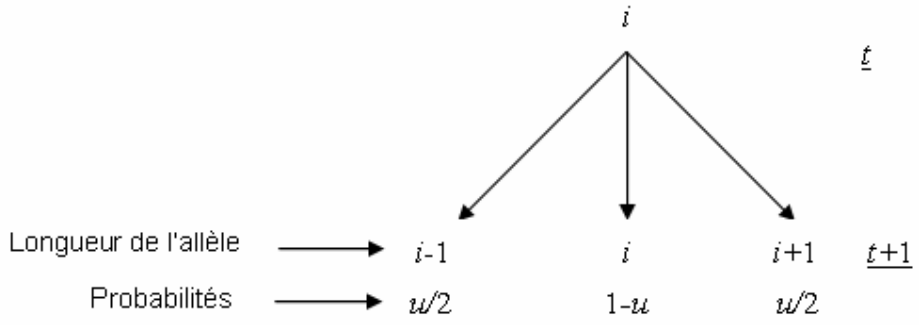
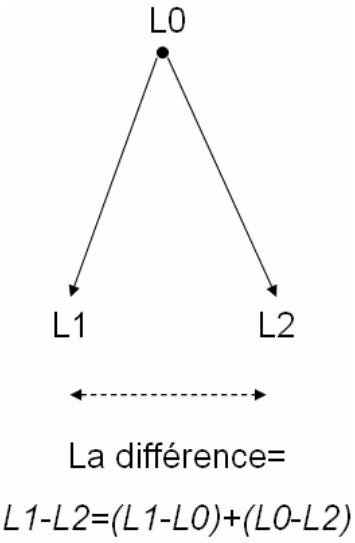


Figure III-21. *Probabilités de mutation à la génération suivante.*

La probabilité d'observer une longueur *i* à la génération suivante peut s'écrire :

$$P(D(g+1)=i)=P(D(g)=i)(1-u)+P(D(g)=i-1)\frac{u}{2}+P(D(g)=i+1)\frac{u}{2}.$$

En travaillant sur la différence (*k*) de longueurs entre deux allèles de la même génération, et non pas sur la différence entre un des 2 allèles et l'ancêtre, on peut représenter les calculs comme suit :



Cette différence peut s'écrire par une transformée de Fourier en fonction d'un nombre imaginaire sur le cercle trigonométrique (α) ($\alpha = e^{ix}$, $x = réel$).

On retrouve ainsi *L1-L0* représenté par :

$$f_1(\alpha)=\left(\frac{1}{2}u\frac{1}{\alpha}+(1-u)+\frac{1}{2}u\alpha\right)^g f_0(\alpha),$$

avec $f_0(\alpha)=1$. Ce qui donne :

$$f_1(\alpha)=\left(1-u\left(1-\frac{1}{2}\left(\alpha+\frac{1}{\alpha}\right)\right)\right)^g.$$

Comme ceci est pareil pour *L0-L2* :

$$f_2(\alpha)=\left(1-u\left(1-\frac{1}{2}\left(\alpha+\frac{1}{\alpha}\right)\right)\right)^g,$$

et sachant que la différence *L1-L2* peut être représenté par le produit $f_1(\alpha) * f_2(\alpha)$, alors on peut généraliser la différence entre deux allèles par :

$$F_g(\alpha) = \left[1 - u\left(1 - \frac{1}{2}\left(\alpha + \frac{1}{\alpha}\right)\right)\right]^{2g}$$

Si u<<1, on peut écrire :

$$F_g(\alpha) = \sum_{k=-\infty}^{k=+\infty} \alpha^k (P(D(g) = k)) \cong e^{-2ug\left(1-\frac{1}{2}\left(\alpha+\frac{1}{\alpha}\right)\right)}$$

En travaillant sur les temps réduits *t=g/No* et, dans le cas de la taille constante, on peut écrire :

$$F_t(\alpha) = \sum_{k=-\infty}^{k=+\infty} \alpha^k (P(D(t) = k)) \cong e^{-2ut\left(1-\frac{1}{2}\left(\alpha+\frac{1}{\alpha}\right)\right)}$$

Comme nous le verrons par la suite cette équation qui peut également se lire comme la transformée des *P(D=k|T)* avec *T* un intervalle de coalescence. On écrira *P(D=k|T,M)* lorsque le taux de mutation (*M*) intervient. Lorsque les tailles varient, cette fonction de Fourier fait intervenir les tailles (*N*) et sera considérée comme la transformée des quantités *P(D=k|T,N,M)*.

3.2.2.2 Modèle démographique

L'évolution des tailles efficaces sera modélisée par une fonction en escalier comme celle appliquée par Drummond et al (2005). À chaque intervalle de génération (*ti*, *ti+1*) correspond une espérance de taille efficace (*Ni*). On désigne dans la suite l'ensemble de ces paramètres (les *ti*, et les *Ni*) par *N*.

3.2.2.3 Résultats théoriques

Les résolutions de la vraisemblance *P(Y|θ)*, font intervenir les probabilités des données (des différences de motifs entre deux allèles) conditionnées par *N* et *M*. Pour résoudre cette vraisemblance, on intègre sur les temps de coalescence (*T*) en étudiant d'abord la probabilité que ces 2 allèles coalescent dans un temps de coalescence inférieur à une valeur donnée (T<*t*), et qu'ils présentent une différence *D* de ± *k* motifs. Elle peut s'écrire :

$$P(D = k|N, M) = \int dP(D = k, t < T < t + dt|N, M) = \int \frac{d}{dt} P(D = k, T < t|N, M)dt\,.$$

Les probabilités conjointes $P(D=k,T<t|N,M)$ sont des résultats nouveaux et sont présentés dans l'article III, intitulé *« Distribution of coalescence times and of distances between microsatellite alleles with changing effective population size »*. Les résolutions se sont faites à partir d'une transformée de Fourier (F) intégrée par rapport aux temps pour chaque intervalle (voir l'équation 12 pour une taille constante et l'équation 17 pour une taille variable (c'est-à-dire pour chaque intervalle de temps)) puis inversée pour obtenir numériquement les probabilités recherchées (voir l'équation 11 pour taille variable).

La formule d'inversion de l'équation de Fourier $F(x,M,N) = \Sigma_k P(D= k|N,M,T) \cos(kx)$:

$$P(D = k|M, N) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(x, N, M) \cos(kx) dx$$

La théorie permet d'obtenir des expressions analytiques explicites de la fonction F, et donc, une seule intégration numérique est requise pour obtenir un résultat exact.

3.2.2.4. Approximation de la vraisemblance

Pour appliquer les probabilités ci-dessus à l'ensemble des marqueurs, une vraisemblance approchée est proposée. Les précédents résultats permettent de calculer la probabilité qu'une paire d'allèles présente une différence donnée, en fonction des paramètres N et M. On peut donc, à partir d'un échantillon de n allèles en plusieurs locus, confronter les fréquences de paires d'allèles qui présentent une différence k à la valeur attendue $P(D=k|N,M)$ selon le modèle (N,M). Cela peut suggérer une première vraisemblance approchée, en considérant des tirages indépendants de paires et en posant un modèle multinomial sur les tirages de paires. Pour un échantillon comprenant $n_0, n_1,\ldots n_k$ paires d'allèles présentant des distances $0, 1,\ldots k$ une première approximation de la vraisemblance s'écrit, pour chaque marqueur :

$$Log\ L_1\ (n_0,n_1,..n_k|N,M) = Cte + \Sigma_k\ \ n_k\ \ log\ P(\ D = k|N,M)$$

Cependant, lorsqu'on applique la théorie de la coalescence, on ne peut pas faire les calculs en supposant que les paires d'allèles sont indépendantes les unes des autres. L'ensemble des n allèles d'un échantillon sont réunis dans un même arbre de sorte que les états possibles d'un groupe d'allèles dépendent de l'état de tout autre groupe. Les fréquences f_k des paires d'allèles à distances k ne sont pas des variables indépendantes. Pour généraliser à

plus de 2 allèles, il faut calculer des probabilités d'identité conjointe et non pas de simples probabilités d'identité entre gènes. Cela permet par exemple de calculer la probabilité que, si l'on tire un échantillon de 4 allèles, ceux-ci diffèrent de *p*, *s* et *t* motifs, c'est-à-dire qu'ils soient des allèles *A(i), A(i+p), A(i+s) et A(i+t)* (quel que soit *i*), où *A(k)* désigne un allèle comportant *k* motifs. Les résultats sont obtenus pour 3 et 4 allèles. Au-delà, plusieurs difficultés surgissent : la combinatoire devient très lourde ; l'obtention de résultats numériques nécessite pour les inversions des intégrales doubles, triples, …; enfin l'obtention même des solutions est limitée par le fait que pour un échantillon de grande taille, il n'est plus correct de supposer qu'un seul événement de coalescence (au plus) est possible à chaque génération. S'il n'est pas possible par cette approche d'aboutir à la vraisemblance exacte d'un échantillon de taille quelconque, une deuxième approximation peut être proposée, en calculant les variances et covariances des fréquences des paires (f_p, f_s). En effet, calculer ces moments du second ordre revient à considérer les probabilités des états concernant 3 et 4 allèles.

Le calcul explicite des $Var(f_p)$ et des $Cov(f_p,f_s)$ est obtenu comme précédemment, en introduisant les fonctions génératrices de ces quantités dont on établit les solutions selon le modèle démographique et le modèle de mutation. On utilise ensuite une inversion finale pour obtenir les valeurs numériques. Il faut noter que ces variances et covariances comprennent des termes « structurels » (correspondant au cas où les fréquences f_p et f_s sont en concordance avec la population totale), et des termes d'échantillonnage liés à la taille limitée de l'échantillon. On peut alors proposer une seconde approximation de la vraisemblance, en admettant une approximation gaussienne pour la distribution des f_k :

$$Log\ L_2(\ f_0, f_1, \ldots f_k/M,N) = Cte\ \ -½\ log\ (|V(M,N)|) - ½\ \ [\ f - P\ ]'\ \ V(M,N)^{-1}\ \ [\ f - P\ ]$$

Où $\boldsymbol{V}(N,M)$ est la matrice des variances et covariances pour les indices *p,s*… présents dans l'échantillon ($f_0,f_1,\ldots f_k$), $|\boldsymbol{V}(M,N)|$ est le déterminant de cette matrice et où $[\ f - P\ ]$ représente le vecteur des différences $(\ f_k - P(D=k|N,M)\ )$.

Pour chaque marqueur correspond une vraisemblance. Ces vraisemblances sont ensuite sommées pour obtenir les Thêta (*N, M*) finaux avec une variance sur l'ensemble des marqueurs. Dans notre modèle l'utilisateur fixe le taux de mutation (*M*), ce qui permet d'en déduire la taille efficace (*N*). Comme nous avons travaillé sur des calculs de probabilités délimités sur des intervalles de temps (*T*), les tailles efficaces peuvent être déduites sur chaque intervalle. Pour cela, l'utilisateur fixe au départ le nombre d'intervalles qu'il souhaite analyser entre la taille actuelle (*No*) et la taille ancestrale (*Nmax*).

3.2.3. L'*a priori* $P_0(\theta)$

Les *a priori* seront également donnés par l'utilisateur sur le taux de mutation *M* et la taille efficace (*N*). Pour *M* une seule valeur pourra être donnée pour l'ensemble des marqueurs. On prévoit une extension pour faire varier le *M* entre marqueurs. Concernant le *N*, plusieurs *a priori* pourront être donnés selon les intervalles délimités par l'utilisateur entre *No* et *Nmax* et la connaissance qu'il a sur la population.

3.3. Autres résolutions pour tester le modèle

Afin de tester notre modèle nous avons envisagé l'histoire d'une population soumise à un goulot d'étranglement intense. La vraisemblance peut être décomposée en deux probabilités pour tester le comportement des temps de coalescence conditionnés par les données, la taille efficace et le taux de mutation. La vraisemblance décomposée :

$$P(D = k, T < t \mid N, M) = P(T < t \mid D = k, N, M) * P(D = k \mid N, M),$$

permet d'obtenir la probabilité des temps de coalescence *T*, conditionnée par *D*, *N* et *M* :

$$P(T < t \mid D = k, N, M) = \frac{P(D = k, T < t \mid N, M)}{P(D = k \mid N, M)}.$$

Au sein de l'article III vous retrouvez les résolutions de $P(D = k \mid N, M)$ à l'équation 32 (qui n'est autre que l'équation 17 pour *T* qui tend vers l'infini $F(\infty, \alpha)$**).** Les espérances de ces distributions sur tous les *T* font intervenir une transformée de Laplace, voir l'équation 28à l'article III.

ARTICLE 3

Distribution of coalescence times and distances between microsatellite alleles with changing effective population size

CLAUDE CHEVALET
NATACHA NIKOLIC

Theoretical Population Biology (Soumis)

Résumé

Cet article décrit les résolutions analytiques pour étudier les effets des changements de la taille effective d'une population sur la diversité allélique actuelle à un marqueur microsatellite. Dans un premier temps, nous décrivons les calculs des probabilités conjointes de la distance entre les deux allèles (la différence du nombre de motifs microsatellite) et du temps de coalescence à l'ancêtre commun le plus récent, compte tenu de l'histoire démographique et du processus de mutation. Les solutions sont données par des moyennes d'analyses d'inversions sous les deux hypothèses, taille de la population constante et taille variables. Ces inversions numériques ont permis de résoudre les distributions des temps de coalescence et celles des distances alleliques pour tous les modèles de mutation démographique et pour toutes les histoires. Des illustrations numériques sont données pour montrer les effets des variations de la taille de la population sur les distributions des temps de coalescence et des fréquences *f(k)* des différences de k motifs entre paires d'allèles. La possibilité de détecter un ancien goulot d'étranglement est discutée. Les extensions de la méthode pour faire face à plus de deux allèles sont proposés, et pour résoudre le cas de quatre allèles. Cela permet d'examiner la répartition de covariance des fréquences *f(k)* et de proposer d'autres approches d'estimation.

Distribution of coalescence times and distances between microsatellite alleles with changing effective population size

Claude Chevalet and Natacha Nikolic

Laboratoire de Génétique cellulaire, UMR 444 INRA/ENVT
Centre de Recherche INRA de Toulouse
BP52627, 31326 CASTANET TOLOSAN CEDEX, France

Corresponding author
Claude Chevalet
Laboratoire de Génétique cellulaire, UMR 444 INRA/ENVT
Centre de Recherche INRA de Toulouse
BP52627, 31326 CASTANET TOLOSAN CEDEX, France
Telephone: +33 (0) 561 285 117
Fax: +33 (0) 561 285 308
Email: claude.chevalet@toulouse.inra.fr

Abstract

We investigate the effects of past changes of the effective size of a population on the present allelic diversity at a microsatellite marker locus. We first derive the analytical expression of the generating function of the joint probabilities of the time to the More Recent Common Ancestor for a pair of alleles and of their distance (the difference in allele size). We give analytical solutions in the case of constant population size and geometrical mutation model. Otherwise numerical inversion allows the distributions to be calculated in general cases. The effects of population expansion or decrease and the possibility to detect an ancient bottleneck are discussed. The method is extended to samples of three and four alleles, which allows investigating the covariance structure of the frequencies $f(k)$ of pairs of alleles with a size difference of k motifs, and suggesting some approaches to the estimation of past demography.

Keywords

Population effective size, Microsatellite, Genetic diversity, Coalescence times, Demography, Fourier transform, Laplace transform.

1. Introduction

The effective population size (N_e) is an important parameter in ecology, evolutionary biology and conservation biology (Wang 2005). Different concepts of N_e have been proposed such as the inbreeding effective size, variance effective size, eigenvalue effective size, mutation effective size and coalescent effective size (Felsenstein 1971; Ewens 1982; Gregorius 1991; Caballero 1994; Whitlock and Barton 1997; Charlesworth *et al.* 2003). In general, this parameter is difficult to estimate mainly because of the highly stochastic nature of the processes of inbreeding and genetic drift for which it is usually defined and measured (Wang 2005). Many methods have been developed in the past three decades to estimate the contemporary and ancestral effective population sizes using different information extracted from some genome fragment in a sample of individuals. Estimating present and ancestral population sizes is based on the joint effects of population size and of mutation process on the distribution of alleles in a sample. Most estimators of N_e are based on models in which it is assumed to be constant, or submitted to a regular (linear or exponential) change, while this is unlikely in natural populations (Waples 2005).

In this paper we investigate the influence of population size changes on the distribution of coalescence time between alleles, and on the distribution of allelic frequencies. We focus on the class of microsatellite DNA markers. Microsatellites are loci that are made of repeated short motifs (one to six base pairs) embedded between locus specific sequences. Alleles are characterized by the numbers of repeats that can be observed from the lengths of PCR amplified fragments, so that they behave as co-dominant markers. These

markers are generally assumed to be neutral, which allows the neutral model of genetic drift to be used. Mutations within microsatellites are frequent, altering their overall length by insertion or deletion of a small number of repeat units. This property allowed inferences to be made about the history of human populations (Bowbock *et al.* 1994), and recent demographic events to be detected, such as the expansion of human populations since the palaeolithic era (Shriver et *al.* 1997; Reich and Goldstein 1998; Reich *et al.* 1999). Although most recent population surveys are making use of genome-wide panels of SNP (Single Nucleotide Polymorphisms) markers, microsatellite markers should remain useful to infer population structures, because of their neutrality and their high polymorphism. For example Sun *et al.* (2009) showed that microsatellite markers are efficient molecular clocks to investigate the history of populations, and to provide better (unbiased) estimates of population differentiation (F_{ST}) than with SNPs. Most results on the distributions of allelic frequencies were developed to investigate the variation across marker loci of the within population diversity and the relationships between specific genetic distances and divergence times (Goldstein *et al.* 1995; Shriver *et al.* 1995; Kimmel *et al.*, 1996). Works considered different mutation models including non symmetrical mutation process (Pritchard and Feldman 1996; Kimmel *et al.* 1996; Kimmel and Chakraborty 1996) and accounting for dependence of the outcome of a mutation on allele length (Watkins 2006).

We extend these analyses to some variants of the Stepwise Mutation Model, and to the case when population size has undergone changes in the past. We focus on the distribution of coalescence times between alleles conditional on the difference between alleles, and on the distribution of differences

in repeat numbers in a set of alleles sampled from a population. We first derive the joint distribution of coalescence times and of differences between alleles. Then we show how the methods provide analytical solutions in the cases when population size changes are modelled by a step function and the mutation process is modelled by a geometric function, and how they allow exact numerical results to be obtained for general mutation models and for any model of population size variation.

2. Models and Methods

2.1. Background and notations

Modelling the mutation process of microsatellite took advantage of the model devised for electrophoretically detectable alleles of genes coding for proteins (Ohta and Kimura 1973). In the simplest version of the Stepwise Mutation Model (SMM), the number of repeats of a microsatellite locus is changed by $+1$ or -1 with equal probabilities if a mutation occurs (single step SMM). In the following, we shall refer to the difference D in repeat numbers between two alleles as the distance between these alleles. In a diploid population of constant effective size N, the distribution of allelic frequencies reaches an equilibrium such that the expected homozygosity can be written:

$$Hom = \frac{1}{\sqrt{1+8N\mu}}, \tag{1}$$

where μ is the mutation rate. As usual we set

$$\theta = 4\,N\,\mu.$$

More generally, the probability that a pair of alleles exhibits a distance k was shown to be equal to:

$$Pr(D = k) = \frac{1}{\sqrt{1+2\theta}}(\frac{1+\theta-\sqrt{1+2\theta}}{\theta})^{|k|}, \quad (2)$$

where k can take any integer value (Wehrhahn 1975). A way to derive the result is to consider the probabilities $Pr(D = k/g)$ that 2 alleles are at distance k conditional on g, the number of generations to their Most Recent Common Ancestor (MRCA). Under the single step SMM the generating function of the probabilities takes the following (approximate) expression:

$$\sum_{k=-\infty}^{k=+\infty} \alpha^k \times Pr(D = k/g) = e^{-2g\mu} \times e^{2g\mu \frac{1}{2}(\alpha+\frac{1}{\alpha})}, \quad (3)$$

where $\alpha = e^{ix}$ (x real) stands for any complex number in the unit circle ($|\alpha| = 1$). The second term is known to be the generating function of the series $I_k(2g\mu)$ where I_k is the modified Bessel function of the first kind. This result is then combined with the distribution of T_{MRCA}, the time to the MRCA. For a population with constant effective size N it is:

$$Pr(T_{MRCA} < t) = 1 - e^{-t}, \quad (4)$$

when the time t is expressed relative to population size and linked to the number of generations g by:

$$t = \frac{g}{2N_0}. \quad (5)$$

In the following, the definitions of t and of θ are related to the population size at the present time (N_0). We consider a monoecious diploid population, with non-overlapping generations, random mating and a finite but changing effective size. Effective size is assumed to be large enough for the continous approximation to be used.

2.2. Mutations

We assume the mutation process can be described by a generalized symmetrical Stepwise Mutation Model. If a mutation occurs, the number of repeats of the allele may be changed by $+r$ or $-r$, with equal probabilities $\frac{1}{2}m_r$ ($\sum_r m_r = 1$). This model is characterized by the generating function:

$$M(\alpha) = \sum_{r>0} m_r \frac{1}{2}(\alpha^r + \frac{1}{\alpha^r}) \;\; = \;\; M_R(x) = \sum_{r>0} m_r \, cos(rx), \tag{6}$$

where $\alpha = e^{ix}$ (x real). In general $M(\alpha)$ would be complex, but for symmetrical models, it is real and equal to its real part we write as $M_R(x)$. Both forms in α or x are used for the generating functions found in the following. The single step model is characterized by $M_R(x) = cos(x)$. Another special model is obtained assuming a geometric progression for the probabilities m_r with a common ratio c, $c < 1$, so that $m_r = (1-c)c^{r-1}$ (Whittaker 2003). The corresponding generating function (Watkins 2006) is:

$$M_R(x) = \frac{(1-c)(cos(x)-c)}{1 - 2 \;\; c \;\; cos(x) + c^2}. \tag{7}$$

Coping with a general mutation model is then straightforward, simply replacing the expression $\frac{1}{2}(\alpha + \frac{1}{\alpha})$ by the appropriate $M(\alpha)$ function in expressions like Eq. (3).

2.3. The joint distribution of coalescence time and of allele distance

We introduce a functional generalization of inbreeding coefficients in order to deal with the joint distributions of coalescence times and of allele differences. The approach follows the methods used to deal with migrations between demes in a discrete infinite lattice (Malécot 1951, 1975, 1981;

Kimura and Weiss 1964). It is an alternative to the combination of *a priori* coalescence time distribution (given by Eq. 4 for constant population size, or by Polanski *et al.* (2002) for time-dependent population size) with the distribution of D conditional on coalescence times (Eq. 3 for the single step SMM). It allows also extensions to the calculation of joint allele frequencies among 3 or 4 alleles, as for higher order probabilities of identity between genes (Gillois 1964, 1965; Chevalet *et al.* 1977).

(Figure 1)

We consider a random pair of alleles at generation (g), and the probability $P_g(k) = Pr(D(g) = k, G \leq g)$ that these alleles derive from a common ancestor gene G generations ago (with $G \leq g$), and that these alleles differ by k repeats. Let $N(g)$ be the effective size of the population at that generation. We calculate the probability of such events for a sample of two alleles drawn at the next generation $(g+1)$.

· With probability $\frac{1}{2N(g)}$, the two alleles are copies of a single ancestor gene in the g-th generation (Figure 1-A). If no mutation occurred (probability $(1-\mu)^2$) both alleles are alike and their distance is 0. If some mutation occurred, the combined probability of the event is of the order of μ/N and will be neglected.

· With probability $1 - \frac{1}{2N(g)}$, the two alleles are copies of two different genes in the g-th generation. Let k be the distance between these two ancestor alleles, which happens with probability $P_g(k)$ for alleles that derive from a single gene less than g generations ago. If no mutation occurred (probability

$(1-\mu)^2$) the two new alleles are at distance k (Figure 1-B). If one mutation occurred (probability $2\mu(1-\mu)$) for one of the alleles, the two new alleles are at distance $k+r$ or $k-r$ with probabilities $\frac{1}{2}m_r$ (Figure 1-C). In the same way as we neglect terms in μ/N, we also neglect contributions involving two mutations (terms in μ^2).

Combining the different cases we can write:

$$\begin{aligned} & P_{g+1}(k) \\ & = \quad \frac{1}{2N(g)}\, P_g(k)\, \delta_{0,k} \\ & + \quad (1-\frac{1}{2N(g)}) \left((1-2\mu)P_g(k) + 2\mu \sum_r m_r (\frac{1}{2}P_g(k-r) + \frac{1}{2}P_g(k+r)) \right), \end{aligned} \tag{8}$$

where $\delta_{s,k}$ stands for the Kronecker function, equal to 1 if $s = k$, and 0 if $s \neq k$.

An alternative way to derive these equations is to consider the second order moments of allele frequencies. Let p_i be the frequency of alleles with i motifs, at some generation. Following the same reasoning one can calculate the expectations of the products of allele frequencies p'_i at the next generation, conditional on the present allele frequencies $\{p\}$, to get:

$$\begin{aligned} E(p'_i p'_{i+k} | \{p\}) \quad = \quad & \frac{1}{2N}\, p_i p_{i+k} \delta_{0,k} \\ + \quad & (1-\frac{1}{2N})[\ (1-2\mu) p_i p_{i+k} \\ + \quad & \mu \sum_r m_r (\frac{1}{2} p_{i-r} p_{i+k} + \frac{1}{2} p_{i+r} p_{i+k}) \\ + \quad & \mu \sum_r m_r (\frac{1}{2} p_i p_{i+k-r} + \frac{1}{2} p_i p_{i+k+r})\ ]. \end{aligned} \tag{9}$$

Summing over all values of i introduces the frequencies

$$f_k = \sum_i p_i p_{i+k} \tag{10}$$

of pairs of alleles at distance k. Taking expectations in Eq. (9) provides the same recursion as in Eq. (8). This alternative approach will be used in the following to work on samples of 3 and 4 alleles.

This infinite set of Eq. (8) can be summarized into a single one, introducing the generating function of the series $P_g(k)$:

$$F(g,\alpha) = \sum_{k=-\infty}^{k=+\infty} \alpha^k P_g(k).$$

Multiplying the previous equations by α^k and summing up yields:

$$\begin{aligned} & F(g+1,\alpha) \\ & = \quad \frac{1}{2N(g)} \\ & + \quad (1-\frac{1}{2N(g)}) \left((1-2\mu)F(g,\alpha) + 2\mu \sum_r m_r (\frac{1}{2}\alpha^r + \frac{1}{2}\frac{1}{\alpha^r}) F(g,\alpha) \right) \\ & = \quad \frac{1}{2N(g)} + \left(1 - \frac{1}{2N(g)} - 2\mu(1-M(\alpha)) \right) F(g,\alpha), \end{aligned} \tag{11}$$

where $M(\alpha)$ describes the mutation process (Eq. 6). We obtain a linear recursion, similar to that describing the change with time of inbreeding coefficients, and enriched with the description of the spectrum of possible alleles carried by genes derived from a single one, g generations ago ($F(0,\alpha) = 0$).

Assuming that population size $N(g)$ remains large enough, we can turn to the continuous approximation, using the reduced time t (Eq. 5). Equation (11) becomes an ordinary differential equation with respect to time t,

for any value of α:

$$\frac{dF(t,\alpha)}{dt} = \frac{N_0}{N(t)} - \left(\frac{N_0}{N(t)} + \theta(1 - M(\alpha)) \right) F(t,\alpha), \tag{12}$$

where $N(t)$ is the population size at time t. The solution of this equation is the Fourier series based on the joint probabilities $Pr(D = k, T_{MRCA} < t)$:

$$F(t,\alpha) \;\; = \;\; \sum_k \alpha^k \, Pr(D = k, T_{MRCA} < t) \, ,$$

so that:

$$Pr(D = k, T_{MRCA} < t) \;\; = \;\; \frac{1}{2\pi} \int_{-\pi}^{\pi} F_R(t,x) \; cos(kx) \, dx \, , \tag{13}$$

where $F_R(t,x)$ stands for the real part of $F(t,\alpha)$.

Equation (12) gives access to the joint distribution of D and T_{MRCA}. We first solve this equation in special cases, then derive the distribution of the T_{MRCA} of two alleles conditional on their distance D and the marginal distribution of D.

2.3.1. Constant population size

For constant population size, $N(t) = N_0$, solving Eq. (12) with initial condition $F = 0$ gives:

$$F(t,\alpha) = \frac{1 - e^{-t(1+\theta(1-M(\alpha))}}{1 + \theta(1 - M(\alpha))}. \tag{14}$$

2.3.2. Variable population size

In order to deal with variable population size, we model the variations of $N(t)$ by a step function, or 'skyline plot' (Pybus *et al.* 2000; Drummond *et al.* 2005): let $t_0 = 0$ be the present time, $0 < t_1 < t_2 < ... < t_i < ... < t_I$ the

times in the past at which the population size underwent some change (Figure 2). In each interval (t_i, t_{i+1}) the population size is assumed to remain constant and equal to N_i, so that Eq. (12) can be easily integrated. This approach can be used to approximate any function $N(t)$ by an appropriate step function.

(Figure 2)

Assume that $t_j < t \leq t_{j+1}$.

One has to integrate the forward Eq. (12) along the intervals (t, t_j), (t_j, t_{j-1}),... (t_2, t_1) and $(t_1, 0)$. Integrating from past time t (at which we set $F = 0$) to past time t_j, we get:

$$F_j(t,\alpha) = \frac{N_0}{N_j} \frac{1 - \exp\left(-(t-t_j)(\frac{N_0}{N_j} + \theta(1 - M(\alpha)))\right)}{\frac{N_0}{N_j} + \theta(1 - M(\alpha))},$$

since the size remains equal to N_j in the interval. Integrating along the next interval (t_j, t_{j-1}) a similar integration yields, taking now F_j as the initial condition:

$$F_{j-1}(t,\alpha) = \frac{N_0}{N_{j-1}} \frac{1 - b_{j-1}}{a_{j-1}} + b_{j-1} F_j(t,\alpha), \tag{15}$$

where a's and b's are functions of α defined by:

$$a_i = \frac{N_0}{N_i} + \theta(1 - M(\alpha)) \tag{16}$$

$$b_i = \exp(-(t_{i+1} - t_i)\, a_i) \tag{17}$$

for $0 \leq i \leq I$ (noting that $t_{I+1} = \infty$ and $b_I = 0$). Going on until the present time $t_0 = 0$ is reached, one gets the value of $F(t, \alpha)$:

$for \qquad t_j < t \leq t_{j+1}$

$$F(t,\alpha) \qquad (18)$$
$$= \frac{1-b_0}{a_0} + b_0\left(\frac{N_0}{N_1}\frac{1-b_1}{a_1} + b_1\left(... + b_{j-2}\left(\frac{N_0}{N_{j-1}}\frac{1-b_{j-1}}{a_{j-1}} + b_{j-1}F_j(t,\alpha)\right)...\right)\right).$$

3. Coalescence times T_{MRCA} conditional on allele differences D

In order to derive properties of the distribution of T_{MRCA}, we introduce the Laplace transform of the function $F(t,\alpha)$, setting (for $u \geq 0$):

$$\begin{aligned}\mathcal{T}(\alpha,u) &= \int_0^\infty e^{-ut} \sum_{k=-\infty}^{k=+\infty} \alpha^k \, dPr(D=k, t < T_{MRCA} \leq t+dt) \\ &= \int_0^\infty e^{-ut} \, \frac{dF(t,\alpha)}{dt} dt.\end{aligned}$$

3.1. Constant population size

From Eq. (14), the Laplace transform is equal to:

$$\mathcal{T}(\alpha,u) = \frac{1}{1+u+\theta(1-M(\alpha))}\,.$$

From this expression, we can derive the Laplace transform $\mathcal{L}_k(u)$ of the distribution of T_{MRCA} conditional on the distance being equal to k. Taking the Fourier inverse we write:

$$\mathcal{T}(\alpha,u) = \sum_k \alpha^k \; H_k(u) = \sum_k \alpha^k \; H_k(0)\, \mathcal{L}_k(u)\,. \qquad (19)$$

In this expression, $H_k(0)$ is the probability $Pr(D=k)$. In the case of a symmetrical geometric model, one gets explicit solutions for $H_k(u)$ because $\mathcal{T}(\alpha,u)$ is a simple rationale expression in α, allowing the calculus of residues to be easily carried out. For the single step model ($M_R(x) = cos(x)$), one obtains:

$$\mathcal{T}(\alpha,u) = -\frac{2\alpha}{\theta}\frac{1}{(\alpha-\alpha_1(u))(\alpha-\alpha_2(u))}$$

with $0 < \alpha_1(u) < 1 < \alpha_2(u)$. The inversion involves only two residues, at 0 and $\alpha_1(u)$. Eventually, the result for Eq. (19) is given as:

$$H_k(u) = \frac{1}{\theta} \ \frac{1}{\Delta(u)} \ \ (\alpha_1(u))^{|k|} \tag{20}$$

where

$$\Delta(u) = \sqrt{\left(1 + \frac{1+u}{\theta}\right)^2 - 1} \ ; \ \alpha_1(u) \ = \ 1 + \frac{1+u}{\theta} - \Delta(u). \tag{21}$$

Setting u to 0 gives the values of $Pr(D = k)$ (Eq. 2).

Equations (20) and (21) show that $\mathcal{L}_k(u)$ is the product of $|k| + 1$ terms, which proves that the time T_{MRCA} , **conditional** on k, is decomposable into the sum of $|k| + 1$ independent variables:

$$T_{MRCA}|k = T_0 + T_1 + ... + T_{|k|}. \tag{22}$$

The Laplace transform of the first one is:

$$\mathcal{L}_0(u) \ = \ \frac{\sqrt{1+2\theta}}{\theta \ \Delta(u)}$$

and, if $|k| > 0$, the $|k|$ additional variables share the same distribution given by their transform:

$$\mathcal{L}_*(u) \ = \ (1 + \frac{1}{\theta} + \frac{1}{\theta}\sqrt{1+2\theta})\left(1 + \frac{1+u}{\theta} - \Delta(u)\right).$$

The densities of these distributions can be expressed using the I_k modified Bessel functions (Abramowitz and Stegun 1965, p 374; Feller 1971, pp 58, 503):

$$g_0(t) = \sqrt{1+2\theta} \ I_0(\theta t) \ e^{-(1+\theta)t}$$

for T_0 and

$$g_k(t) = (1 + \frac{1}{\theta} + \frac{1}{\theta}\sqrt{1+2\theta})^{|k|} \ \frac{|k|}{t} \ I_{|k|}(\theta t) \ e^{-(1+\theta)t}$$

for the sum $T_1 + ... + T_{|k|}$. Differentiating the Laplace transforms gives direct access to the moments of these conditional distributions, for example:

$$E(T_{MRCA}|k) = \frac{1}{2} + \frac{1}{2} \ \frac{1}{1+2\theta} \ + \ |k| \ \frac{1}{\sqrt{1+2\theta}}, \tag{23}$$

$$Var(T_{MRCA}|k) = \frac{1}{2} + \frac{1}{2} \ \frac{1}{(1+2\theta)^2} \ + \ |k| \ \frac{1}{\sqrt{1+2\theta}} \ \frac{1+\theta}{1+2\theta}. \tag{24}$$

These results (Eq. 21, 22, 23 and 24) can be extended to the symmetrical geometric model (7) with slight modifications through the replacement of θ by:

$$\theta_c = \frac{\theta(1+c)+2c}{(1-c)^2},$$

in the calculation of $\alpha_1(0)$ (Eq. 21) and by the addition of a weight for the case $D = 0$:

$$Pr(D = k) = \frac{\theta\,(1+c)}{\theta\,(1+c)+2c} \ \frac{1}{\sqrt{1+2\theta_c}} \ (\frac{1+\theta_c-\sqrt{1+2\theta_c}}{\theta_c})^{|k|}, \tag{25}$$

for $k \neq 0$, and:

$$Pr(D = 0) = \frac{2c}{\theta\,(1+c)+2c} + \frac{\theta\,(1+c)}{\theta\,(1+c)+2c} \ \frac{1}{\sqrt{1+2\theta_c}}. \tag{26}$$

Il can also be noted that an approximation can be derived from Eq. (2):

$$Pr(|D| \leq k) \simeq 1 - (1 - \frac{1}{\sqrt{2\,\theta}}) \ \exp(-\,k\,\sqrt{\frac{2}{\theta}}), \tag{27}$$

$(k \geq 0)$ which holds for large values of θ.

3.2. Variable population size

For variable population size, getting the conditional distribution is slightly more complex. First one notes in Eq. (18) that only the last term depends on time t, so that the derivative dF/dt takes a simpler expression:

$$\begin{array}{rcl} for \quad t_j & < & t \ \leq \ t_{j+1} \\ \dfrac{dF(t,\alpha)}{dt} & = & b_0 b_1 ... b_{j-1} \, \dfrac{N_0}{N_j} \, \exp\left(-(t-t_j)(\dfrac{N_0}{N_j} + \theta(1-M(\alpha)) \right). \end{array} \tag{28}$$

This allows the Laplace transform $\mathcal{T}(\alpha, u)$ to be given an explicit expression. Integrating in t from 0 to infinity involves the $I+1$ intervals of time, to get:

$$\mathcal{T}(\alpha,u) \ = \ \sum_{j=0}^{j=I} \psi_j \, \frac{N_0}{N_j} \, \frac{\exp(-ut_j) - b_j \exp(-ut_{j+1})}{u + a_j}, \tag{29}$$

where $\psi_0 = 1$ and $\psi_j = b_0 b_1 ... b_{j-1}$ for $j \geq 1$. From this, moments of the conditional distribution of T_{MRCA} can be obtained from the derivatives in u of $\mathcal{T}(\alpha, u)$. For example, from Eq. (19) we can write:

$$\frac{\partial \mathcal{T}}{\partial u}(\alpha, 0) = \sum_k \alpha^k \ H_k(0) \, \frac{d\mathcal{L}_k}{du}(0).$$

Then taking the Fourier inverse gives:

$$\begin{array}{rcl} E(T_{MRCA}|D=k) \ Pr(D=k) & = & H_k(0) \, \dfrac{d\mathcal{L}_k}{du}(0) \\ & = & \dfrac{1}{2\pi} \displaystyle\int_{-\pi}^{\pi} \frac{\partial \mathcal{T}_R}{\partial u}(x,0) \, cos(kx) \, dx \end{array} \tag{30}$$

Obtaining the result requires in general to carry out a numerical integration; this holds for the moments of the distribution as well as for the distribution itself using Eq. (13) with F defined by Eq. (18).

(Table 1)

(Figures 3)

Effects of a systematic change in population size on the conditional cumulative distributions of T_{MRCA} are illustrated in Figure 3, considering exponential changes. Table 1 gives the probabilities of the different conditions illustrated in Figure 3. Figure 4 shows the conditional expectations (Eq. 30) under the same conditions. Figure 5 is similar to Figure 4, for scenarii involving a transient reduction of population size (botteleneck).

4. Distribution of distances between alleles in a finite sample

The preceding analysis gives a complete description of the distributions involved with two alleles. To get insights into the distribution of larger samples of alleles we look at the distribution of distance frequencies between alleles and at the possible extensions to samples of several alleles.

Numerical calculations of probabilities $Pr(D = k)$ are obtained from Eq. (18) with $t = \infty$ (so that $j = I$ for the case of variable population size), or from Eq. (29) at $u = 0$.

4.1. Within population distribution of distances between alleles

The distribution of absolute differences $|D|$ in repeat numbers between alleles and of their squares allow global measures of diversity to be defined, for example the stepwise weighted genetic distance (D_{sw}) of Shriver *et al.* (1995), and the Average Squared Distance (ASD) of Goldstein *et al.* (1995). Another way to characterize the genetic diversity is to consider the frequencies of pairs of alleles that show a distance equal to k, the "P_k *distribution*"

considered by Shriver *et al.* (1997). We use here the notation f_k (Eq. 10), noting that the P_k value of Shriver *et al.* (1997) is equal to $2\,f_k$ for $k > 0$, so that : $f_0 + 2\sum_{k>0} f_k = 1$.

In practice, such measures are calculated from a sample of alleles drawn from the population. However, we will only consider here their population values, as if based on an exhaustive sampling. At the population level, the within population measures of diversity corresponding to D_{sw} and ASD can be written as:

$$\mathcal{D}_1 = \sum_{k>0} 2\,k\,f_k \;\; ; \;\; \mathcal{D}_2 = \sum_{k>0} 2\,k^2\,f_k.$$

Their expectations depend on the probabilities that two alleles are at some distance:

$$\begin{aligned} E(\mathcal{D}_1) = E(|D|) &= \sum_{k>0} k\,Pr(|D| = k)\,, \\ E(\mathcal{D}_2) = E(D^2) &= \sum_{k>0} k^2\,Pr(|D| = k)\,. \end{aligned}$$

For a constant population size, Eq. (25) and (26) allow exact results to be obtained for the geometric model:

$$\begin{aligned} E(\mathcal{D}_1) = E(|D|) &= \frac{1}{\sqrt{1+2\theta_c}} E(D^2)\,, \\ E(\mathcal{D}_2) = E(D^2) &= \frac{(1+c)}{(1-c)^2}\theta\,. \end{aligned} \tag{31}$$

A generalization to any mutation model is:

$$E(D^2) = \theta\,M^{(2)}(1).$$

where $M^{(2)}(\alpha)$ is the second derivative of the $M(\alpha)$ function (Kimmel and Chakraborty 1996).

Obtaining the variance of these measures (their variance over a set of genetically independent loci) needs considering second moments, hence probabilities involving four alleles drawn from the same population. Expressions for the variance of $\mathcal{D}_2$ were obtained by Pritchard and Feldman (1996) and by Kimmel and Chakraborty (1996), giving:

$$Var(\mathcal{D}_2) = \frac{4}{3}\,\theta^2 + \frac{1}{3}\,\theta \tag{32}$$

for the single step SMM in a population of constant size. We consider in the next section the general calculation of the second moments of frequencies f_k's.

(Tables 2 and 3)

Imbalance indices.

Global measures of diversity have been used to search for indication of population size changes. Using the expressions relating the theoretical parameter θ to observable quantities provide various estimates of it: θ may be estimated from homozygosity Hom (Eq. 1), or from $\mathcal{D}_2$ (Eq. 31 with $c = 0$):

$$\begin{aligned} \hat{\theta_0} &= \frac{1}{2}(\frac{1}{Hom^2} - 1), \\ \hat{\theta_2} &= \mathcal{D}_2. \end{aligned}$$

This suggested using an *imbalance index* defined as the ratio of both estimates - or its natural logarithm:

$$i2 = \ln(\hat{\theta_2}) - \ln(\hat{\theta_0}),$$

to check for departure from constant size (Reich *et al.* 1999; King *et al.* 2000; Bobrowski and Kimmel 2004). Using the $\mathcal{D}_1$ measure suggests another

estimate of θ,

$$\hat{\theta_1} = \mathcal{D}_1(\mathcal{D}_1 + \sqrt{\mathcal{D}_1^2 + 1})$$

and allows another imbalance index to be defined:

$$i1 = \ln(\hat{\theta_1}) - \ln(\hat{\theta_0}).$$

Given the mutation rate, the three measures: homozygosity, $\mathcal{D}_1$ and $\mathcal{D}_2$ provide three values $N(0)$, $N(1)$ and $N(2)$ of population size. Effects of population size changes on imbalance indices are illustrated in Table 2 for the cases of exponential increase or decrease of population size as considered in Figure 3. Effects of a recent or ancient bottleneck are outlined in Table 3. Two mutation rates were considered (0.001 and 0.0001) in these tables.

(Figures 4 and 5)

The "P_k *distribution*"

The previous results allow the effects of population size changes on the full "P_k *distribution*" to be considered. One can generalize the idea of imbalance indices, characterizing the shape of this distribution by estimates of θ - or of effective size - for each value of the distance between alleles. Under the single step SMM (Eq. 2), all ratios $Pr(|D| = k+1)/Pr(|D| = k)$ are equal to $\alpha_1(0)$ (Eq. 21) and linked to θ according to this equation. For a general P_k *distribution*, each ratio $Pr(|D| = k+1)/Pr(|D| = k)$ can be used to give a special value of θ (or of population size) that depends on k.

Results are shown in Figures 4 and 5 for cases of population size changes. In Figure 4, only conditions such that $Pr(|D| = k) > 0.01$ are considered, as

in Table 1.

4.2. Correlations between pairs of alleles

The frequencies of pairs of alleles at distance k, as defined by the f_k's (Eq. 10), are not independent variables. The expected value of a product $f_k\, f_s$ is equal to the probability that among the four alleles A, B, C and D considered in two pairs, the first two alleles A and B are at distance k and the last two alleles C and D are at distance s. The two events considered in this joint probability are not independent since the four alleles belong to the same population and derive from a common ancestor. Such relationships generate correlations between these f_k frequencies.

These correlations can be experimentally observed in a survey involving several marker loci sharing the same mutation process. Getting their expected values provides the variability of the diversity measures between loci, which permits determining the number of independent loci needed to get estimates of parameters with some precision, or to test departure from a demographic hypothesis with some confidence.

The way to get Eq. (8) from the expectation of frequencies f_k (Eq. 9 and 10) can be generalized to samples of four alleles, allowing the covariances between two frequencies f_k and f_s to be derived. As in Eq.(9), let p_i be the frequency of alleles with i motifs at some generation, and p'_i their frequency at the next generation. In this next generation, the product $f'_k\, f'_s$ can be written as:

$$f'_k\, f'_s = \sum_i \sum_j p'_i\, p'_{i+k}\, p'_j\, p'_{j+s}.$$

In each term of the sum, the four alleles may derive either from 3 different

alleles of the previous generation (there are 6 combinations, each with probability $1/2N$) or from 4 different alleles (probability $= 1 - 6/2N$), assuming that no more than 1 coalescence event may happen. We obtain:

$$\begin{aligned} E(p'_i\, p'_{i+k}\, p'_j\, p'_{j+s}/\{p\}) = \frac{1}{2N} \quad & (\quad \delta_{k,0}\, p_i\, p_j\, p_{j+s} + \delta_{i,j}\, p_i\, p_{i+k}\, p_{j+s} \\ & \qquad + \ four\ similar\ terms \quad) \\ + \ (1 - \frac{6}{2N}) \quad & (\quad (1-4\mu)\, p_i\, p_{i+k}\, p_j\, p_{j+s} \\ & \qquad + \ \mu\ [\sum_r m_r (\frac{1}{2}\, p_{i-r}\, p_{i+k}\, p_j\, p_{j+s} + \frac{1}{2}\, p_{i+r}\, p_{i+k}\, p_j\, p_{j+s}) \\ & \qquad\qquad + \ three\ similar\ terms\ ]\ ). \end{aligned}$$

Summing in i and j introduces the values of f_k, f_s, f_{k+r}, of their products $f_k\, f_s$, $f_k\, f_{s+r}$,... in the previous generation and sums of products involving three genes, such as $\sum_i\ p_i\, p_{i+k}\, p_{i+s}$. Taking expectations gives a series of equations in the expected values of products, $\mathcal{H}_{k,s} = E(f_k\, f_s)$, and introduces the quantities:

$$\mathcal{G}_{k,s} \quad = \quad E(\sum_i p_i\, p_{i+k}\, p_{i+s}).$$

$\mathcal{G}_{k,s}$ is the probability that in a sample of three alleles A, B and C, the distances are equal to k between A and B and to s between A and C. As before, we turn to the generating functions ($\alpha = e^{ix}$, $\beta = e^{iy}$; x, y real):

$$\begin{aligned} H(\alpha,\beta) \quad &= \quad \sum_{k=-\infty}^{k=+\infty} \sum_{s=-\infty}^{s=+\infty} \alpha^k \beta^s\, \mathcal{H}_{k,s}, \\ G(\alpha,\beta) \quad &= \quad \sum_{k=-\infty}^{k=+\infty} \sum_{s=-\infty}^{s=+\infty} \alpha^k \beta^s\, \mathcal{G}_{k,s}, \end{aligned}$$

and write the change of G and H after one generation:

$$G'(\alpha,\beta) \quad = \quad (1 - \frac{3}{2N} - 3\mu) G(\alpha,\beta)$$

$$
\begin{aligned}
&+ \quad \mu\,(M(\alpha)+M(\beta)+M(\alpha\beta))\,G(\alpha,\beta)\\
&+ \quad \frac{1}{2N}\,(F(\alpha)+F(\beta)+F(\alpha\beta)),
\end{aligned}
$$

$$
\begin{aligned}
H'(\alpha,\beta) &= (1-\frac{6}{2N}-4\mu)H(\alpha,\beta)\\
&+ \quad \mu\,(2M(\alpha)+2M(\beta))\,H(\alpha,\beta)\\
&+ \quad \frac{1}{2N}\,(F(\alpha)+F(\beta)+2G(\alpha,\beta)+2G(\alpha,\frac{1}{\beta})).
\end{aligned}
$$

In the case of a constant population size, the ultimate values of $H(\alpha,\beta)$, $G(\alpha,\beta)$ and $F(\alpha)$ are given by:

$$
\begin{aligned}
\bar{H}(\alpha,\beta) &= \frac{1}{6}\,\frac{\bar{F}(\alpha)+\bar{F}(\beta)+2\bar{G}(\alpha,\beta)+2\bar{G}(\alpha,1/\beta)}{1+\frac{\theta}{3}(1-\frac{1}{2}(M(\alpha)+M(\beta))}\\
\bar{G}(\alpha,\beta) &= \frac{1}{3}\,\frac{\bar{F}(\alpha)+\bar{F}(\beta)+\bar{F}(\alpha\beta))}{1+\frac{\theta}{2}(1-\frac{1}{3}(M(\alpha)+M(\beta)+M(\alpha\beta))}\\
\bar{F}(\alpha) &= \frac{1}{1+\theta(1-M(\alpha))}
\end{aligned}
$$

The symmetry ($\mathcal{H}_{k,s}=\mathcal{H}_{|k|,|s|}$) allows the inversion formula to be simplified as:

$$
E(f_k\,f_s)=\mathcal{H}_{k,s}=\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\bar{H}_R(x,y)\,cos(kx)\,cos(sy)\,dx\,dy,
$$

where $\bar{H}_R(x,y)$ stands for the real part of $\bar{H}$.

This provides the way to obtain the correlation structure of the distance frequencies f_k by numerical integration. However, since $\bar{H}(\alpha,\beta)$ is a rationale expression in α and in β for the geometric mutation model, it would be possible to get an analytical expression of $E(f_k\,f_s)$. More easily, the variance of $\mathcal{D}_2$ can be derived analytically from the equilibrium value $\bar{H}$. Noting that:

$$
\mathcal{D}_2^2=\sum_{k=-\infty}^{k=+\infty}\sum_{s=-\infty}^{s=+\infty}k^2\,s^2\,f_k\,f_s\,,
$$

the expectation of the square of $\mathcal{D}_2$ becomes:

$$\begin{aligned} E(\mathcal{D}_2^2) &= \sum_{k=-\infty}^{k=+\infty} \sum_{s=-\infty}^{s=+\infty} k^2\, s^2\, E(f_k\, f_s) \\ &= \frac{\partial^4 \bar{H}}{\partial\alpha^2 \partial\beta^2}(1,1). \end{aligned}$$

Differentiating, and setting $\alpha = \beta = 1$, we get:

$$E(\mathcal{D}_2^2) = \frac{7}{3}\,(\theta\, M^{(2)}(1))^2 + \frac{1}{3}\,\theta\,(4\, M^{(2)}(1) + 5\, M^{(3)}(1) + M^{(4)}(1)), \qquad (33)$$

where $M^{(2)}$, $M^{(3)}$ and $M^{(4)}$ are the second, third and fourth derivatives of the M function. Under the single step model ($M(\alpha) = \frac{1}{2}(\alpha + 1/\alpha)$), this leads to Eq. (32).

(Figure 6)

Figure 6 shows the structure of the correlation matrix, for some values of the θ parameter.

4.3. Distance frequencies among 3 or 4 alleles

The previous analysis showed that obtaining the variances and covariances of distance frequencies f_k needed that probabilities of states among three and four alleles are known. We already introduced the quantities

$$\mathcal{G}_{k,s} = E(\sum_i p_i\, p_{i+k}\, p_{i+s})$$

which give the frequencies of triplets of alleles showing distance differences of k and s. Deriving the corresponding recursion formula followed the same line as for pairs of alleles, noting that three sampled alleles may be copies of

two or three alleles in the previous generation. Considering the variances and covariances of f_k's led to the introduction of special fourth order moments. More generally, one can derive the expectation of sums involving 4 alleles:

$$\mathcal{K}_{k,s,v} = E(\sum_i p_i\, p_{i+k}\, p_{i+s}\, p_{i+v})$$

to get the expected frequency of quadruplets of alleles showing distance differences equal to k, s and v. Setting

$$K(\alpha,\beta,\gamma) = \sum_{k=-\infty}^{k=+\infty}\sum_{s=-\infty}^{s=+\infty}\sum_{v=-\infty}^{v=+\infty} \alpha^k\beta^s\gamma^v E(\sum_i p_i\, p_{i+k}\, p_{i+s}\, p_{i+v}),$$

the same reasoning as before leads to the following recursion for the transform K between two generations:

$$\begin{aligned} K'(\alpha,\beta,\gamma) &= (1-\frac{6}{2N}-4\mu)\,K(\alpha,\beta,\gamma) \\ &+ \mu\,(M(\alpha)+M(\beta)+M(\gamma)+M(\alpha\beta\gamma))K(\alpha,\beta,\gamma) \\ &+ \frac{1}{2N}\,(\,G(\beta,\gamma)+G(\alpha,\gamma)+G(\alpha,\beta) \\ &\quad + G(\alpha\beta,\gamma)+G(\alpha\gamma,\beta)+G(\alpha,\beta\gamma)\,). \end{aligned}$$

Combining this equation with those in F and in G (previous sections) and rewriting as differential equations, we have the complete system:

$$\begin{aligned} \frac{dF(t,\alpha)}{dt} &= \frac{N_0}{N(t)} - \left(\frac{N_0}{N(t)} + \theta(1-M(\alpha))\right) F(t,\alpha) \\ \frac{dG(t,\alpha,\beta)}{dt} &= \frac{N_0}{N(t)}(F(t,\alpha)+F(t,\beta)+F(t,\alpha\beta)) \\ &- 3\left(\frac{N_0}{N(t)} + \frac{\theta}{2}(1-\frac{M(\alpha)+M(\beta)+M(\alpha\beta)}{3})\right) G(t,\alpha,\beta) \\ \frac{dK(t,\alpha,\beta,\gamma)}{dt} &= \frac{N_0}{N(t)}(\,G(t,\beta,\gamma)+G(t,\alpha,\gamma)+G(t,\alpha,\beta) \\ &\quad + G(t,\alpha\beta,\gamma)+G(t,\alpha\gamma,\beta)+G(t,\alpha,\beta\gamma)\,) \\ &- 6\left(\frac{N_0}{N(t)} + \frac{\theta}{3}(1-\frac{M(\alpha)+M(\beta)+M(\gamma)+M(\alpha\beta\gamma)}{4})\right) K(t,\alpha,\beta,\gamma). \end{aligned}$$

In addition, one may write the equation in H for the expectations of products of frequencies:

$$\begin{aligned}\frac{dH(t,\alpha,\beta)}{dt} &= \frac{N_0}{N(t)}(\,F(t,\alpha)+F(t,\beta)+2G(t,\alpha,\beta)+2G(t,\alpha,\frac{1}{\beta}) \\ &- 6\left(\frac{N_0}{N(t)}+\frac{\theta}{3}(1-\frac{M(\alpha)+M(\beta)}{2})\right)\,H(t,\alpha,\beta).\end{aligned}$$

Although cumbersome, it would be possible to get an explicit solution of this system if population size changes are modelled by a step function. Solving it between two times 0 and t, starting from initial F, G and K zero values, would provide the joint probability that a set of 4 alleles shows distance differences of k, s and v and that their T_{MRCA} is smaller than t.

Retrieving the values of $\mathcal{G}_{k,s}$ and $\mathcal{K}_{k,s,v}$ is obtained from the inversion of $G(t,\alpha,\beta)$ and $K(t,\alpha,\beta,\gamma)$. The symmetry $\mathcal{G}_{k,s}=\mathcal{G}_{-k,-s}$ makes it possible to write the inversion formula as:

$$\mathcal{G}_{k,s}(t) = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} G_R(t,x,y)\,cos(kx+sy)\,dx\,dy,$$

and allows all values to be expressed from those defined with $(s,k)=(0,0)$; $(0,k), k>0$ (with 6 equiprobable combinations); $(0,k,2k), k>0$ (6 combinations) and: $0<s<k$ $(s<k/2)$ (12 combinations). Similarly, the $\mathcal{K}_{k,s,v}$ values need a triple integral:

$$\mathcal{K}_{k,s,v}(t) = \frac{1}{8\pi^3}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K_R(t,x,y,z)\,cos(kx+sy+vz)\,dx\,dy\,dz.$$

5. Discussion

We have obtained results that allow the distribution of coalescent times between microsatellite alleles to be better known, analytically in the case

of constant population size, and through numerical integration in more general cases when population size has changed in the past. In the first case, we showed that the T_{MRCA} conditional on $D = k$ was the sum of $|k| + 1$ independent terms (Eq. 22). Although restricted to the case of constant population size, this qualitative behaviour seems quite robust to population size changes (Fig. 4 and 5) except for the first few terms if the population has undergone a strong and recent size decrease. From Eq. (23) and (24), the ratio of the difference between expectations of two successive distributions ($E(T_{MRCA}|k+1) - E(T_{MRCA}|k)$), to the standard error of ($T_{MRCA}|k$) is a decreasing function of k and of θ, taking its maximum ($= 1/\sqrt{2}$) for small θ and approximately equal to: $1/\sqrt{\theta + k\sqrt{\theta/2}}$ for large θ. While the mean and variance of conditional variables ($T_{MRCA}|k$) increase with k, their distributions become more and more overlapping, so that distances between alleles provide weak information on the structure of the coalescence tree of a set of alleles.

In the case of population expansion (Figure 3-A, 3-B, 3-C) coalescence times tend to show an upper limit, as far as only common situations are considered (such that $Pr(|D| = k) > 0.01$). In parallel, at least for fast expansion (Figure 3-B and 3-C), all the conditional T_{MRCA} show a minimal value linked to the recent expansion that makes recent coalescence unlikely, even for homozygous pairs of alleles ($T_{MRCA}|0$). In the case of a regular (exponential) reduction of population size (except for very low rate of decrease, Figure 3-D), one observes a mixture of the conditional distributions (Figure 3-E and 3-F): pairs of alleles with small distances (0, 1, ...) are the most common and show a small T_{MRCA}, while for larger k values the distributions are

quite similar to those expected for the large ancestral population size. This may give access to the ancestral size from the distribution of the frequencies f_k, for largest k, provided the mutation rate is large enough for making these large k values observable or the number of markers is large enough for rare events to be observed. On the contrary, a very slow decrease (Figure 3-D) would not be observable since almost all coalescence events occurred when the population size was small. We discuss in the following how these features could become observable, using our results on the "P_k *distribution*".

In previous works on microsatellite diversity, focus has been mainly given to the value of homozygosity (the value of $Pr(D = 0)$, Eq. 1) and to global indicators of population expansion such as the imbalance index (Reich *et al.* 1999; King *et al.* 2000). Table 2 shows how the expected values of indices behave, so that $i2 < i1 < 0$ in the case of a growing population, and $i2 > i1 > 0$ if population size has undergone a reduction in the past. Tables 2 et 3 show how these measures are less sensible to very fast growth or to very slow decrease (Table 2-C and 2-D). Effects of a recent or ancient bottleneck are outlined in Table 3, showing that except for a recent and severe bottleneck (Table 3-B) the effects of a transient bottleneck are not clear, since indices are low. In these Tables, two mutation rates were considered (0.001 and 0.0001) and showed the greater efficiency of more variable markers to detect variations of population size. However, a statistical analysis remains to be carried out to check the efficiency of these indices to detect reduction of population size. In addition to these indicators, we have also considered the complete distribution of distance frequencies to check how it depended on demography. Extending the idea of imbalance index is illustrated in Figures

4 and 5 (curves with gray squares). The results suggest that, depending on the rate of decrease of population size, it is more or less feasible to access the size of the ancestor population (Figure 4-E, 4-F) or impossible if the decrease is very slow: in the case of Figure 4-D, only the current size is accessible because it is unlikely to observe alleles distant enough for their T_{MRCA} to be representative of the ancestor population. On the other hand, only rather slow increase of population size seems detectable (Figure 4-A, 4-B), while accessing the value of the ancestral population size remains unlikely. Effects of a bottleneck could be observed with the same criterion, mainly through an increase of the expected homozygosity relative to the rate of decrease of the probabilities $Pr(D = k)$ (Figure 5-E). This effect is extended to the first values of k if the bottleneck is strong enough (Figure 5-A, 5-B). One notes that the effect of a recent bottleneck is similar to a medium or very fast regular decrease (Figures 4-E and 5-B, 4-F and 5-A). If the bottleneck is more ancient (Figure 5-C, 5-D), only strong ones should be detectable. In addition Figure 5-D indicates that such a bottleneck prevents one to access the size of the ancestor population, even if the analysis suggests that the population had suffered a transient decline of its size. Assuming a mutation rate lower than that used in Figures 3, 4 and 5 ($\mu = 0.001$) makes the signals less clear (Tables 2 and 3, second lines calculated with $\mu = 0.0001$). For example, the effect of a strong recent bottleneck (Figure 5-B) would look like the effect expected from a short bottleneck with higher mutation rate (Figure 5-A). With a lower mutation rate, accessing the ancestor population size would need a larger number of markers.

The previous results were concerned only with expected values. Extend-

ing the analysis to samples of 3 and 4 alleles allowed us to derive the covariances between pairwise distance frequencies and the variance of the $\mathcal{D}_2$ (Eq. 33) for a general mutation model. As seen in Figure 6, where the coefficients of correlation are given for values of k and s up to about $3\sqrt{1+2\theta}$, the correlation matrix shows a strong structure independent of θ, after normalization. In particular, negative covariances occur between, on the one hand the frequency f_0 (and frequencies f_k with small k, when θ is large), and frequencies f_s, with s of the order of $\sqrt{1+2\theta}$. This may correspond to the approximation given by Eq. (27). The variances of distance frequencies f_k decrease with k, in a fashion which is approximately exponential, except for the smallest values of k. An empirical approximation can be proposed from the numerical calculations:

$$Var(f_k) \;\simeq\; 0.053 \;\; \theta^{-1.14} \;\; \exp(-\,k\,\sqrt{\frac{2}{\theta}}) \;\; .$$

This strong covariance structure should be important to be taken into account in order to develop statistical analysis aimed at detecting population size variations based of the "P_k *distribution*", in the same way as the variance of $\mathcal{D}_2$ was necessary to get the power of tests based on the imbalance index. If the pairwise distance frequencies f_k are obtained from a finite sample of alleles, additional terms must be added to the expression given above for the second moments of $\mathcal{D}_2$ and of the f_k's. Their calculation makes use of the general probabilities $\mathcal{K}_{k,s,v}$ defined for subsets of 4 alleles (not shown).

As far as numerical integrations are needed, there is no limit to the choice of a mutation model, provided the mutation rate of an allele does not depend on its size and there is no constraint on the number of microsatellite motifs. The assumption of symmetry in our setting (Eq. 6) is in fact not needed,

as quoted by Kimmel and Chakraborty (1996). On the contrary, any dependence of the mutation rate on the size of the allele (Watkins, 2006) or constraints on the range should need a different approach. The recent analysis of Sun *et al.* (2009) suggests however that ignoring range constraints does not seem a critical assumption, even working for rather large generation numbers. Changes in population size was modelled using only the 'skyline plot', because it allows analytical derivations of the generating functions (Eq. 28 and 29) and limits the need for numerical integrations to the last step in order to recover probabilities. Modelling changes of size by piecewise linear demographic functions is an interesting alternative developed by Heled and Drummond (2008) in the same context. When considering linear trends in our framework, integrating Eq. (12) along linear pieces leads to recursions involving modified partial Gamma integrals, instead of the simple explicit formulae (15), (16) and (17). It seems more efficient to split a linear piece into successive steps, then apply expressions (28) and (29). Our results are restricted to samples of 2, 3 and 4 alleles. Extending the approach to any sample size does not seem feasible, for two reasons. First, the usual assumption that zero or only one coalescence event may happen each generation is not valid for a large sample size (the assumption is that the square of the sample size is small with respect to the effective size of the population). Second, even with low values of the sample size, huge combinatorial problems arise. As a consequence, the joint distribution of the numbers of alleles of each type cannot be given a concise analytical expression for current sample sizes.

Recently, a similar work was developed concerning the analysis of the

number of segregating sites, corresponding to the Infinite Sites Model for mutations, which can be convenient to deal with clusters of tightly linked Single Nucleotide Polymorphism (SNP) markers (Notohora and Umeda 2006). These authors developed also the joint distribution of the coalescence time and of the number of segregating sites, and extended their results to cope with the effect of population structure (island model, stepping stone model). The present analysis devoted to microsatellite markers would also deserve such extensions. In particular, its extension to some versions of the stepping stone model (infinite one- or two- dimensional discrete lattice), should be technically feasible with the use of Fourier-like transforms in the spatial variable.

References

Abramowitz, M., Stegun, I., 1965. Handbook of mathematical functions. Electronic release (http://www.math.sfu.ca/ cbm/aands/toc.htm).

Bobrowski, A., Kimmel, M., 2004. Asymptotic behavior of joint distributions of characteristics of a pair of randomly chosen individuals in discrete-time Fisher-Wright models with mutations and drift. Theor. Popul. Biol. 66, 355-367.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E. Kidd, J.R., Cavalli-Sforza, L.L., 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455-457.

Caballero, A., 1994. Developments in the prediction of effective population size. Heredity 73, 657-679.

Charlesworth, B., Charlesworth, D., Barton, N. H., 2003. The effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. Syst. 34, 99-125.

Chevalet, C., Gillois, M., Nassar, R., 1977. Identity coefficients in finite populations : Evolution of identity coefficients in a random mating dioecious population. Genetics 86, 697-713.

Drummond, A. J., Rambaut, A., Shapiro, B., Pybus, O. G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185-1192.

Ewens, W. J., 1982 On the concept of effective population size. Theor. Popul. Biol. 21, 373-378.

Feller, W., 1971. An Introduction to Probability Theory and its Applications, Volume II (second edition). John Wiley & Sons, New York.

Felsenstein, J., 1971. Inbreeding and variance effective numbers in populations with overlapping generations. Genetics 68, 581-597.

Gillois, M., 1964. La relation d'identité en génétique. Thèse, Université

de Paris, 294p.

Gillois, M., 1965. Relation d'identité en génétique. Ann. Inst. Henri Poincaré B, 2, 1-94.

Goldstein, D. B., Ruiz-Linares, A., Cavalli-Sforza, L. L., Feldman, M. W., 1995. An Evaluation of Genetic Distances for Use With Microsatellite Loci. Genetics 139, 463-471.

Gregorius, H. R., 1991 On the concept of effective number. Theor. Popul. Biol. 40, 269-283.

Heled, J., Drummond, A. J., 2008. Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology 8, 289.

Kimmel, M., Chakraborty, R., 1996. Measures of Variation at DNA Repeat Loci under a General Stepwise Mutation Model Theor. Popul. Biol. 50, 345-367.

Kimmel, M., Chakraborty, R., Stiverst, D. N., Deka, R., 1996. Dynamics of Repeat Polymorphisms Under a Forward-Backward Mutation Model: Within- and Between-Population Variability at Microsatellite Loci. Genetics 143, 549-555.

Kimura, M, Weiss, G. H., 1964. The Stepping Stone Model of Population

Structure and the Decrease of Genetic Correlation with Distance. Genetics 49, 561-576.

King, J. P., Kimmel, M., Chakraborty, R., 2000. A power analysis of microsatellite-based statistics for inferring past population growth. Mol. Biol. Evol. 17, 1859-1868.

Malécot, G., 1951. Un traitement stochastique des problèmes linéaires (mutations, linkage, migration) en génétique de population. Annales Université de Lyon, section Sciences 14, 79-117.

Malécot, G., 1975. Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. 8, 212-241.

Malécot, G., 1981. Evolution, Parentés, Migrations. in : Chevalet C, Micali A, eds. Modèles mathématiques en biologie, Lecture Notes in Biomathematics 41, 95-119.

Notohara, M., Umeda, T., 2006. The coalescence time of sampled genes in the structured coalescent model Theor. Popul. Biol. 70, 289-299.

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

Polanski, A., Bobrowski, A., Kimmel, M., 2002. A note on distributions of times to coalescence, under time-dependent population size. Theor. Popul. Biol. 63, 33-40.

Pritchard, J. K., Feldman, M. W., 1996. Statistics for Microsatellite Variation Based on Coalescence. Theor. Popul. Biol. 50, 325-344.

Pybus, O. G., Rambaut, A., Harvey, P. H., 2000. An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. Genetics 155, 1429-1437.

Reich, D. E., Feldman, M. W., Goldstein, D. B., 1999. Statistical properties of two tests that use multilocus data sets to detect population expansions. Mol. Biol. Evol. 16, 453-466.

Reich, D. E., Goldstein, D. B., 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA 95, 8119-8123.

Shriver, M. D., Jin, L., Boerwinkle, E., Deka, R., Ferrell, R. E., Chakraborty, R., 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol. Biol. Evol. 12, 914-920.

Shriver, M. D., Jin, L., Ferrell, R. E., Deka, R., 1997. Microsatellite data support an early population expansion in Africa. Genome Res. 7, 586-591.

Sun, J. X., Mullikin, J. C., Patterson, N., Reich D. E., 2009. rosatellites are molecular clocks that support accurate inferences about history. Mol. Biol. Evol. 26, 1017-1027.

Wang, J. L., 2005. Estimation of effective population sizes from data on genetic markers. Phil. Trans. R. Soc. Lond. B Biol. Sci. 360, 1395-1409.

Waples, R. S., 2005. Genetic estimates of contemporary effective population size: to what time periods do the estimates apply. Mol. Ecol. 2005 14, 3335-3352.

Watkins, J. C., 2006. Likelihood-Based Estimation of Microsatellite Mutation Rates. Theor. Popul. Biol. 71, 147-159.

Wehrhahn, C. F., 1975. The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics 80, 375-394.

Whitlock, M. C., Barton, N. H., 1997. The effective size of a subdivided population. Genetics 146, 427-441.

Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., Sibly, R. M., 2003. Likelihood-Based Estimation of Microsatellite Mutation Rates. Genetics 164, 781-787.

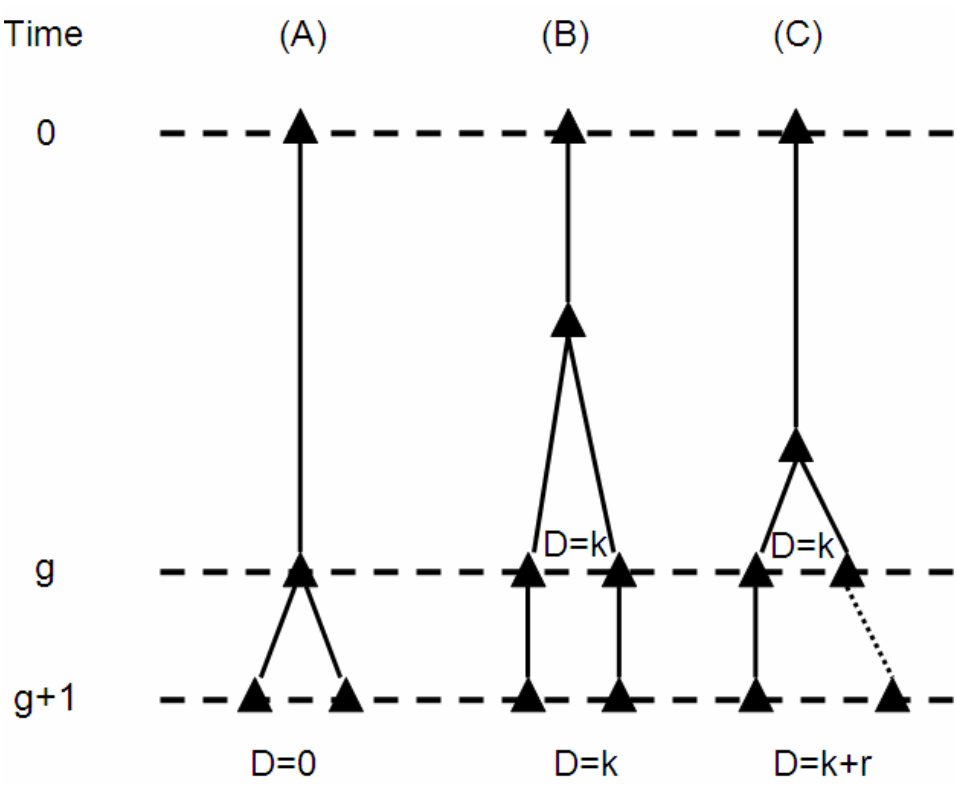


Figure 1. Possible origins of a pair of alleles.

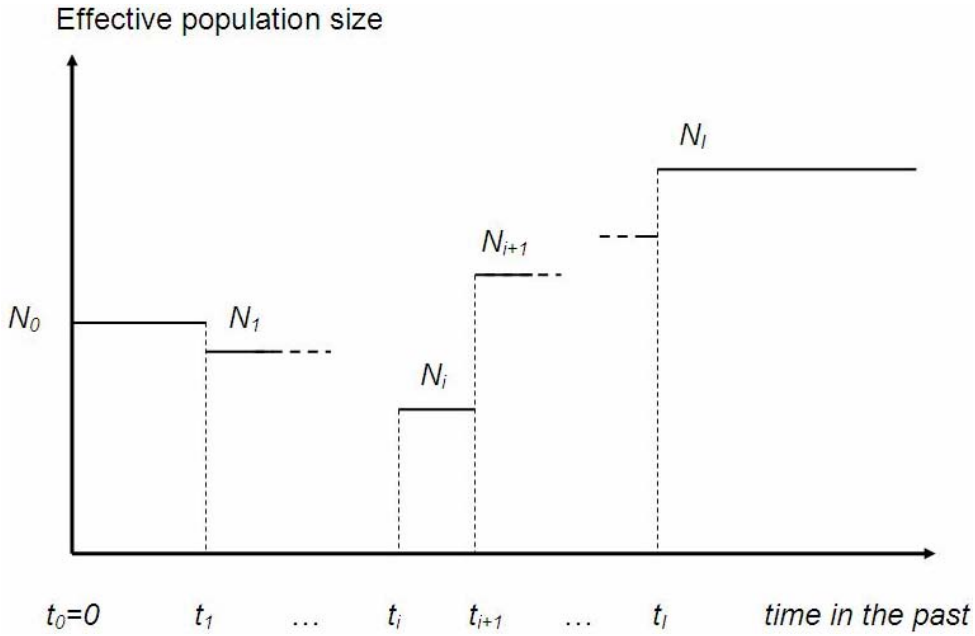


Figure 2. Modelling variations of population size in the past.

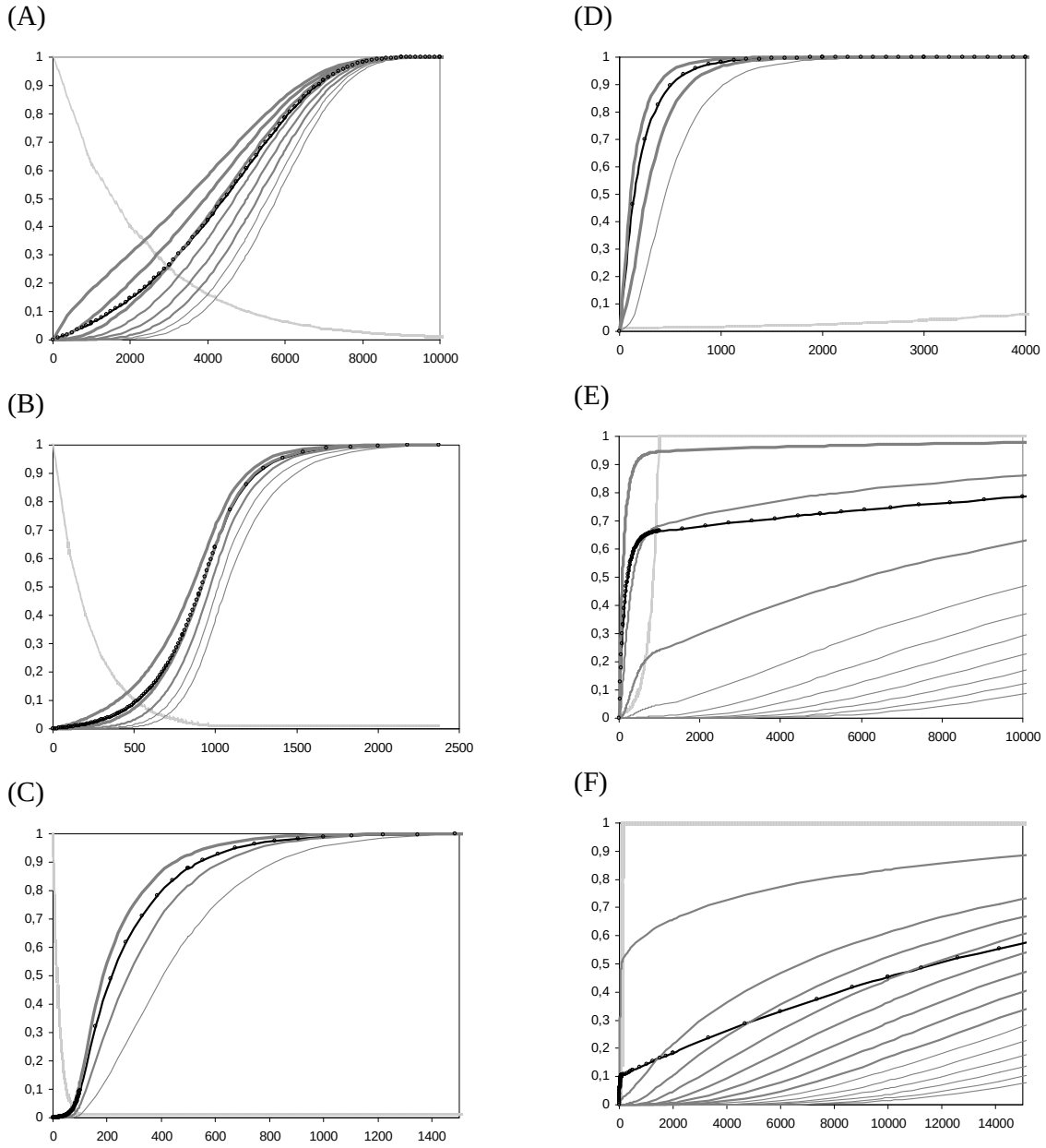


Figure 3. Cumulative distribution of the Time to the Most Recent Common Ancestor, conditional on the difference *D* between alleles, after exponential growth or decrease of population size.

Definition of the six considered cases:

A, B, C: exponential growth from N=100 to N=10000 with 10000 (A), 1000 (B) or 100 (C) generations and a mutation rate at μ=0.001.

D, E, F: exponential decrease from N=10000 to N=100 with 10000 (A), 1000 (B) or 100 (C) generations and a mutation rate at μ=0.001.

Abscissa: time in the past (number of generations).

Ordinates: straight lines are the conditional distributions for increasing values of *D* from left to right. Only values of *k* such that $Pr(|D|=k)>0.01$ are reported (see Table 1); black lines with points are the unconditioned distribution; light gray lines are the population size divided by 10000 (the largest values).

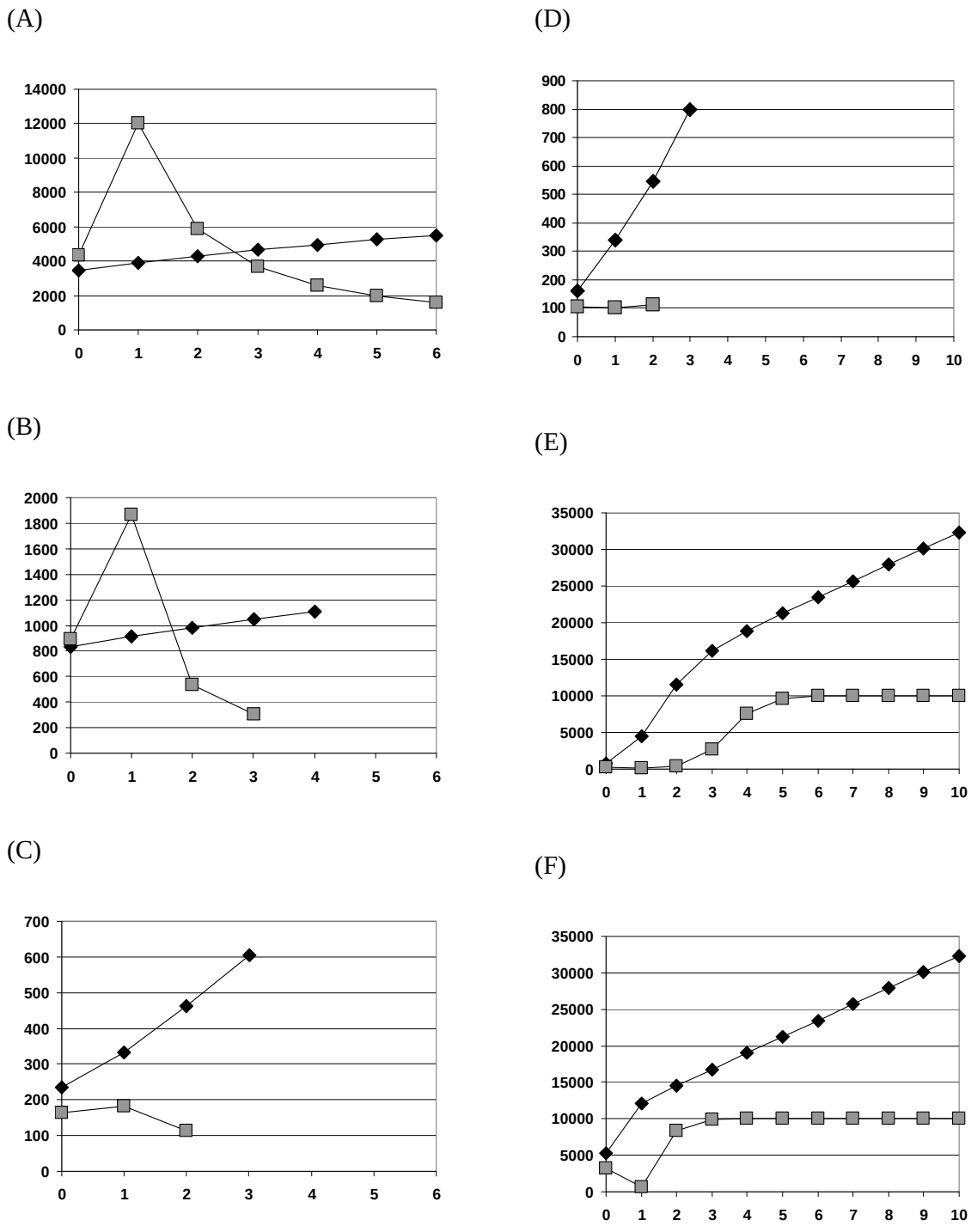


Figure 4. Effects of population size changes on the expected values of T_{MRCA} and on effective size conditional on the distance D between alleles. Cases A to F as defined in Figure 3. Only conditions such that $Pr(|D|=k)>0.01$ are reported.
Abscissa: distance k between two alleles.
Ordinates: black diamonds are expected values of T_{MRCA} conditional on k (number of generations in the past); gray squares are effective population size; derived from the ratio f_{k+1}/f_k (see the text).

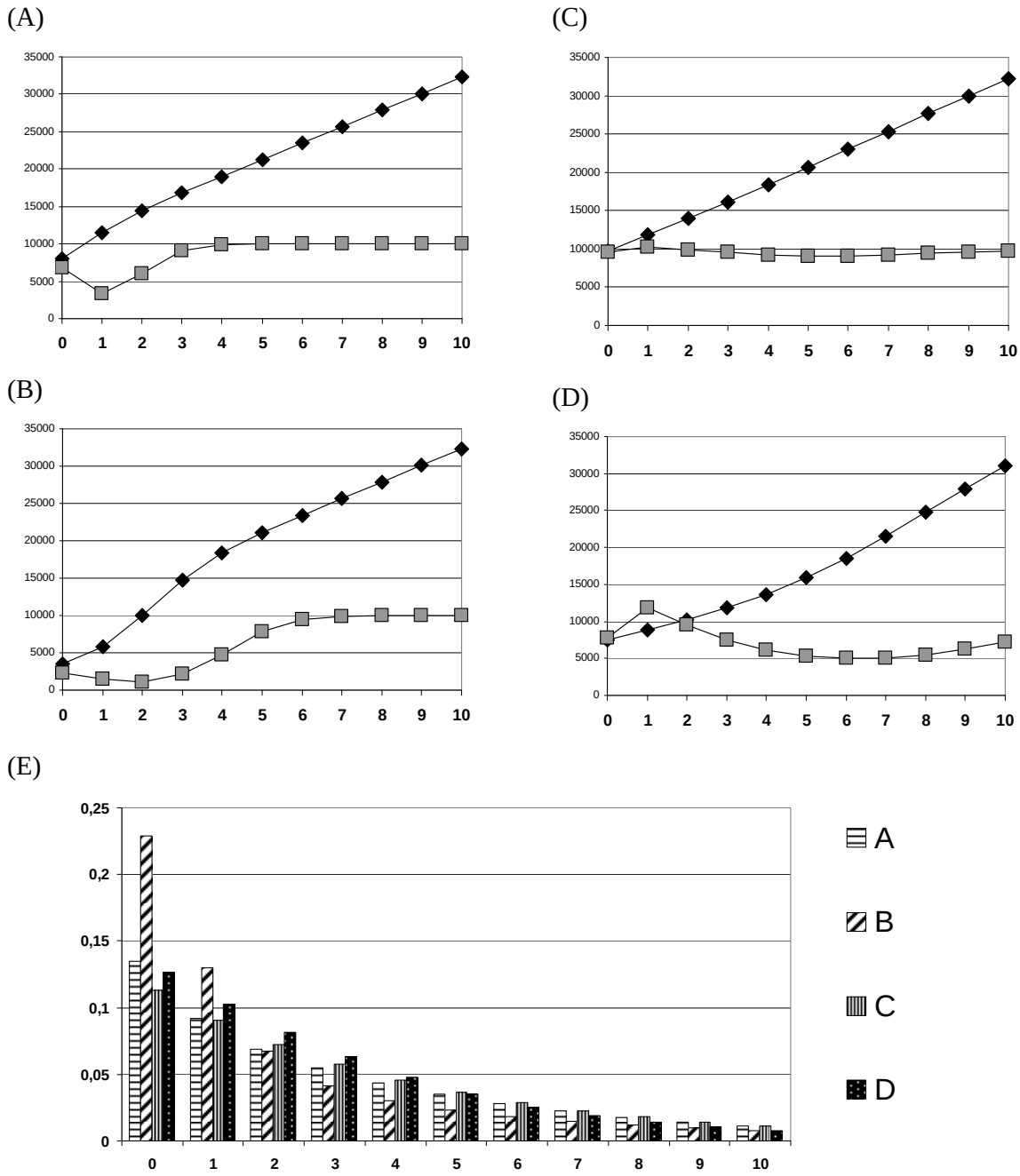
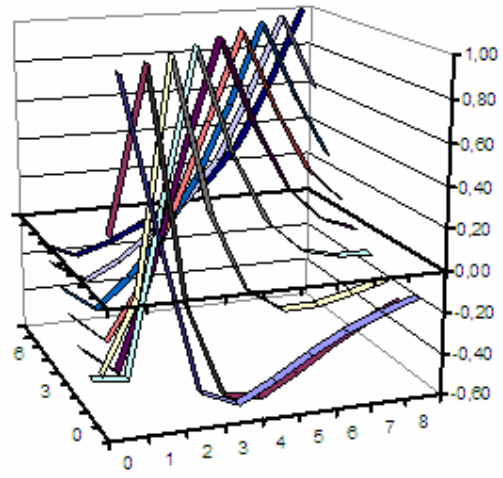
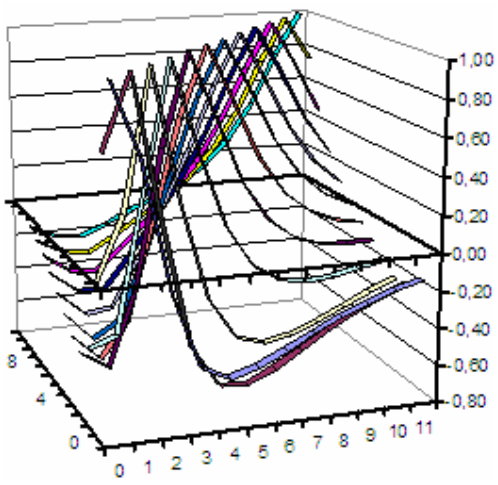


Figure 5. Effects of a bottleneck on the expected values of T_{MRCA}, on effective size conditional on the distance *D* between alleles and on the frequencies f_k.
Figures 5-A, 5-B, 5-C, 5-D: Abscissa and Ordinates as in Figure 4.
Figure 5-E: abscissa is the distance *k* between alleles; ordinate is the frequencies f_k.
Definition of the four considered cases:
Present and ancient population sizes are N_0= N_2=10000 in the four cases. The mutation rate is set to μ=0.001.
Cases A and B: recent bottleneck with reduced population size set to N_1=1000 from generations 200 to 300 (A) or from generations 200 to 1200 (B) in the past.
Cases C and D: ancient bottleneck with population size set to N_1=1000 from generations 5000 to 5100 (C) or from generations 5000 to 6000 (D) in the past.

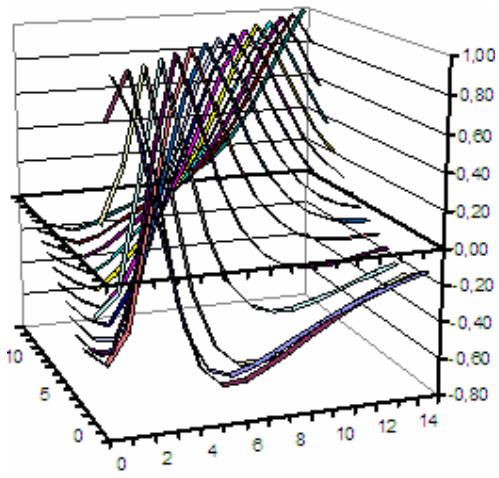
(A)



(B)



(C)



(D)

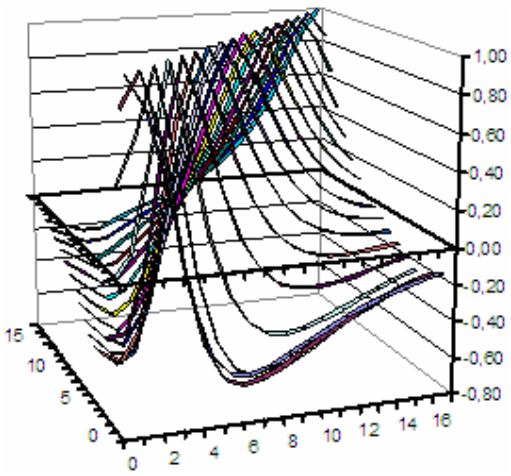


Figure 6. Coefficients of correlation between pairwise distances of allele frequencies f_k and f_s under constant population size.
Horizontal coordinates: the distances *k* and *s* between alleles.
Ordinates: the coefficients of correlations $corr(f_k, f_s)$. For each value of *k* (the left coordinate), the Figure shows the values $corr(f_k, f_s)$, *s= 0,1,...* as a function of *s* (the right coordinate).
Definition of the four considered cases:
(A): $\theta = 4$, distances from *0* to *8*
(B): $\theta = 8$, distances from *0* to *11*
(C): $\theta = 12$, distances from *0* to *14*
(D): $\theta = 16$, distances from *0* to *16*

Table 1. Expected frequencies of pairs of alleles at distance (± *k*) in the six cases considered in Figure 3. Only the frequencies larger than 0.01 are shown. The last line gives the total of remaining smaller frequencies (f_k).

k	A	B	C	D	E	F
0	0.1673	0.3500	0.6565	0.7373	0.5536	0.1947
± 1	0.2730	0.4196	0.2892	0.2203	0.1790	0.1638
± 2	0.2041	0.1655	0.0463	0.0350	0.0642	0.1284
± 3	0.1414	0.0496			0.0419	0.1026
± 4	0.0916	0.0122			0.0324	0.0821
± 5	0.0557				0.0258	0.0657
± 6	0.0320				0.0206	0.0525
± 7	0.0175				0.0165	0.0420
± 8					0.0132	0.0336
± 9					0.0106	0.0269
± 10						0.0215
± 11						0.0172
± 12						0.0138
± 13						0.0110
Total smaller f_k	0.0174	0.0032	0.0080	0.0074	0.0422	0.0441

Table 2. Expected imbalance indices (*i*) in growing (A, B, C) or decreasing (D, E, F) populations with a mutation rate at μ=0.001 (First line) and at μ=0.0001 (Second line).
Cases A to F as in Figure 3.

	i1	*i2*		*i1*	*i2*
A	-0,519	-0,698	**D**	0,019	0,045
	-0,200	-0,356		0,003	0,009
B	-0,455	-0,680	**E**	1,733	2,538
	-0,046	-0,115		0,679	1,279
C	-0,086	-0,170	**F**	0,952	1,052
	-0,007	-0,019		0,227	0,308

Table 3. Expected imbalance indices (*i*) after a recent (A, B) or ancient (C, D) bottleneck with a mutation rate at μ=0.001 (First line) and at μ=0.0001 (Second line).
Cases A to D as in Figure 5.

	i1	*i2*		*i1*	*i2*
A	0,311	0,347	**C**	-0,007	0,001
	0,101	0,134		0,009	0,018
B	0,789	1,057	**D**	-0,107	-0,066
	0,460	0,723		0,007	0,055

DONNÉES

COMPLÉMENTAIRES

Cette partie est dédiée aux données qui n'ont pas été présentées au sein du troisième article intitulé *« Distribution of coalescence times and of distances between microsatellite alleles with changing effective population size »*. Le premier paragraphe concerne des données complémentaires sur l'étude des temps de coalescence *T*, selon la distance *D* entre allèles ; le second paragraphe concerne des détails mathématiques supplémentaires à l'article pour une meilleure compréhension de certains passages d'une équation à l'autre.

3.1. *T* selon la distance *D*

Dans cet article ont été présentés les cumuls des *P(T|D,N,M)* pour seulement *D* égal à zéro (*P(T|D=0,N,M)*), car les distributions se comportent de la même façon sur tous les *D*, comme vous pouvez le voir sur les figures ci-dessous représentant les résultats de *P(T|D,N,M)* avec différents scénarios et un taux de mutation M à 0.0003 (taux de mutation trouvé pour les populations étudiées par les 37 marqueurs nucléaires microsatellites). Du scénario A à D, l'intensité du goulot d'étranglement est de plus en plus forte sur la durée (A par rapport à B, puis C à D) et en intensité (A par rapport à C, puis B à D). Les distributions *P(D|N,M,T)* traduisent l'histoire de la population puisque le goulot d'étranglement est perceptible et se traduit par une diminution de la variabilité du à l'augmentation des processus de coalescence (fusion) entre allèles au moment du goulot.

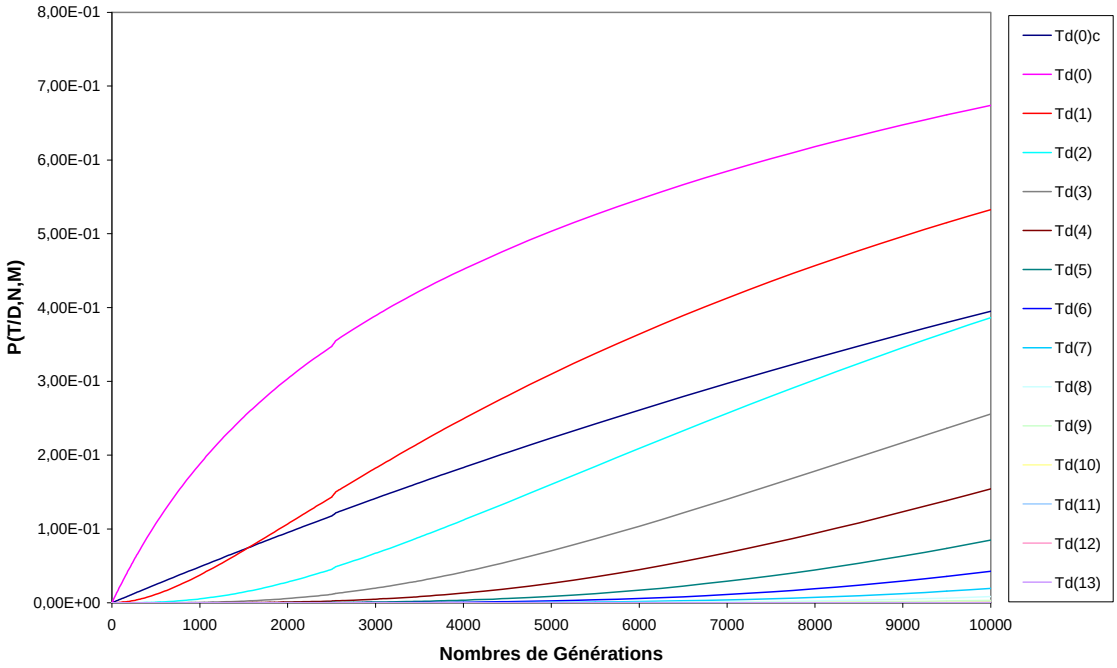


Figure III-22. *Distribution des temps de coalescence T conditionnés par D, N et M pour le scénario A. Le scénario A retrace une population ayant subi un goulot d'étranglement d'un rapport 1/2 (passage de 10 000 individus à 5000) pendant 50 générations: N0 et N2=10 000, N1=5000, et l'intervalle T1-T2=50.*

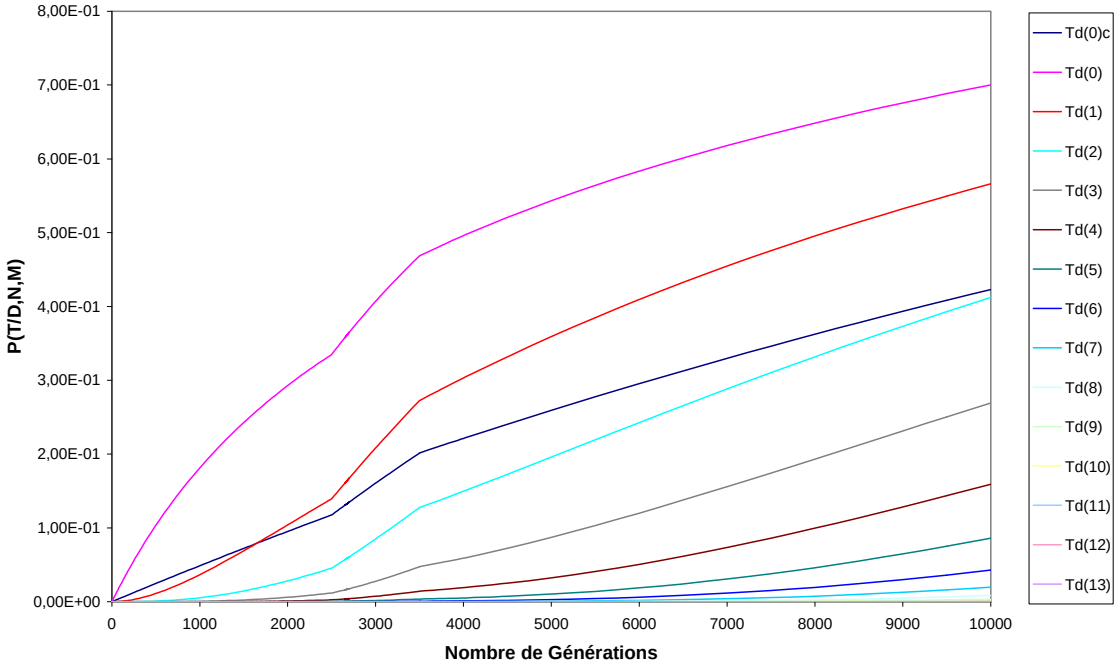


Figure III-23. *Distribution des temps de coalescence T conditionnés par D, N et M pour le scénario B. Le scénario B retraçant une population ayant subi un goulot d'étranglement d'un rapport 1/2 (passage de 10 000 individus à 5000) pendant 1000 générations: N0 et N2=10 000, N1=5000, et l'intervalle T1-T2=1000.*

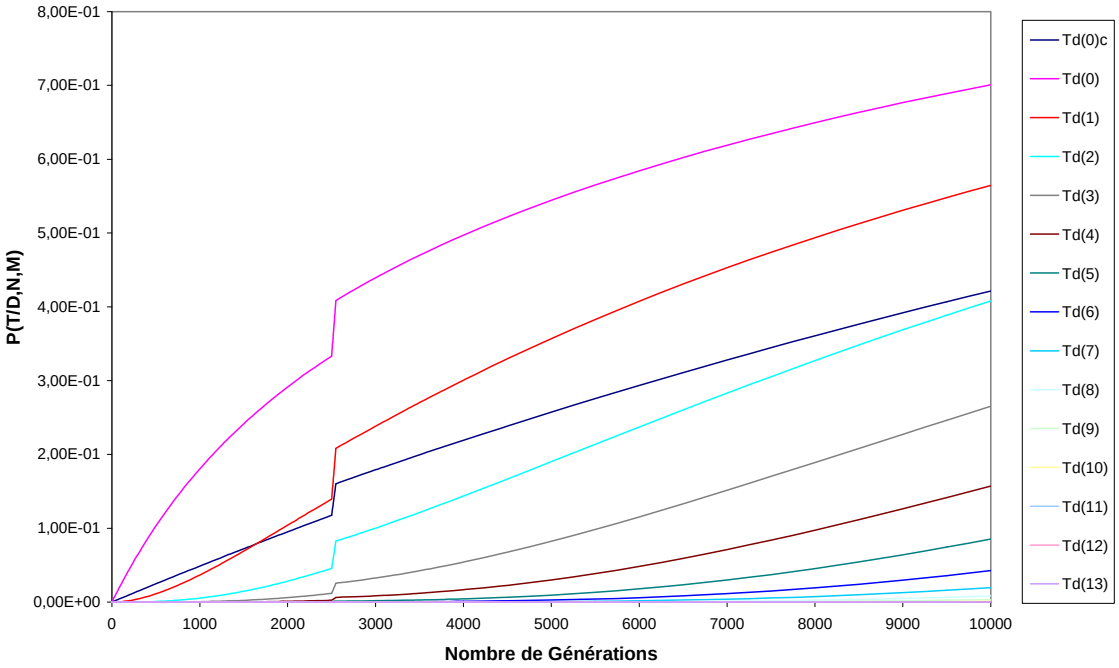


Figure III-24. *Distribution des temps de coalescence T conditionnés par D, N et M pour le scénario C. Le scénario C retraçant une population ayant subi un goulot d'étranglement d'un rapport 1/20 (passage de 10 000 individus à 500) pendant 50 générations: N0 et N2=10 000, N1=500, et l'intervalle T1-T2=50.*

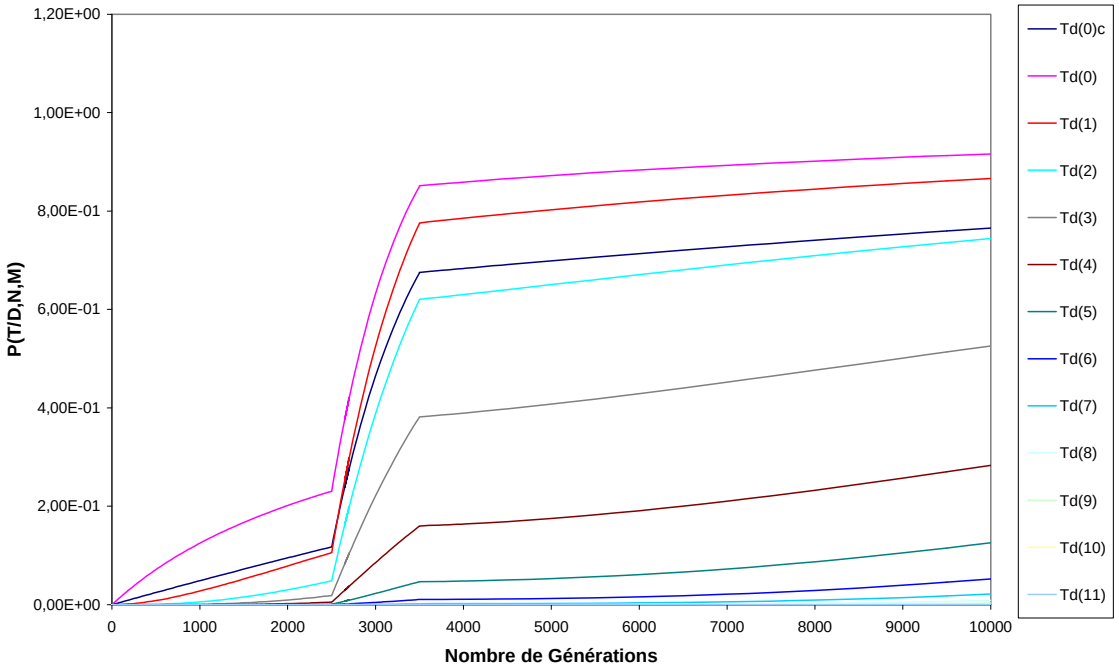


Figure III-25. *Distribution des temps de coalescence T conditionnés par D, N et M pour le scénario D. Le scénario D retraçant une population ayant subi un goulot d'étranglement d'un rapport 1/20 (passage de 10 000 individus à 500) pendant 1000 générations: N0 et N2=10 000, N1=500, et l'intervalle T1-T2=1000.*

3.2. Détails mathématiques supplémentaires d'une équation à l'autre

Les numéros des équations sont ceux correspondant à l'article III de ce mémoire.

3.2.1. Passage de l'équation 8 à 10

Équation 8 :

$$F(g+1,\alpha) = \frac{1}{2N(g)} + \left(1 - \frac{1}{2N(g)} - 2\mu(1-M(\alpha))\right) * F(g,\alpha).$$

Développement :

$$F(g+1,\alpha) = \frac{1}{2N(g)} + \left(1 - \frac{1}{2N(g)} - 2\mu + 2\mu M(\alpha)\right) * F(g,\alpha).$$

Comme on a dans l'idée de faire apparaître θ, on met en facteur $\frac{1}{2N(g)}$

$$F(g+1,\alpha) = \frac{1}{2N(g)} + \left(1 - \frac{1}{2N(g)}(1 + 4N(g)\mu(1-M(\alpha)))\right) * F(g,\alpha).$$

On introduit le temps réduit ($t = g / 2N_0$) en écrivant *dF/dt*, avec $dt = 1/2N_0$, qui revient à écrire la différentielle des *F*

$$\frac{F(g+1,\alpha) - F(g,\alpha)}{\frac{1}{2N_0}} = 2N_0\left(\frac{1}{2N(g)} - \frac{1}{2N(g)}(1 + 4N(g)\mu(1-M(\alpha)))\right) * F(g,\alpha),$$

$$\frac{F(g+1,\alpha) - F(g,\alpha)}{\frac{1}{2N_0}} = \frac{N_0}{N(g)} - \left(\frac{N_0}{N(g)} + \theta(1-M(\alpha))\right) * F(g,\alpha).$$

Equation 10 :

$$\frac{dF(t,\alpha)}{dt} = \frac{N_0}{N(t)} + \left(\frac{N_0}{N(t)} + \theta(1-M(\alpha))\right) * F(t,\alpha).$$

3.2.2. Passage de l'équation 10 à 12 (taille constante)

Équation 10 :

$$\frac{dF(t,\alpha)}{dt} = \frac{N_0}{N(t)} - \left(\frac{N_0}{N(t)} + \theta(1-M(\alpha))\right) * F(t,\alpha).$$

Développement :

Comme nous travaillons sur taille constante, l'équation 10 $N_0 = N(t)$ devient une équation avec second membre constant :

$$\frac{dF(t,\alpha)}{dt} = 1 - \left(1 + \theta(1 - M(\alpha))\right) * F(t,\alpha),$$

$$F'(t,\alpha) + \left(1 + \theta(1 - M(\alpha))\right) * F(t,\alpha) = 1.$$

Pour passer au premier membre, nous la considérons égale à zéro

$$F'(t,\alpha) + \left(1 + \theta(1 - M(\alpha))\right) * F(t,\alpha) = 0,$$

$$\frac{F'(t,\alpha)}{F(t,\alpha)} = -\left(1 + \theta(1 - M(\alpha))\right).$$

Pour simplifier, nous écrivons

$$1 + \theta(1 - M(\alpha)) = a.$$

sachant que

$$\log F = -at + K \quad (K \text{ étant une constante}),$$

ce qui donne

$$F = K\exp(-at) = \frac{1}{a} + K\exp(-at),$$

et comme

$$K = -\frac{1}{a},$$

nous obtenons

$$F = \frac{1 - \exp(-at)}{a}.$$

Équation 12 :

$$F(t,\alpha) = \frac{1 - \exp\left(-t(1 + \theta(1 - M(\alpha))\right)}{1 + \theta(1 - M(\alpha))}.$$

3.2.3. Passage de l'équation 10 à 13 (taille variable)

Nommons dans l'équation 10, $a = \frac{N_0}{N(t)}$ et $b = \frac{N_0}{N(t)} + \theta(1 - M(\alpha))$:

$$\frac{dF(t,\alpha)}{dt} = a + b * F(t,\alpha)\,.$$

Le premier membre de cette différentielle s'écrit alors :

$$\overline{F(t,\alpha)} = -a/b\,,$$

et le second membre avec la variable muette τ :

$$\frac{dF(\tau,\alpha)}{d\tau} + bF(\tau,\alpha) = o\,,$$

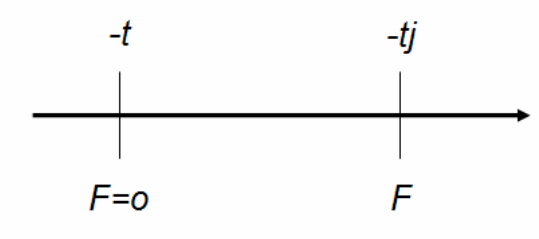
$$\frac{dF(\tau,\alpha)}{d\tau} = -bF(\tau,\alpha)\,.$$

Lorsque l'on intègre sur le second membre, on obtient :

$$\ln(F(\tau,\alpha)) = -b\tau + K\,,$$

$$F(\tau,\alpha) = K\exp(-b\tau)\,.$$

Si on se place tel que la représentation ci-dessous :



La constante vaut :

$$F(t,\alpha) = \frac{a}{b} + K\exp(bt) = o\,,$$

$$K = -\frac{a}{b}\exp(-bt)\,.$$

Ce qui donne la solution :

$$F(tj,\alpha) = \frac{a}{b} - \frac{a}{b}\exp(-bt) * \exp(btj)\,,$$

$$F(tj,\alpha) = \frac{a}{b} - (1 - \exp(-b(t - tj))\,.$$

IV

DISCUSSION GÉNÉRALE
ET
CONCLUSION

Face à la croissance économique mondiale amenuisant les ressources vivantes utilisées et les palliers trophiques sous-jacents, certaines solutions peuvent être trouvées pour protéger et conserver ces ressources dès lors que nous pouvons établir des constats d'état. Les évolutions démographiques sont régulièrement utilisées pour révéler l'état de ces ressources. Cependant ces estimations peuvent être difficiles d'accès dans des milieux de grandes dimensions, avec peu de visibilité et/ou hostiles. C'est le cas des stocks des espèces halieutiques. Pour pallier ces difficultés, différentes techniques ont vu le jour pour prévoir leurs chutes et anticiper les crises économiques. Mais il n'y a que très peu de temps que l'estimation des stocks a été reliée au concept de biodiversité. Le terme de biodiversité est apparu en 1985 sous la plume de Walter G. Rosen (Meine et al. 2006). Il désigne la diversité des organismes vivants à plusieurs échelles, avec la diversité des gènes, des espèces ou encore de l'organisation et de la répartition des écosystèmes. Avec la prise de conscience de l'extinction des espèces au cours des dernières décennies, il est très utilisé depuis 1986. Il laisse transparaître que la survie de l'espèce humaine dépend de la diversité des autres organismes vivants et que l'on ne peut pas se détacher du milieu qui nous entoure. Rien n'est cloisonnable, tout est complexement lié. Nous ne pouvons pas ainsi négliger qu'estimer cette biodiversité seulement par le dénombrement d'espèces ou d'individus ait une grande part d'incertitude.

Pour pallier cette part d'incertitude, les paramètres reflétant la taille et la viabilité d'une population peuvent s'avérer d'une grande utilité. Ces paramètres sont pour la plupart issus de l'information contenue dans le matériel génétique. Le matériel génétique, code gérant le développement et l'organisation d'un individu, évolue suivant l'histoire de la population à laquelle il appartient et de ses capacités d'adaptation au milieu environnant. De ce fait, il peut être utilisé pour retracer l'histoire d'une population, accéder à sa viabilité et son évolution dans l'espace-temps. Parmi les paramètres génétiques reflétant cette viabilité, se trouve la taille efficace (*Ne*). Elle est régulièrement mentionnée comme étant primordiale en conservation car elle donne une idée du niveau d'abondance et de l'état de santé d'une population. Vu sous le regard évolutionniste, l'état de santé représente la capacité de réaction et donc d'adaptation d'une population face à un évènement soudain dans son milieu. La taille efficace est liée aux taux de perte de diversité génétique, de fixation d'allèles délétères et de l'efficacité de la sélection naturelle à maintenir les allèles avantageux (Berthier et al. 2002). Si cette taille efficace subit une diminution trop élevée, la perte de variabilité génétique résultant de la force de dérive génétique devrait mettre l'espèce ou la population en risque d'extinction,

puisque le matériel nécessaire sur lequel opère la sélection est perdue. *Ne* peut donc être un outil important pour accéder au statut de vulnérabilité des populations (Mace & Lande 1991). Des tailles efficaces minimales ont été proposées pour maintenir les populations. *Ne* = 50 a été suggéré pour minimiser les effets négatifs de la dépression consanguinité et *Ne* = 500 pour conserver suffisamment de possibilités d'évolution (Franklin 1980; Lande 1988; Franklin & Frankham 1998; Lynch & Lande 1998). Le rôle des facteurs génétiques dans l'extinction des espèces a été controversé notamment depuis le papier de Lande (1988). Néanmoins récemment, les preuves de la contribution des facteurs génétiques dans le risque d'extinction ont été regroupées et confirmées (Frankham 2005). Parmi ces preuves, on retrouve la dépression consanguine et la perte de la diversité génétique comme facteurs augmentant le risque d'extinction. L'auteur confirme même que si la caractérisation génétique n'est pas incluse dans les projets de conservation pour évaluer l'état des populations, le risque d'extinction de la population ou de l'espèce sera sous-estimé, et les stratégies de restauration inappropriées. Estimer l'état des stocks à partir d'un paramètre génétique tel que la taille efficace semble être indispensable pour appréhender le devenir d'une population.

Malgré les progrès en biologie moléculaire, en statistique et en bioinformatique, l'estimation de la taille efficace continue de rencontrer des difficultés face aux bases génotypiques hétérogènes et aux forces évolutives qu'elle fait intervenir. Lorsque l'on cherche à intégrer la taille efficace dans un projet de conservation d'une espèce en danger, un travail préliminaire doit être effectué sur la sensibilité des modèles vis-à-vis des données génétiques, de la diversité et la structure des populations. C'est ce que nous avons cherché à faire dans la première partie de ce travail sur le Saumon atlantique, espèce inscrite dans la liste rouge des espèces en danger en Europe (Porcher & Baglinière 2001). Pour ce faire, un effort dans la construction des données génétiques, du choix des populations et des analyses s'est imposé.

Parmi les séquences nucléiques reflétant au mieux la variabilité d'une population, utilisée pour évaluer la taille efficace, nous retrouvons majoritairement les séquences non codantes. Notre choix s'est porté sur les marqueurs nucléaires microsatellites, car ce sont des marqueurs neutres très polymorphes, qui ont fait leur preuve en génétique quantitative et des populations. Ces séquences, dites également silencieuses, permettent généralement d'éviter de prendre en considération les effets de la sélection dans les modèles. Les autres forces agissant sur la variabilité génétique sont la dérive, les mutations et/ou la migration. Le phénomène du

homing chez le Saumon laisse à penser que la migration peut être négligée. Il était donc naturel de tester des estimateurs de *Ne* négligeant la migration chez cette espèce.

Plusieurs modèles ont été proposés pour estimer *Ne*, mais la méthode la plus utilisée reste celle basée sur la différence des fréquences alléliques entre deux échantillons de la même population, à quelques générations d'écart, appelée la méthode des moments. La tendance actuelle tend vers des modèles plus complexes qui cherchent à retracer la généalogie des gènes de l'échantillon pour remonter à l'ancêtre commun. Ces modèles nommés les modèles de coalescence dérivent de la théorie de l'identité par descendance de Malécot (1941) développée par Kimura & Crow (1964) et généralisée par Kingman (1982) pour l'appliquer à un échantillon de plusieurs gènes pour lesquels on considère un ancêtre commun. Cette reconstruction d'arbre de gènes permet d'estimer l'espérance de *Ne* entre l'ancêtre commun et la population actuelle. Trois méthodes de coalescence négligeant la migration ont donc été choisies pour être comparées. La première (TM3) utilise deux échantillons alors que les deux autres (MSVAR et DIY ABC) utilisent un seul échantillon par population. Parmi ces deux dernières, l'une fonctionne sous statistique d'approximation bayésienne (ABC), alors que les deux autres recourent au calcul de vraisemblance dans une approche bayésienne.

La comparaison empirique de méthodes d'analyses génétiques nécessite un travail préliminaire sur le panel de marqueurs génétiques sélectionnés. Parce que beaucoup de marqueurs microsatellites ont été développés chez le Saumon atlantique, il n'était pas utile d'en développer de nouveaux. Cependant, il paraissait indispensable de proposer un panel optimal pour les études en génétique des populations parmi la plus grande base comprenant plus de 850 marqueurs. Pour cela, la carte génétique du Saumon atlantique fournie par Bjorn Hoyheïm (données non publiées) a été utilisée pour sélectionner des marqueurs indépendants, c'est-à-dire espacés sur les bras chromosomiques, polymorphes, et fortement utilisés dans des projets européens tels qu'ASAP. Cette sélection a fourni un panel de 73 marqueurs microsatellites qui a ensuite été testé sous amplification M13 (Schuelke 2000). Sur ces 73 marqueurs, un panel de 37 s'est révélé facile d'amplification, polymorphe et optimal pour les analyses en diversité et différenciation génétique chez le Saumon atlantique. Cette étude a également confirmé que l'analyse de nos populations nécessitait un minimum de 25-28 marqueurs polymorphes pour une bonne différenciation des populations. Malgré cette constatation, le panel entier des 37 marqueurs a été utilisé dans les analyses de comparaison

des trois méthodes de coalescence, pour pouvoir analyser leur comportement, en fonction du nombre et de la variabilité des marqueurs utilisés.

Les données empiriques du Saumon atlantique ont été établies à partir de quatre populations Européennes sauvages et anadromes qui font l'objet de suivis biologiques ou d'études sur leur fonctionnement et leur évolution sur le long terme. Ces quatre populations (Oir et Scorff au Nord-Ouest de la France, et Spey et Shin au Nord-Est de l'Ecosse) ont été choisies en raison de leurs différences de niveau d'abondance et de statut, avec deux vulnérables (Françaises) et deux en préoccupation mineure (Ecossaises). Les analyses génétiques faites sur ces populations suggèrent une histoire commune avec un ancêtre commun datant de la dernière glaciation. Elles détectent également la présence d'un goulot d'étranglement récent dont la datation ne peut pas être confirmée avec les modèles existants utilisés. L'identification de cette datation sera envisagé dans la continuité de ce travail en utilisant le nouveau modèle que nous avons développé (*VarEff*) et permettant de décrire l'évolution des tailles efficaces au cours du temps. Enfin, d'une manière surprenante, les analyses ont montré que toutes les populations possédaient une forte variabilité génétique malgré un faible flux génique. Pour comprendre la forte diversité génétique au sein des plus petites populations (Oir et Scorff), un nouveau modèle de prédiction de la diversité génétique actuelle (*DemoDivMS*), en fonction de l'histoire démographique de la population a également été développé sur les bases théoriques du modèle *VarEff*. Son utilisation suggère que la forte variabilité au sein de ces petites populations est la conséquence d'un goulot d'étranglement récent et que le faible flux génique observé ne peut à lui seul expliquer cette forte variabilité génétique. Cependant, nous ne pouvons pas exclure l'hypothèse de la « compensation génétique » propre aux salmonidés, mentionnée par plusieurs auteurs (Fleming 1996, 1998 ; Jones et Hutchings 2001, 2002 ; Palstra et Ruzzante 2008), qui consiste à l'apparition de tacons mâles matures (précoces) se reproduisant avec les femelles (Palstra et Ruzzante 2008). Ces mâles précoces féconderaient une large proportion d'oeufs dans les populations de Saumon atlantique sauvage d'Europe du Sud, ce qui contribuerait à l'augmentation de la variabilité génétique en balançant le sexe ratio, en augmentant le brassage et la taille efficace (*Ne*) (Garcia-Vazquez et al. 2001). Les proportions élevées d'œufs fécondés par les tacons matures dans les deux populations Françaises (Oir et Scorff) (Baglinière et Maisse 1985 ; Baglinière et al. 1993) pourraient ainsi contribuer au maintien de la forte diversité génétique observée dans ces populations.

Le point fort de l'étude sur la comparaison des trois estimateurs de *Ne* est l'utilisation de populations très différentes en terme de structure populationnelle et de mode de gestion. Les analyses de ces populations avec 9 panels microsatellites au sein desquels varient le nombre et le polymorphisme (5 marqueurs hautement (H+) et faiblement (H-) hétérozygotes puis 10, 20 et 28 marqueurs H+ et H-, et les 37) ont révélé divers points majeurs à considérer dans le cadre du développement de programmes de conservation de ces populations.

Tout d'abord les estimateurs de *Ne* sont sensibles au nombre de marqueurs et à leurs polymorphismes. Bien que l'incertitude autour de la médiane soit importante pour les modèles de coalescence liés à un échantillon, notamment avec MSVAR, celle-ci diminue tout en augmentant le nombre de marqueurs polymorphes et ceci pour les trois méthodes. L'utilisation d'une trentaine de marqueurs neutres polymorphes pourrait être préconisée. Toutefois, l'incertitude des méthodes à deux échantillons est également influencée par la taille de l'échantillon (*S*) en fonction de la taille globale (*Nt*). Plus ce rapport est petit et plus l'incertitude devrait diminuer. Dans notre cas d'étude, au-delà d'un rapport *Nt/S* de 20, l'incertitude augmente fortement.

Certaines distributions des fréquences allèliques peuvent fournir des valeurs totalement erronées. Une surreprésentation d'un allèle va avoir tendance à surestimer la taille. Une mauvaise répartition des allèles entre échantillons va également surestimer la taille avec les méthodes des moments comme TM3.

Le mode de gestion des populations, tel que les introductions d'individus natifs et non natifs, peuvent également modifier les estimateurs et ne doivent pas être négligées lors des choix des méthodes. Les valeurs des *Ne* médians sont sous-estimées par MSVAR et surestimées par TM3 dans la Shin, subissant des introductions natives, alors que TM3 sous-estime la Scorff, en ayant subi une introduction non native.

À la différence de l'étude de Palstra et Ruzzante (2008), l'ensemble des résultats ne montre pas que plus la taille globale est petite et plus la taille efficace est proche de la taille globale. Le lien entre la taille efficace et la taille globale est très différent d'une méthode à l'autre et il est donc difficile de conclure avec les résultats obtenus. Cependant, il apparaît que la plus petite population, Oir, est nettement surestimée par MSVAR. L'explication pourrait rejoindre l'hypothèse de la compensation génétique mentionnée plus haut, puisque cette population présente un nombre élevé de tacons mâles (33-79% de 1+) se reproduisant avec les femelles adultes (Baglinière et al. 1993).

Pour finir, l'ensemble des méthodes semble sensible à la constitution des *a priori*. Cette sensibilité se traduit soit par une augmentation de l'incertitude en utilisant MSVAR, soit par

une estimation erronée dépendant fortement de l'*a priori* sur toutes les populations en utilisant DIY ABC et sur seulement les grandes populations (*Nt/S* élevé >20) en utilisant TM3. Il serait donc préférable que la méthode DIY ABC ne soit pas utilisée sur des populations de Saumon pour lesquelles on manque d'information pour formuler un *a priori* réaliste.
Les méthodes de coalescence à un échantillon ne se comportent pas de la même façon que celles à deux échantillons. Il semble plus sûr de coupler la méthode MSVAR avec une méthode comme TM3.

Le fort homing chez le Saumon atlantique incite à négliger la migration comme force majeure d'évolution de la variabilité génétique. Cependant, comme nous avons pu le voir, les introductions semblent avoir un impact sur la variabilité des populations suivant l'origine des souches. La taille efficace peut être sur- ou sous-estimée si l'introduction est respectivement non native ou native, avec les méthodes à un échantillon et inversement avec les méthodes à deux échantillons. Finalement il semble que les méthodes de coalescence sous dérive, mutation et migration pourraient être les mieux adaptées. Néanmoins, l'utilisation de l'une de ces méthodes de coalescence mêlant la dérive, la mutation et la migration (LAMARC), a révélé plusieurs inconvénients. Les temps de calculs se sont avérés très longs (4 à 6 semaines) pour atteindre la convergence et les résultats se sont révélés incohérents et différents selon les marqueurs. Bien que les temps de calculs par les autres méthodes soient un peu plus courts, les méthodes de coalescence actuelles restent gourmandes en temps, à cause de la simulation des arbres de coalescence qu'elles entreprennent pour faire leurs calculs.

Pour d'une part limiter ce temps de calcul et diminuer les incertitudes, et d'autre part répondre aux questions des tendances évolutives sur l'espèce, un nouveau modèle de coalescence (*VarEff*) sous dérive et mutation a été développé dans ce travail. En effet, bien que les trois modèles de coalescence que nous avons comparés sur les populations étudiées aillent dans le sens des conclusions de l'IUCN sur le statut des populations, c'est-à-dire vulnérables pour les deux Françaises et à préoccupations mineures pour les deux Ecossaises, les résultats ne permettent pas de conclure sur le devenir des populations. Il était donc indispensable de développer un nouveau modèle capable de retracer l'évolution de ces populations. *VarEff* est ainsi un modèle de coalescence qui ne se contente pas d'estimer seulement la taille actuelle et ancestrale. Il cherche à retracer les fluctuations des tailles efficaces par paliers de générations passées jusqu'à l'ancêtre le plus lointain. Pour cela et pour réduire les temps de calculs, et c'est ce qui fait l'originalité de ce modèle, la vraisemblance de

Thêta (*4NeM*) est calculée directement, sans passer par les arbres, à partir des observations génétiques des différences de longueurs des microsatellites (nombre de répétitions) (*D*) et des distributions théoriques des temps de coalescence sous le modèle SMM. Les résolutions analytiques permettent d'exprimer les probabilités des *D* et des temps de coalescence en fonction des tailles *Ne* de la population dans le passé et des taux de mutation *M*, c'est-à-dire des paramètres d'intérêt que l'on cherche à estimer. Ces calculs ont conduisent ensuite à une approximation de la vraisemblance des observations.

Le modèle *VarEff* utilise une partie des résultats analytiques présentés dans les travaux du troisième article (Chevalet et Nikolic, 2009), concernant les calculs des probabilités conjointes des distances entre allèles (*D*) en fonction du modèle de mutation et des tailles efficaces. Dans l'article, les solutions analytiques concernent les transformées de Fourier des distributions des différences de longueur *D* et des temps de coalescence entre allèles, leur inversion pouvant être traitée numériquement dans les situations où la taille de la population est variable. Ainsi, on peut travailler séparément sur les distributions des temps de coalescence d'un côté et des distances *D* de l'autre, et voir l'influence des mutations et des tailles efficaces sur ces distributions. Les distances *D* sont directement liées aux taux de mutation *M* et à *Ne*, et plus ceux-ci sont élevés plus il y aura de valeurs observables de *D* (*Do, D1*…) avec des fréquences élevées pour les grands *D*. Concernant les temps de coalescence conditionnés par les *D*, on peut voir que les différences entre ces temps seront d'autant plus importantes que *M* et *Ne* sont petits car les fréquences des grandes différences alléliques sont plus faibles. Pour les grandes valeurs de *M* ou de *Ne*, les petites distances sont le reflet des dernières manifestations de la coalescence. Cela donne à penser que pour détecter les variations démographiques récentes les marqueurs doivent être assez variables. En regardant les distributions des temps de coalescence non conditionnés, qui représentent la superposition des distributions des temps conditionnés, on peut voir que ces distributions sont plates et que les estimations de taille efficace devraient être plus précisées en s'appuyant sur les temps conditionnés, surtout en présence de perturbation démographique légère. Tous ces résultats laissent à penser que *VarEff*, par ses fondements sur les calculs analytiques, transcrira l'histoire d'une population plus fidèlement que les autres modèles.

Ce modèle a bien évidement des limites. Tout d'abord le modèle ne prend pas en compte la migration compte tenu de sa complexité à l'intégrer à l'échelle évolutive et sa grande variabilité dans le milieu naturel. Néanmoins dans les cas où les taux de migration ont

pu être estimés, en supposant cette migration homogène et ancienne, une alternative est de les intégrer comme une force de mutation dans notre modèle. C'est ce que nous avons fait dans le modèle *DemoDivMS*, pour étudier les fortes diversités génétiques observées au sein des petites populations. Les résultats montrent que, dans ces cas, la migration n'était sans doute pas la principale explication de la forte diversité génétique observée. Ensuite, les analyses sur données théoriques montrent que lors d'un goulot d'étranglement trop intense et long, les allèles coalescent rapidement, ce qui diminue l'échelle du temps sur laquelle nous pouvons retracer les fluctuations des *Ne* dans les générations passées. Néanmoins, la comparaison des temps de coalescence conditionnés par les distances entre allèles, comme ceux utilisés dans *VarEff*, et des temps non conditionnés, suggère que *VarEff* permet de rendre compte de l'histoire d'une population.

Ce modèle qui a été conçu pour retracer l'évolution des tailles efficaces du Saumon atlantique, afin de comprendre sa diminution mondiale, peut être utilisé sur tous les organismes diploïdes, à partir de marqueurs microsatellites, si la migration est faible. Le travail de cette thèse se poursuivra par l'analyse des populations étudiées afin de retracer leur évolution et également évaluer les faiblesses et les performances de *VarEff*.

Ignorer les paramètres génétiques, comme la taille efficace mais également la diversité et la structure des populations, dans les programmes de conservation pourrait avoir divers effets sur l'estimation de l'état, l'identification du statut et la gestion des populations. Tout d'abord le risque d'extinction pourrait être sous-estimé dans de nombreux taxons (Brook et al. 2002). La réintroduction pourrait être inadaptée, notamment si cette introduction se fait avec des individus très apparentés ou avec une autre population ayant un nombre de chromosomes différent produisant des hybrides stériles. Les flux de gènes ne pourraient pas être détectés et une population pourrait se voir fragmentée. Les problèmes associés à l'incompatibilité génétique dans une petite population ne pourraient pas être pris en compte (Young et al. 2000). Il est donc essentiel de lier les méthodes actuelles avec les outils génétiques pour compléter les mesures existantes. De plus, l'absence des données démographiques du Saumon atlantique en mer confirme la nécessité pour cette espèce d'estimer la taille efficace afin d'affiner les évaluations des états des stocks.

RÉFÉRENCES

BIBLIOGRAPHIQUES

Allendorf FW, Thorgaard GH. 1984. Tetraploidy and the evolution of salmonid fishes. In Evolutionary Genetics of Fishes, ed. BJ Turner: 1-53. Plenum, New York.

Altukhov YP, Salmenkova EA, Omelchenko VT. 2000. Salmonid Fishes: Population biology, genetics and management. Blackwell Science, Oxford.

Aprahamian MW, Davidson IC, Cove RJ. 2008. Life history changes in Atlantic salmon from the River Dee, Wales. Hyrobiologia, 602: 61-78.

ASF. 2004. Atlantic Salmon Aquaculture: A primer. Atlantic Salmon Federation. (Disponible: http://www.asf.ca:Aquaculture/ASFaquaculture2004.pdf).

Ayllon F, Martinez JL, Juanes F, Gephard S, Garci-Vasquez E. 2006. Genetic history of the population of Atlantic salmon, Salmo salar L., under restoration in the Connecticut River, USA. ICES Journal of Marine Science, 63: 1286-1289.

Baglinière JL, Denais L, Rivot E, Porcher JP, Prévost E, Marchand F, Vauclin V. 2004. Length and age structure modifications of the Atlantic salmon (*Salmo salar*) populations of Brittany and Lower Normandy from 1972 to 2002. Technical Report, INRA-CSP, p24.

Baglinière JL, Marchand F, Vauclin V. 2005. Interannual changes in recruitment of the Atlantic salmon (*Salmo salar*) population in the River Oir (Lower Normandy, France): relationships with spawners and instream habitat. ICES Journal of Marine Science, 62: 695-707.

Baglinière JL, Thibault M, Dumas J. 1990. Réintroductions et soutiens de populations du Atlantic salmon atlantique (*Salmo salar* L.) en France. La Terre et la Vie, Suppl. 5: 299-323.

Baglinière JL, Maisse G, Nihouarn A. 1993. Comparison of two methods of estimating Atlantic salmon, Salmo salar, wild smolt production. In GIBSON R.J., CUTING R.E. (eds), Production of juvenile Atlantic salmon, Salmo salar, in natural waters, Canad. Spec. Publ. Fish. Aquat. Sci., 118, 189-201.

Baglinière JL, Maisse G. 1985. Precocious maturation and smoltification in wild atlantic salmon in the armoricain Massif, France. Aquaculture 45(1-4): 249-263.

Baudouin L, Lebrun P. 2000. An operational bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. Acta Horticulturae 546: 81-93.

Barton NH and Slalkin M, 1986. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity 56: 409-415.

BBC. 2005. Salmon storm escape figure given. (http://news.bbc.co.uk/1/hi/scotland/4287407.stm).

Boeuf G. 1992. Salmonids smolting : a pre-adaptation to the oceanic environment. In : RANKIN J.C., JENSEN F.B. (Eds.), Fish Ecophysiology : 105-135, Chapman and Hall, London.

Beall E. 1994. Les phases de reproduction. In « Le Saumon Atlantique », J.C. Gueguen and Prouzet P., Eds IFREMER: 123-140.

Beaumont MA. 2001. Conservation Genetics, in Balding, D.J., Bishop M., Cannings, C. (eds) Handbook of Statistical Genetics. John Wiley, New York.

Beaumont MA. 1999. Detecting population expansion and decline using licrosatellites. Genetics 153:1414-1422.

Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B Biol Sci. 263:16191626.

Berthier P, Beaumont MA, Cornuet JM, Luikart G, 2002. Likelihood-based estimate of the effective population size using temporal changes in allele frequency: a genealogical approach. Genetics 160: 741–751.

Billard R, Jalibert M, Marcel J, Carpentier P. 1987. La gestion des géniteurs, des gamètes et des œufs chez le Saumon atlantique et divers salmonidés. In La restauration des rivières à saumons, INRA, M. Thibault et R. Billard éditeurs: 251-264.

Blaxter JHS. 1953. Sperm storage and cross-fertilization of spring and autumn spawning herring. Nature 172: 1189–1190.

Brook BW, Tonkyn DW, O'Grady JJ, Frankham R. 2002. Contribution of inbreeding to extinction risk in threatened species. Conservation Ecology 6(1), 16.

Brown AHD, 1970. The estimation of Wright's fixation index from genotypic frequencies. Genetica 41: 399-406.

Caballero A. 1994. Developments in the prediction of effective population size. Heredity 73: 657–679.

Carnac P. 1988. Le saumon Atlantique : sa biologie et son élevage en France. Thèse en Pharmacie.

Caron F, Fontaine PM. 2003. L'état des stocks de Saumon atlantique au Québec en 2002. Société de la faune et des parcs du Québec. Direction sur la Faune, p48.

Carter TJ, Pierce GJ, Hislop JRG, Houseman JA, Boyle PR. 2001. Predation by seals on salmonids in two Scottish estuaries. Fisheries Management and Ecology 8: 207-225.

Chevalet C, Nikolic N. 2009. Distribution of coalescent times and distances between microsatellite alleles with changing effective population size. Theoretical Population Biology (*En révision*).

Cockerman CC. 1973. Analysis of gene frequencies. Genetics 74: 679-700.

Cockerman CC. 1969. Variance of gene frequencies. Evolution 23: 72-83.

Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ, Guillemaud T, Estoup A. 2008. Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation, Bioinformatics 24(23): 2713-2719.

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.

Consuegra S, Garcìa de Leàniz C, Serdio A, Gonzalez Morales M, Straus LG, Knox D, Verspoor E. 2002. Mitochondrial DNA variation in Pleistocene and modern Atlantic salmon from the Iberian glacial refugium. Molecular Ecology 11: 2037-2048.

Consuegra S, Verspoor E, Knox D Garcìa de Leàniz C. 2005. Asymmetric gene flow and the evolutionary maintenance of genetic diversity in small, peripheral Atlantic salmon populations. Conservation Genetics 6: 823-842.

Consuegra S, Nielsen EE. 2007. Population size reductions. *In :* Verspoor E, Stradmeyer L, Nielsen JL. The Atlantic Salmon: Genetics, Conservation and Management. Blackwell Publishing: 239-269.

Cuinat R. 1980. Rejets de matières ou suspension par les exploitations de granulats dans la rivière Allier. Effets sur la vie aquatique. CSP, 6e D.R., coll. FAO, CECFI Vichy.

Cury P, Miserey Y. 2008. Une mer sans poissons. Calmann-lévy.

Crespi B, Fulton MJ. 2004. Molecular systematics of Salmonidae: combined nuclear data yields a robust phylogeny. Molecular Phylogenetics and Evolution 31: 658-679.

C.T.G.R.F. 1976. Bases de gestion de l'eau en salmoniculture intensive. Etud. Cent. Tech. Génie Rural Eaux Forêts, Bordeaux 4, p82.

Davaine P, Prouzet P. 1994. La vie marine du saumon Atlantique dans son aire géographique. In « Le Saumon Atlantique », J.C. Gueguen and Prouzet P., Eds IFREMER: 64-85.

Dickson RR, Turrell WR. 1999. The NAO: The dominant atmospheric process affecting variability in home, middle and distant waters of European Atlantic salmon. In The Ocean Life of Atlantic Salmon: 92-115. Mills, D. K. (Ed.). Oxford: Fishing News Books, Blackwell Science.

Dittman AH, Quinn TP. 1996. Homing in Pacific salmon: mechanisms and ecological basis. Journal of Experimental Biology, 199, 83–91.

Dodson J, Laroche J, Lecompre F. 2009. Contrasting Evolutionary Pathways of Anadromy in Euteleostean Fishes. American Fisheries Society Symposium 69 (*in press*).

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology Evolution 22: 1185-1192.

Dukes JP, Deaville R, Bruford M W, Youngson A F, Jordan W C. 2004. Odorant receptor gene expression changes during the parr-smolt transformation in Atlantic salmon. Molecular Ecology 13(9):2851-7.

Dulvy NK, Sadovy Y, Reynolds JD. 2003. Extinction vulnerability in marine populations. Fish and Fisheries 4: 25-64.

Dumas J, Prouzet P. 2003. Variability of demographic parameters and population dynamics of Atlantic salmon (*Salmo salar*) in a south west French river. ICES Journal of Marine Science, 60: 356-370.

Eackles MS, King TL. 2002. Aquatic Ecology Branch, U.S. Geological Survey, 11700 Leetown Road, Kearneysville, WV 25430, USA.

Estoup A, Largiadèr C-R, Perrot E, Chourrout D. 1996. One-tube rapid DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. Molecular Marine Biology and Biotechnology 5: 295-298.

Excoffier L. 2001, Analysis of population subdivision, pp.217-324 in Handbook of Statistical Genetics, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. John Wiley & Sons, Chichester, UK.

Excoffier L, Smouse PE, and 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.

FAO. 2006. THE STATE OF WORLD FISHERIES AND AQUACULTURE 2006. ISSN 1020-5489.

Fergusson A, Fleming IA, Hindar K, Skaala Ø, McGinnity P, Cross T, Prodöhl P. 2007. Farm escapes. *In :* Verspoor E, Stradmeyer L, Nielsen JL. The Atlantic Salmon: Genetics, Conservation and Management. Blackwell Publishing: 357-398.

Fleming IA. 1996. Reproductive strategies of Atlantic salmon: ecology and evolution. Reviews in Fish Biology and Fisheries 6: 379–416.

Fleming IA. 1998. Pattern and variability in the breeding system of Atlantic salmon (*Salmo salar*), with comparisons to other salmonids. Canadian Journal of Fisheries and Aquatic Sciences 55: 59–76.

Frankel OH, Soulé ME. 1981. Conservation and evolution. Cambridge Univ. Press, New York. NY.

Frankham R. 2005. Genetics and extinction. Biological Conservation 126: 131-140.

Frankham R, Ballou JD, Briscoe DA. 2002. Introduction to Conservation Genetics. Cambridge University Press: Cambridge, UK.

Franklin IR, Frankham R. 1998. How large must populations be to retain evolutionary potential? Animal Conservation, 1: 69–73.

Franklin IR. 1980. Evolutionary change in small populations. In: M.E. Soulé and B.A. Wilcox (Ed.) Conservation Biology: An evolutionary-ecological perspective: 135-149. Sinauer Associates, Sunderland, MA.

Friedland KD, Reddin DG, Castonguay M. 2003. Ocean thermal conditions in the post-smolt nursery of North American Atlantic salmon. ICES Journal of Marine Science, **60**, 343-355.

Garcia-Vazquez E, Moran P, Martinez JL, Perez J, de Gaudemar B, Beall E. 2001. Alternative mating strategies in Atlantic salmon and brown trout. Journal of Heredity 92(2): 146-9.

Gharbi K. 2001. Construction d'une carte génétique partielle du génome tétraploide de la truite commune (*Salmo trutta*). Cartographie comparée des régions paralogues et alignements avec les cartes du saumon atlantique (*Salmo salar*) et de la truite arc-en-ciel (*Oncorhynchus mykiss*). Thèse, Sciences animales, Génétique et hérédité. INRA.

Gilbey J, Knox D, O'Sullivan M and Verspoor E. 2005. Novel DNA markers for rapid, accurate and cost-effective discrimination of continental origin of Atlantic salmon (*Salmo salar L.*). ICES Journal of Marine Science 62(8): 1609-1616.

Graham M. 1935. Modern theory of exploiting a fishery and application to North Sea trawling. J. Cons. Int. Explor. Mer. 10: 264–274.

Gueguen JC, Prouzet P. 1994. Le Saumon atlantique : Biologie et gestion de la ressource", (Eds), IFREMER, Brest, p330.

Guyomard R. 1994. La diversité génétique des populations de saumon atlantique. *In.* Le saumon atlantique par Gueguen J.C et Prouzet P. *Ed*. Infremer: 141-151.

Hartley SE, Horne MT. 1984. Chromosome relationships in the genus *Salmo*. Chromosoma 90: 229-237.

Hartley SE. 1987. The chromosomes of salmonid fishes. Biol. Rev. 62: 197-214.

Hasler AD. 1966. Underwater Guideposts: Homing of Salmon. University of Wisconsin Press, Madison.

Heland M, Dumas J. 1994. Ecologie et comportement des juvéniles. In « Le Saumon Atlantique », J.C. Gueguen and Prouzet P., Eds IFREMER: 29-46.

Hewitt GM. 1999. Post-glacial recolonization of European biota. Biol. J. Linn. Soc. 68: 87-112.

Hewitt GM. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. Biological Journal of the Linnean Society 58: 247-276.

Hoar WS. 1976. Smolts transformation: evolution behavior and physiology. J. Fish. Res. Bd Canada 33(5): 1233-1252.

Hoar WS. 1988. The physiologie of smolting salmonids. In “Fish Physiologie, W.S. Hoar and D.J. Randall eds, vol XIB, Academic Press, New York: 275-343.

Hoyheim B. and Jorgensen S.M. 2000. MGA-Genetics, Norwegian School of Veterinary Science, PO Box 8146 DEP, Oslo NO-0033, Norway.

Hoyheim B. 2000 Atlantic salmon microsatellites. MGA-Genetics, Norwegian School of Veterinary Science, PO Box 8146 DEP, Oslo NO-0033, Norway.

Iannuccelli E, Woloszyn N, Arhainx J, Gellin J, Milan D. 1996. *Proc. XXVth ISAG*, Tours, France, 88.

ICES. 2001. Report of the Working Group on the Atlantic Salmon. ICES Document, CM2001/ACFM: 15. p199.

ICES. 2003. Report of the Working Group on the Atlantic Salmon. ICES Document, CM2001/ACFM: 19. p297.

Irigoien X, Harris RP, Head RN, Harbour D. 2000. North Atlantic Oscillation and spring bloom phytoplankton composition in the English Channel. Journal of Plankton Research 22(12): 2367-2371.

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

Jones MW, Hutchings JA. 2001. The influence of male parr body size and mate competition on fertilization success and effective population size in Atlantic salmon. Heredity 86: 675–684.

Jones MW, Hutchings JA. 2002. Individual variation in Atlantic salmon fertilization success: implications for effective population size. Ecological Applications 12: 184–193.

Jonsson B, Jonsson N, Hansen LP. 2003. Atlantic salmon straying from river Imsa. Journal of Fish Biology 62: 641-657.

Jonsson B, Jonsson N. 2004. Factors affecting marine production of Atlantic salmon (*Salmo salar*). Canadian Journal of Fisheries and Aquatic Sciences 61: 2369–2383.

Jordan WC, Verspoor E, Youngson AF. 1997. The effect of selection on estimates of genetic divergence among populations of the Atlantic salmon (*Salmo salar*). Journal of Fish Biology 51: 546-560.

Kelly LA, Stellwagen J, Bergheim A. 1996. Waste loadings from a fresh-water Atlantic salmon farm in Scotland. Water Research Bulletin 32(5): 1017-1025.

Klemetsen A, Amundsen PA, Dempson JB et al. 2003. Atlantic salmon Salmo salar L., brown trout Salmo trutta L. and Arctic charr Salvelinus alpinus (L.): a review of aspects of their life histories. Ecology of Freshwater Fish 12: 1–59.

King TL, Verspoor E, Spidle AP, Gross R, Philips RB, Koljonen M-L, Sanchez JA, Morrison CL. 2007. Biodiversity and population structure. *In :* Verspoor E, Stradmeyer L, Nielsen JL. The Atlantic Salmon: Genetics, Conservation and Management. Blackwell Publishing: 117-166.

King TL, Kalinowski ST, Schill WB, Spidle AP, Lubinski BA. 2001. Population structure of Atlantic salmon (*Salmo salar L.*): a range-wide perspective from microsatellite DNA variation. Molecular Ecology 10:807–821.

Kingman JFC. 1982. The coalescent. Stoch. Proc. Appl. 13: 235-248.

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

Knockaert C. 2006. Salmonidés d'aquaculture. De la production à la consommation. Ed. Quae.

Koljonen M-L, Tähtinen J, Säisä M, Koskiniemi J. 2002. Maintenance of genetic diversity of Atlantic salmon by captive breeding programmes and the geographic distribution of microsatellite variation. Aquaculture 212 : 69-92.

Johnson KR, Wright JE Jr, May B**.** 1987. Linkage relationships reflecting ancestral tetraploidy in salmonid fish. Genetics 116: 579-591.

Jones MW, Hutchings JA. 2001. The influence of male parr body size and mate competition on fertilization success and effective population size in Atlantic salmon. Heredity 86: 675–684.

Jones MW, Hutchings JA. 2002. Individual variation in Atlantic salmon fertilization success: implications for effective population size. Ecological Applications 12: 184–193.

Lage C, Kornfield I. 2006. Reduced genetic diveristy and effective population size in an endangered Atlantic salmon (*Salmo salar*) by captive breeding programmes and the geographic distribution of microsatellite variation. Aquaculture 212: 69-92.

Lande R. 1995. Mutation and conservation. Conservation Biology 9: 782-791.

Lande R. 1988. Genetics and demography in biological conservation. Science 241: 1455–1460.

Le François NR, Lamarre SG, Blier PU**.** 2003. La cryoconservation du sperme de loup de mer : recherche d'un médium de conservation efficace et évaluation de la qualité du sperme après décongélation. Dans : Activités 2001-2002. Direction de l'innovation et des technologies. MAPAQ: 55-56.

Lynch M, Lande R. 1998. The critical effective size for a genetically secure population. Animal Conservation 1: 70–72.

Mace GM, Lande R. 1991. Assessing extinction threats – toward a reevaluation of IUCN threatened species categories. Conservation Biology 5: 148–157.

MacCrimmon HR, Gots BL. 1979. World distribution of Atlantic salmon, Salmo salar. Journal of Fisheries research Board of Canada 36: 422-457.

Malécot G. 1941. Etude mathématique des populations "mendéliennes". *Ann*. Univ. Lyon, Sci., A. 4: 45-60.

McConnell SK, O'Reilly P, Hamilton L, Wright JM, Bentzen P. 1995. Polymorphic microsatellite loci from Atlantic salmon (*Salmo salar*): genetic differentiation of North American and European populations. Canadian J. Fisheries and Aquatic Sciences 52:1863-1872.

Meffe G, Carroll CR. 1997. Principles of conservation biology. Sinauer Associates, Sunderland, MA.

Meine C, Soulé M & Noss RE. 2006. « A mission-driven discipline » : the growth of conservation biology. Conservation Biology 20: 631-651.

Middlemas SJ, Armstrong JD, Thompson PM. 2003.The significance of marine mammal predation on salmon, and sea trout p 43-60. *In* " Salmon at the edge" D. Mills (Ed), Blackwell Science Ldt, Oxford.

Mills D. 1989. Ecology and Management of Atlantic salmon. Chapman and Hall, London. p351.

Mills D. 1987. Succès du IIIème Symposium international du Saumon atlantique. Saumons 59: 10-12.

Mills D. (Ed.) 2000. The ocean life of Atlantic salmon, Fishing News Books, Blackwell Science: 75-87.

Mills D. (Ed). (2003). Salmon at the Edge. Blackwell Science, Oxford.

Mounib MS. 1978. Cryogenic preservation of fis hand mammalian spermatozoa. J. Report. Fert. 53: 13-18.

NASCO. 1996. The Atlantic salmon as the predator and the prey. Report of the special session of the Council. Publication CNI (96) 59. North Atlantic Salmon Conservation Organisation.

Narum SR, Banks M, Beacham TD, Bellinger MR, Campbell MR, Dekoning J, Elz A, Guthrie Iii CM, Kozfkay C, Miller KM, Moran P, Phillips R, Seeb LW, Smith CT, Warheit K, Young SF, Garza JC. 2008. Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. Mol Ecol. 25

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

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89: 583-590.

Nei M. 1973. Analysis of gene diversity in subdivised populations. Proc. Natl. Acad. Sci. USA 70: 3321-3323.

Nei M. 1972 Genetic distances between populations. American Naturalist 106: 283-92.

Nielsen EE, Hansen MM, Loeschcke V. 1997. Analysis of microsatellite DNA from old scale samples of Atlantic salmon *Salmo salar*: a comparison of genetic composition over 60 years. Molecular Ecology 6: 487-492.

Nielsen EE. 1999. The evolutionary history of steelhead (*Onchorynchus mykiss*) along the US Pacific Coast: developing a conservation strategy using genetic diversity. ICES Journal of Marine Science 56: 449-458.

Nikolic N, Butler J, Baglinière J-L, Laughton R, McMyn I.A.G, Chevalet C. 2009a. An examination of genetic diversity and effective population size in Atlantic salmon. Genetics Research (*En révision*).

Nikolic N, Fève F, Chevalet C, Hoyheim B, Riquet J. 2009b. A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon (*Salmo salar*) populations. Journal of Fish Biology 74: 458-466.

Nikolic N, Park Y-S, Sancristobal M, Lek S, Chevalet C. 2009c. What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations Genet. Res., Camb. 91: 121–132.

Nunney L, Elam DR. 1994. Estimating the effective size of conserved populations. Conservation Biology 8: 175–184.

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. Hereditas 59: 169-187.

Ohno S, Muramoto J, Steinus C, Christian L, Kittrell WA et al. 1969. Diploid-tetraploid relationship in clupeoid and salmonoid fish. In Chromosomes Today, Vol. 21, eds CD Darlington and KR Lewis, pp. 139-147. Oliver and Boyd, Edinburgh.

Ohno S. 1970a. Evolution by Gene Duplication. Springer-Verlag, New York.

Ohno S. 1970b. The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. Trans. Am. Fish. Soc. 99: 120-130.

Ollivier L. 2002. Eléments de génétique quantitative, 2ème edition revue et augmentée, INRA, Mieux Comprendre, Masson.

Oosterhout CV, Hutchinson WF, Wills DPM, Shipley P, 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4: 535–538.

O'Reilly PT, Hamilton LC, Wright JM. 1997. Biology, Dalhousie University, Halifax, NS B3H 4J1, Canada.

O'Reilly PT, Hamilton LC, McConnell SK, Wright JM. 1996. Rapid analysis of genetic variation in Atlantic salmon (*Salmo salar*) by PCR multiplexing of dinucleotide and tetranucleotide microsatellites. Canadian Journal of Fisheries and Aquatic Sciences 53: 2292–98.

O'Reilly PT, Hamilton LC, McConnell SKJ Wright JM. 1995. Patrick T. O'Reilly, Biology, Dalhousie University, Halifax, N.S. B3H 4J1, Canada.

Otto SP, Whitton J. 2000. Polyploidy incidence and evolution. Annu. Rev. Genet. 34: 401- 437.

Paetkau D, Calvert W, Stirling I, Strobeck C. 1995. Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology 4: 347-354.

Palstra FP, Ruzzante DE. 2008. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? Molecular Ecology 17: 3428-3447.

Parrish BB & Shearer WM. (1977) Effects of Seals on Fisheries. International Council for the Exploration of the Sea, Council Meeting 1977. Document CM, 1977/M: 14 (mimeo).

Parrish DL, Behnke RJ, Gephard SR, McCormick SD, Reeves GH. 1998. Why aren't there more Atlantic salmon (*Salmo salar*) ? Canadian Journal of Fisheries and Aquatic Sciences 55: 281-287.

Paterson S, Piertney SB, Knox D, Gilbey J, Verspoor E. 2004. Characterization and PCR multiplexing of novel highly variable tetranucleotide Atlantic salmon (*Salmo salar* L.) microsatellites. Molecular Ecology Notes 4: 160–162.

Porcher JP, Baglinière JL. 2001. Le Atlantic salmon atlantique. In "Atlas des poisons d'eau douce de France", Keith P. et J. Allardi (coords.). Patrimoines Naturels 47: 240-243.

Potter ECE, Crozier WW. 2000. A perspective on the marine survival of Atlantic salmon. In The ocean life of Atlantic salmon: environmental and biological factors influencing survival. Edited by D.H. Mills. Fishing News Books, Blackwell Science, Oxford, U.K: 19-36.

Pottinger TG, Pickering AD. 1997. Genetic basis to the stress response: selective breeding for stress-tolerant fish. In : Fish stress and health in aquaculture, G.K. Iwama, A.D. Pickering, J.P. Sumpter, C.B. Schreck (eds), 171-193. Cambridge University Press, Cambridge, UK.

Prévost E. 1987. Recherches sur le Saumon atlantique (Salmo salar L.) en France. Thèse Docteur-Ingénieur, ENSA Rennes / Univ. Rennes I., p103.

Prévost E, Chaput G. (Eds.) 2001. Stock, recruitment and reference points. Assessment and management of Atlantic salmon. INRA, Paris, p223.

Pritchard JK, Stephens M, Donnelly PJ. 2000. Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Quinn TP. 1993. A review of homing and straying of wild and hatchery-produced salmon. Fisheries Research 18: 29-44.

Rae BB. 1960. Seals and Scottish fisheries. Marine Research 2: 1-39.

Rae BB. 1968. The food of seals in Scottish waters. Marine Research 2: 1-23.

Rae BB & Shearer WM. 1965. Seal damage to salmon fisheries. Marine Research 2: 1-39.

Ramsey J, Shemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. 29: 467-501.

Rannala B, Mountain JL. 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences USA 94: 9197-9201.

Reed DH, Bryant EH. 2000. Experimental tests of minimal viable population size. Animal Conservation 3: 7-13.

Reed DH, Frankham R. 2003. Correlation between fitness and genetic diversity. Conservation Biology 17: 230-237.

Reddin DG & Shearer WM. 1987. Sea surface temperature and distribution of Atlantic salmon in the Northwest Atlantic Ocean: In Common Strategies of Anadromous and Catadromous Fishes. Am. Fish. Soc. Symposium A: 262-275.

Roule L. 1920. Etude sur le saumon des eaux douces de la France considéré au point de vue de son état naturel et du repeuplement de nos rivières. 1 vol, imprimerie nationale, Ministère Agriculture, Paris, Ed., p178.

Säisä M, Koljonen M-L, Tähtinen J. 2003. Genetic changes in Atlantic salmon stocks since historical times and effective population size of a long-term captive breeding programme. Conservation Genetics 4: 613-627.

Saunders RL. 1981. Atlantic salmon (*Salmo salar*) stocks and mangement implications in the Canadian Atlantic provinces and New England, USA. Canadian Journal of Fisheries and Aquatic Sciences 38: 1612-1625.

Schaefer MB. 1957. A study of the dynamics of the fishery for yellow fin tuna in the Eastern Tropical Pacific. Ocean. Bull. Int. Amer: Trop. Tuna Comm. 2(6): 247-285.

Schaefer MB. 1954. Some aspects of the dynamics of populations, important for the management of the commercial fisheries. Bull. Inter-American Trop. Tuna Comm. 1(2):26-56.

Schindler DW. 2001. Cumulative effects of climate warming and other human stresses on Canadian freshwaters in the new millennium. Canadian Journal of Fisheries and Aquatic Sciences 58: 18-29.

Schlüchter C, Kelly M. 2000. Das Eiszeitalter in der Schweiz. Eine schematische Zusammenfassung . Uttigen, Stiftung Landschaft und Kreis.

Shannon CE. 1948. A mathematical theory of communication. Bell Syst. Tech. 27: 379-423.

Shearer WM. 1992. The Atlantic salmon: natural history, exploitation, and future management. John Wiley & Sons, Inc., New York.

Slatkin M, 1985. Rare alleles as indicators of gene flow. Evolution 39: 53-65.

Slettan A. 1995. Norwegian College of Veterinary Medicine, Dept. of MGA, Division of Genetics, Oslo, Norway, N-0033.

Soltis DE, Soltis PS, 1999. Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* 14: 348-352.

Stabell OB. 1984. Homing and olfaction in salmonids: a critical review with special reference to the Atlantic salmon. Biological Reviews 59: 333-338.

Sun JX, Mullikin JC, Patterson N, Reich DE. 2009. Microsatellites are molecular clocks that support accurate inferences about history. Molecular Biology and Evolution. 26(5): 1017-27.

Svärdson G. 1945. Chromosome studies of Salmonidae. Rep. Swedish State Inst. Freshwater Fish. Res. 23: 1-151.

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Molecular Ecology 7(4): 453-64.

Tagu D. 1999. Principes des techniques de biologie moléculaire. Mieux Comprendre. Ed. INRA.

Taylor AC, Sherwin WB, Wayne RK. 1994. The use of simple sequence of loci to measure genetic variation in bottlenecked species: the decline of the hairy-nosed wombat (*Lasiorhinus krefftii*). Molecular Ecology 3: 277-290.

Thibault M. 1985. Eléments de la problématique du Saumon atlantique en France. Restauration des rivières à saumons. INRA. M. Thibault et R. Billard éditeurs: 413-427.

Thibault M. 1994. Aperçu historique sur l'évolution des captures et des stocks, p. 175-195. *In* "Le saumon atlantique : Biologie et gestion de la ressource", J.C. Guegen et P. Prouzet (Eds), IFRESEA, Paris.

Thompson PM, Mackey B, Barton TR, Duck C, Butler JRA. 2007. Assessing the potential impact of salmon fisheries management on the conservation status of harbour seals (*Phoca vitulina*) in north-east Scotland. Animal Conservation 10(1): 48–56.

Utter FM, Allendorf FW. 1994. Phylogenetic relationships among species of *Oncorhynchus*: a consensus view. Conservat. Biol. 8: 864-867.

Vibert R. 1994. Le saumon Atlantique : Origine et caractéristiques essentielles. In « Le Saumon Atlantique », J.C. Gueguen and Prouzet P., Eds IFREMER: 11-25.

Verspoor E. 1994. The evolution of genetic divergence at protein coding loci among anadromous and nonanadromous populations of Atlantic salmon Salmo salar. In: A. Beaumont (Ed.) Genetics and Evolution of Aquatic Organisms: 52-67. Chapman & Hall, London.

Verspoor E, Stradmeyer L, Nielsen JL. 2007. The Atlantic Salmon: Genetics, Conservation and Management. Blackwell Publishing.

Verspoor E, Beardmore JA, Consuegra S, Garcia de Leaniz C, Hindar K, Jordan WC, Koljonen M-L, Mahkrov AA, Paaver T, Sanchez JA, Skaala O, Titov S, Cross TF. 2005. Population structure in the Atlantic salmon: insights from 40 years of research into genetic protein variation. Journal of Fish Biology 67: 3-54

Verspoor E., Piertney S., Patterson S. and Knox D. 2002. Fish Genetics, FRS Marine Laboratory, Victoria Road, Aberdeen, Scotland AB11 9DB, U.K.

Wanner H, Brönnimann S, Casty C, Gyalistras D, Luterbacher J, Schmutz C, Stephenson D.B, Xoplaki E. 2001. North Atlantic Oscillation – Concepts And Studies. Surveys in Geophysics, Volume 22, Number 4(61): 321-381.

Waples RS. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121: 379-391.

Waples RS. 1990. Conservation genetics of Pacific salmon. III. Estimating effective population size. Journal of Heredity 81: 277-289.

Waples RS, Winans GA, Utter FM, Mahnken C. 1990. Genetic approaches to the management of Pacific Salmon. Fisheries 15: 19-25.

Waples RS. 1991. Pacific salmon, Orchorynchus spp., and the definition of 'species' under the Endangered Species Act. Marine Fisheries Reviews 53: 11-21.

Webb J, Verspoor E, Aubin-Horth N, Romakkaniemi A, Amiro P. 2007. The Atlantic salmon. *In :* Verspoor E, Stradmeyer L, Nielsen JL. The Atlantic Salmon: Genetics, Conservation and Management. Blackwell Publishing: 117-166.

Weir BS. 1990. Genetic data analysis. Sinauer, Sunderland, MA.

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

Withler RE, Supernault K J, Miller KM. 2005. Genetic variation within and among domesticated Atlantic salmon broodstocks in British Columbia, Canada. Animal Genetics, 36: 43-50.

Workman PL, Niswander JD. 1970. Population studies on Southwestern indian tribes. II. Local genetic diferentiation in the Papago. Am. J. hum. Genet. 22: 24-49.

Wright JE Jr, Johnson K, Hollister A, May B. 1983. Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. Isozyme Curr. Top. Biol. Med. Res. 10: 239-260.

Wright S. 1978. Evolution and the Genetics of Populations. Vol. 4. Variability within and among Natural Populations. University of Chicago Press, Chicago.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16: 97-159.

Young AG, Brown AHD, Murray BG, Thrall PH, Miller CH. 2000. Genetic erosion, restricted mating and reduced viability in fragmented populations of the endangered grassland herb Rutidosis leptorrhynchoides. In: Young, A.G., Clarke, G.M. (Eds.), Genetics, Demography and Viability of Fragmented Populations. Cambridge University Press, Cambridge: 335–359.

Youngson AF, Jordan WC, Verspoor E, McGinnity P, Cross TF, Fergusson A. 2003. Management of salmonid fisheries in the British Isles: towards a practical approach based on population genetics. Fisheries Research 62: 193-209.

GLOSSAIRE

Partie 1.

VOCABULAIRE

Anadrome

Se dit d'une espèce qui remonte de la mer vers les cours d'eau pour se reproduire (≠ Catadrome).

Amphibiotique

Espèce vivant successivement en mer et en eau douce.

Dérive génétique

Changement à l'intérieur d'une population des fréquences alléliques d'une génération à l'autre, dû à l'échantillonnage d'un nombre limité de gènes, ce qui est inévitable dans des populations de taille finie. Plus petite est la population, plus grande est la dérive génétique, avec comme conséquence la perte de certains allèles, et la réduction de la diversité génétique.

Gène

Unité physique et fonctionnelle de l'hérédité, qui transmet l'information d'une génération à l'autre. C'est un segment d'ADN qui inclut une section transcrite et des éléments régulateurs qui permettent sa transcription.

Génome

L'effectif complet du matériel génétique d'un organisme.

Potamotoque

Organisme amphibiotique migrant en eau douce pour s'y reproduire.

Partie 2.

INDICES EN GÉNÉTIQUE DES POPULATIONS

Fréquence génotypique

Une fréquence génotypique est le rapport du nombre d'individus d'un génotype (gène avec une combinaison allélique précise) sur le nombre total d'individus de la population.

Fréquence allélique

Une Fréquence allélique est le rapport du nombre d'un allèle sur le nombre total d'allèles présents dans la population au locus considéré.

Hétérozygotie attendue

Cette hétérozygotie fait intervenir les lois d'Hardy-Weinberg (voir Test de Hardy-Weinberg). Supposons que l'allèle *A* soit représenté par la fréquence *p=1-q* et l'allèle *a* par la fréquence *q=1-p.* Alors un individu est homozygote pour ce locus si il possède les allèles *aa* ou *AA* et aura une fréquence génotypique de q^2 ou p^2 respectivement. Il sera hétérozygote si, il possède la combinaison Aa dont la fréquence génotypique est :$1-q^2-p^2$.

L'hétérozygotie attendue dans la population peut se calculer de façon générale comme :

$$He = 1 - \sum_{i=1}^{k} p_i^2 ,$$

avec *k* allèles *Ai* de fréquences *pi* dans la population.

Hétérozygotie attendue totale

L'hétérozygotie attendue totale (*Ht*) est l'hétérozygotie attendue sur un ensemble de populations, en supposant que les populations forment une seule population.

Si $\overline{pi}$ est la fréquence du *i-ème* allèle moyennée sur l'ensemble des populations, *Ht* s'écrit :

$$Ht = 1 - \sum_{i=1}^{k} \overline{pi}^2 .$$

Hétérozygotie observée

L'hétérozygotie observée (*Ho*) est la fréquence des locus hétérozygotes sur l'ensemble des locus dans la population.

F-statistiques de Wright

Les F-statistiques (*Fst, Fis, Fit*) de Wright (1978) sont largement utilisées pour caractériser la structure génétique des populations à partir des fréquences génétiques.

Prenons le cas de populations divisées en sous-populations. Trois niveaux hiérarchiques peuvent être observés: celui de l'individu (i), celui de la sous-population à laquelle il appartient (s), et enfin celui de la population considérée dans son ensemble (t). Wright a défini l'hétérozygotie génétique de ces niveaux par les paramètres suivants :

Hi : Hétérozygotie observée dans la population globale.

Hs : Hétérozygotie attendue pour une sous-population supposée à l'équilibre Hardy-Weinberg.

Ht : Hétérozygotie attendue dans la population globale (hétérozygotie attendue totale) supposée à l'équilibre Hardy-Weinberg.

Le premier indice, *Fis*, est alors défini par rapport à la moyenne des *Hs* sur les *k* sous-populations ($\overline{Hs}$) :

$$Fis = (\overline{Hs} - Hi) / \overline{Hs} .$$

Cet indice mesure la diminution d'hétérozygotie à l'intérieur des sous-populations, due à la consanguinité ou à l'homogamie, voire à la sélection. Il prend ses valeurs entre –1 et 1. Si $Fis>0$ les sous-populations présentent un déficit d'hétérozygotes, si $Fis<0$ au contraire un excès en hétérozygotes (lié, par exemple, à un flux migratoire important provoquant un effet d'hétérogamie). Si $Fis=0$ alors les sous-populations sont à l'équilibre de Hardy-Weinberg.

L'effet de la subdivision ou de cohésion entre les sous-populations et la population totale est exprimé par l'indice de fixation suivant :

$$Fst = (Ht - \overline{Hs}) / Ht .$$

Ce paramètre décrivant la différenciation génétique entre les populations peut être également utilisé comme distance génétique. Il prend ses valeurs entre –1 et 1. Si toutes les sous-populations ont la même fréquence allélique et sont à l'équilibre, *Fst* est nul. Dans le cas contraire, les sous-populations ont des fréquences alléliques moyennes différentes. Ces différences sont généralement induites par l'action conjuguée de la dérive, qui diversifie les populations, et de la migration, qui tend à les homogénéiser. Notons que dans le cas d'un effet Wahlund *Ht* sera plus grande que *Hs*, et *Fst* sera positif.

Pour finir, la réduction d'hétérozygotie globale, qui prend en compte l'effet de la reproduction non panmictique et la différenciation génétique, est donnée par l'indice de fixation:

$$Fit = (Ht - Hi) / Ht .$$

Cet indice peut s'écrire également en mettant en relation avec les deux autres indices :

$$(1 - Fit) = (1 - Fis)(1 - Fst) .$$

Lorsque toutes les populations sont à l'équilibre de Hardy-Weinberg (*Fis=0*), et si elles ont les mêmes fréquences alléliques alors *Fst= Fit=0*.

F-statistiques de Weir & Cockerham

En 1984, Weir et Cockerham ont publié un ensemble d'équations sur le paramètre *Fst* ou θ décrivant la structure génétique des populations. Les auteurs prennent en compte l'effet de la taille des échantillons utilisés pour estimer les fréquences alléliques et génotypiques. Ainsi les indices de fixation de Wright peuvent être calculés d'après les formules de Weir et Cockerham (1984 et 1990). Les paramètres *Fit*, *Fis* et *Fst* de Wright correspondent aux paramètrex *F*, *f* et θ de Weir et Cockerham respectivement.

$$F = 1 - C/(A + B + C)$$

$$f = 1 - C/(B + C)$$

$$\theta = A/(A + B + C)$$

Les 3 quantités *A*, *B* et *C* sont définies dans une étude précédente sur des analyses de variance pondérées (Cockerham, 1973). En notant $\tilde{p}_i$ la fréquence d'un allèle ***a*** dans l'échantillon de taille n_i de la population *i* (i=1,2,…,r) et $\tilde{h}_i$ est la proportion d'individus hétérozygotes observé pour l'allèle ***a***, alors

$$A = \frac{\bar{n}}{n_c}\left\{ s^2 - \frac{1}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{1}{4}\bar{h} \right] \right\},$$

$$B = \frac{\bar{n}}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{2\bar{n}-1}{4\bar{n}}\bar{h} \right],$$

$$C = \frac{1}{2}\bar{h},$$

avec:

$\bar{n} = \sum_i n_i / r$, taille moyenne de l'échantillon,

$n_c = \left(r\bar{n} - \sum_i n_i^2 / r\bar{n} \right) / (r-1) = \bar{n}(1 - c^2 / r)$, avec c^2 le carré du coefficient de la variance de la taille de l'échantillon ,

$\bar{p} = \sum_i n_i \tilde{p}_i / r\bar{n}$, la moyenne de la fréquence de l'allèle a dans l'échantillon,

$s^2 = \sum_i n_i (\tilde{p}_i - \bar{p})^2 / (r-1)\bar{n}$, la variance de la fréquence de l'allèle au travers des échantillons,

$\bar{h} = \sum_i n_i \tilde{h}_i / r\bar{n}$, la moyenne des hétérozygoties pour l'allèle ***a***.

Gst

Le *Gst* de Nei (1973) est l'analogue du *Fst* de Wright, il est utilisé dans le cas de locus multialléliques.

Rst

Le *Rst* de Slatkin (1995) ou de Excoffier (2001) est un analogue du *Fst* de Wright (1946, 1951). Cette mesure a été conçue pour les marqueurs microsatellites et se base sur les différences du nombre de répétitions du motif répété.

Lorsque l'on se place sous un modèle en îles avec des populations diploïdes sous équilibre dérive et migration (grand nombre de populations de taille efficace (*N*) recevant à chaque génération une proportion m de gènes pris au hasard des autres populations) *Fst* est généralement utilisé sous la forme suivante:

$$Fst \approx 1/(1+4N(m+u)),$$

avec *u* le taux de mutation.
Cette relation suppose que $u << m$. Néanmoins, il a été montré que les microsatellites ont un *u* élevé entre 10^{-5} et 10^{-2} (Jarne et Lagoda, 1996), il est donc préférable d'utiliser l'indice Rst.

Le *Rst* de Excoffier s'écrit :

$$Rst = (Sb - Sw)/Sb,$$

avec *Sw* la moyenne au carré des différences de taille entre allèles de la même population et *Sb* entre allèles de différentes populations.

Flux génique

Dans le modèle en « îles » de Wright, les populations ont un effectif limité, elles évoluent sans sélection et les immigrants sont supposés provenir, au hasard, des autres populations avec un taux de migration égal à *m*. Dans cette situation la migration s'oppose à l'effet de la dérive résultant de l'effectif limité. Il a été montré que *Fst* ne tend plus alors vers 1 (fixation totale des gènes) mais vers une limite égale approximativement à :

$$Fst=1/(4Nm+1).$$

Nous pouvons en déduire un flux génique *Nm* qui représente le nombre approximatif de migrants à chaque génération :

$$Nm=[(1/Fst)-1]/4.$$

Cette limite, lorsque les migrations s'opposent à la dérive, peut s'appliquer pour le cas des mutations. Cependant les mutations s'opposent moins efficacement que les migrations ce qui entraîne que, plus les effectifs sont grands, plus les mutations empêcheront *Fst* de tendre vers 1. Par contre, un taux de migration très supérieur à *1/4N* fait que les populations se comportent génétiquement comme une seule population panmictique infinie (*Fst*=0) (Ollivier, 2002).

Distance génétique de Nei

Les mesures de distance génétique résument la quantité des divergences génétiques entre les couples d'individus, de populations ou d'espèces. Il existe plusieurs mesures de distance génétique dont beaucoup sont reliées au *Fst*. La distance de Nei (1972) (*D*) est la plus utilisée pour estimer les distances génétiques entre deux populations. Cette distance compare deux populations *x* et *y* en calculant l'opposé du logarithme naturel de l'identité normalisée (*I*):

$$D = -\ln[I].$$

La distance de Nei est ainsi fondée sur les fréquences alléliques :

$$I = \frac{Jxy}{\sqrt{(JxJy)}},$$

$$Jxy = \sum_{m} \sum_{i} p_{mix}\, p_{miy}\ ,$$

$$Jx = \sum_{m} \sum_{i} p_{mix}{}^{2} \qquad Jy = \sum_{m} \sum_{i} p_{miy}{}^{2}\ ,$$

avec *m* = nombre de locus ; p_{mix} = fréquence de l'allèle *i* du locus *m* pour la population *x* ;

$\sum_m$ = somme sur les locus,

$\sum_i$ = somme sur les allèles.

Il existe une autre distance de Nei établie en 1978 qui calcule la distance *D* en introduisant une correction pour le biais de l'échantillonnage d'individus (voir Nei 1978).

Distance Euclidienne

Soit deux point *x* et *y* définie par les vecteurs (*xi*,…,*xn*) et (*yi*,..,*yn*) respectivement. La distance euclidienne entre ces deux points s'écrit alors:

$$d = \sqrt{\sum_{i=1}^{n} \left| x_i - y_i \right|^2} \ .$$

Entropie

En 1948 Shannon a développé un indice *Hs* dans le cadre de la théorie de l'information. Cet indice qualifié d'entropie suppose que la diversité peut être mesurée de la même façon que l'information contenue dans un code ou un message. Si p_i est la fréquence de l'allèle *i*, en un locus qui compte *k* allèles, alors l'indice de Shannon s'écrits:

$$Hs(p) = -\sum_{i=1}^{k} p_i \log p_i \ .$$

Partie 3.
MODÈLES D'ANALYSE

ACP

L'Analyse en Composantes Principales (ACP) est une méthode d'ordination classique basée sur les distances Euclidiennes. Elle représente un ensemble de *n* objets, qui se trouvent à l'origine dans un espace de *p* descripteurs, dans un espace réduit de k dimensions ($k << p$). Pour cela, elle va chercher les axes orthogonaux (indépendants) qui représentent le maximum de la variance projetée des objets sur l'espace de dimension *k*.

AFC

Comme toutes les méthodes d'ordination, sa finalité est de trouver la meilleure représentation possible dans un espace de dimensions réduites. Elle va estimer les axes qui maximisent l'inertie projetée pour obtenir une représentation simultanée des lignes et des colonnes dans l'espace de dimensions réduites cherché. A la différence de l'ACP, l'analyse des correspondances multiples (AFC) est basée sur une métrique du Khi-2. Elle considère les lignes et les colonnes de la matrice de façon symétrique, ce qui entraîne une dualité entre l'espace des colonnes et l'espace des lignes.

AMOVA

L'AMOVA (Analyse de Variance Moléculaire) (Excoffier et al., 1992) répond à des attentes de la biologie moléculaire. Elle a été développée pour des plans d'échantillonnage hiérarchique. Cette analyse définie au préalable par Cockerham (1969, 1973) étudie la structure génétique des populations en évaluant les variances entre et à l'intérieur. Pour cela, elle calcule une distance génétique entre individus et fait une analyse par permutations.

ASSIGNATION

Un test d'assignation est une méthode qui « assigne » chaque individu à la population où son génotype a la plus forte probabilité de se trouver. Plus la proportion d'individus assignés à une population autre que leur population d'origine est faible et plus le niveau de différenciation entre populations est fort. Généralement cette méthode utilise les rapports de vraisemblance : par exemple le rapport de vraisemblance entre la population dans laquelle l'individu a été échantillonné (population d'origine) et la plus forte vraisemblance parmi toutes les populations. Pour calculer cette vraisemblance, une méthode bayésienne (Rannala et Mountain (1997) ; Baudouin et Lebrun (200)), une méthode basée sur les fréquences (Paetkau et al., 1995) ou encore une méthode basé sur les distances génétiques peut être utilisée. Les

méthodes Bayésiennes et fréquencielles sont souvent plus performantes que celles basées sur les distances (voir Cornuet et al. 1999 pour une étude comparative).

BOOTSTRAP

Les méthodes de bootstrap se sont rapidement développées dès la fin des années 1980 et sont largement utilisées de nos jours. Le terme de ré-échantillonnage, ou en anglais, bootstrap désigne un ensemble de méthodes qui consiste à affiner l'inférence faite sur les observations initiales, en analysant l'ensemble des nouvelles observations obtenues par tirage au sort de nouveaux échantillons à partir d'un échantillon destiné à donner une certaine information sur la population.

COALESCENCE

La théorie de la coalescence (Kingman, 1982) décrit le processus de fusion (coalescence) des lignages généalogiques des copies alléliques présentes dans une population. Dans cette théorie, les mutations surviennent le long des lignages selon un processus aléatoire de Poisson.
L'idée est de retracer l'histoire généalogique d'une population jusqu'à l'ancêtre commun le plus récent (MRCA). Pour cela, elle considère que 2 gènes à la génération 'g' sont copies d'un même gène à la génération précédente 'g-1' avec une probabilité de un sur deux fois la taille efficace (1/2Ne) comme dans les approches classiques de Wright (1931) ou de Malécot (1948), et généralise ce raisonnement au cas d'un échantillon de plusieurs gènes.

HARDY-WEINBERG

La loi de Hardy-Weinberg a été établie en 1908 par l'anglais Hardy et l'allemand Weinberg et s'énonce comme suit :

Dans une population fermée, d'effectif illimité, non soumise à la sélection ni aux mutations, et où les unions se font au hasard, les fréquences géniques restent constantes de génération en génération, de même que les fréquences génotypiques, celles-ci pouvant se déduire de la connaissance des fréquences géniques.

Ainsi le test de Hardy-Weinberg est un test d'équilibre entre les fréquences génotypiques attendues et les fréquences génotypiques observées dans la population. Une population sous équilibre Hardy-Weinberg est une population à effectif illimité, qui conserve une fréquence génique constante au cours des générations, en l'absence de migration, de mutation et de

sélection. Ainsi les fréquences génotypiques attendues par la loi de Hardy-Weinberg sont évaluées en supposant une population fermée et un régime panmictique. Connaissant les fréquences de chaque allèle pour un locus donné, la proportion des individus hétérozygotes et homozygotes doit respecter cet équilibre. Ainsi les fréquences génotypiques attendues pour un locus donné à deux allèles A_1 et A_2 peuvent s'écrire:

$A_1A_1 : p^2$

$A_1A_2 : 2pq$

$A_2A_2 : q^2$

Avec p la fréquence de l'allèle A_1 et q la fréquence de l'allèle A_2 .

Ainsi pour une fréquence allélique A_1 de 0.6 et A_2 de 0.4, une population à l'équilibre Hardy-Weinberg aura la constitution génotypique suivante :

$$A_1A_1 = 0.36 \qquad A_1A_2 = 0.48 \qquad A_2A_2 = 0.16\,.$$

Si les fréquences génotypiques observées dans la population sont différentes des fréquences attendues alors on dit que la population n'est pas à l'équilibre Hardy-Weinberg.
Le calcul des écarts est généralement évalué par le test du Khi2 qui pose l'hypothèse nulle Ho=la population se reproduit au hasard. Si la probabilité d'obtenir les valeurs du khi2 observées est plus petite que 0.05 (dans le rang 0<P<0.05), nous pouvons conclure que les résultats sont significatifs et que l'hypothèse nulle est rejetée en faveur de l'hypothèse H1=la population ne se reproduit pas au hasard.

IAM (Infinite Allele Model)

Dans ce modèle à nombre infini d'allèles, chaque nouvelle mutation donne naissance à un nouvel état allélique. Ce modèle ne tient donc pas compte des homoplasies et des mutations réverses. Il suppose également que chaque nouvelle mutation est indépendante des autres et qu'une mutation n'en favorise pas une autre. L'exemple des réparations géniques préférentielles dans le sens AT vers CG montre que ce dernier point est faux. Ainsi le modèle IAM est une approximation du processus de mutation réel.

INFERENCE BAYESIENNE

On nomme inférence bayésienne la démarche permettant de calculer la probabilité d'une hypothèse. Cette démarche est régie par le théorème de Bayes qui revient à dire que des probabilités peuvent être déterminées à partir d'observations mais aussi à partir de connaissance a priori exprimée par une loi de probabilité a priori sur des paramètres représentant une hypothèse. Les données statistiques (Y) sont combinées aux hypothèses d'intérêts (θ) (paramètres inconnus sur lesquels ont dispose d'a priori $Po(\theta)$) :

$$P(\theta|Y)=(P(Y|\theta)*Po(\theta))/P(Y).$$

Ainsi, l'approche bayésienne permet de modéliser les observations et les paramètres du modèle au moyen de probabilités qui tiennent compte de l'a priori dont l'utilisateur dispose sur l'ensemble des paramètres. Cependant, l'implémentation pratique de cette approche se heurte régulièrement à des problèmes analytiques et/ou numériques. Cela conduit à coupler la méthode bayésienne avec l'algorithme de simulation MCMC (Markov Chain Monte Carlo) qui permet de résoudre ces problèmes rencontrés.

MARKOV CHAIN

En probabilité un processus stochastique vérifie la propriété markovienne si la distribution conditionnelle de probabilité des états futurs, étant donné l'instant présent, ne dépend que de ce cet état présent (ou d'un nombre fini d'états passés). Un processus qui possède cette propriété ne dépendre que de l'état présent est appelé processus de Markov « du premier ordre ». La notion se généralise si le futur ne dépend que d'un nombre fini d'états antérieurs.

MAXIMUM DE VRAISEMBLANCE

Cette méthode a été développée par Ronald Fisher entre 1912 et 1922.

Soit $p(x/\theta)$ la distribution d'une quantité observée x, fonction d'un paramètre θ. Pour chacune des observations xi, on mesure la valeur $p(xi/\theta)$ puis on fait le produit de toutes ces valeurs. La vraisemblance des observations désignée traditionnellement par la lettre L, en raison de l'appellation anglo-saxone "Likelihood" s'écrit:

$$L(p(x|\theta)) = \prod i\ p(xi|\theta).$$

Le principe de maximum de vraisemblance consiste à prendre pour estimation du paramètre θ la valeur qui maximise L.

MONTE CARLO

On désigne par « méthode de Monte-Carlo » toute méthode numérique utilisant le tirage de nombres aléatoires, plus précisément en utilisant des suites de nombres pseudo-aléatoires générées par des algorithmes spécialisés. Ces méthodes furent longtemps une des utilisations majeures des techniques de simulation, dont les premières études remontent à la seconde guerre mondiale avec les recherches sur la bombe atomique par John Von Neumann et Stanislas Ulam. Elles étaient notamment utilisées pour résoudre des équations aux dérivées partielles.

On peut par exemple approcher la valeur de Pi par une méthode de Monte Carlo : Imaginons un cercle de rayon 1 inscrit dans un carré de coté 2. L'aire du carré est donc 4. Pour trouver l'aire du cercle, il suffit de lancer un ensemble de point au hasard dans le carré et compter le nombre de points qui se trouvent dans le disque et le nombre de points total dans le carré. En multipliant par 4 ce rapport (nombre de points dans le disque/ nombre de points total) nous retrouvons l'aire approximative du disque qui est Pi.

En pratique la méthode de monte Carlo permet de simuler des distributions de probabilités qu'on ne sait pas calculer analytiquement. Elle connaît un usage étendu selon la puissance des ordinateurs.

MCMC (Monte Carlo par Chaînes de Markov)

Les algorithmes de Monte Carlo par Chaînes de Markov sont apparus récemment et sont actuellement les plus utilisés pour calculer l'inférence bayésienne des modèles multiparamétriques. En effet, dans la pratique, la connaissance sur les a priori ($Po(\theta)$) n'a pas de structure propre et la taille des échantillons est généralement faible (la normalité asymptotique n'est pas envisageable). Historiquement, on a d'abord développé les méthodes de Monte Carlo pour approcher les densités. Les méthodes MCMC quand à elles génèrent des suites aléatoires de valeurs des paramètres d'intérêt (θ), produites par des trajectoires de chaînes de Markov homogènes. L'algorithme de génération de ces séquences converge vers une limite stationnaire qui donne les densités a posteriori (θ /x) des paramètres..

La distribution a postériori peut s'écrire :

$$P(\theta|x) = \frac{P(x|\theta) * Po(\theta)}{P(x)},$$

où ($P(x/\theta)$) est la vraisemblance des donnés (x) au vu des paramètres (θ).

PERMUTATIONS

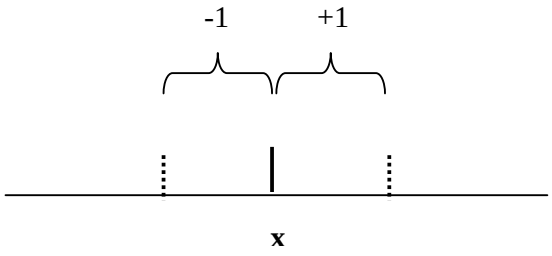
Cette procédure permet de réaliser des tests de significativité des valeurs observées des estimateurs en effectuant un certain nombre de tirages sans remise sur l'ensemble des valeurs observées. Ces tirages permettent d'obtenir la distribution de l'estimateur sous une hypothèse nulle (Ho). La valeur observée sur les données réelles est ensuite comparée à la distribution obtenue, et permet d'obtenir une estimation de la probabilité d'obtenir sous H0 une valeur supérieure ou égale à la valeur observée. Les permutations permettent d'éviter les inconvénients signalés avec le test de Khi², leur usage a été proposé par Brown (1970) pour Fis et Workman & Niswander (1970) pour Fst.

Le test de permutations peut être effectué par un tirage

- sur l'ensemble des allèles toutes populations confondues : procédure utilisée pour tester l'hypothèse nulle *Fit* = 0.
- sur l'ensemble des individus toutes populations confondues : procédure utilisée pour tester l'hypothèse nulle *Fst* = 0
- sur l'ensemble des allèles de chaque population : procédure utilisée pour tester l'hypothèse nulle *Fis* = 0.
- sur les locus.

SMM (Stepwise Mutation Model)

Chaque nouvelle mutation provoque l'ajout (+1) ou la suppression (-1) d'une répétition à un site donné (x).



TEST DE MANTEL

Le test de Mantel est un test de corrélation entre deux matrices. Ces matrices doivent être de même rang et en relation avec les mêmes vecteurs objets. Ce test est communément utilisé en Ecologie pour évaluer la corrélation entre les distances génétiques et les positions géographiques d'espèces ou de populations étudiées. Lorsque les matrices ne sont pas symétriques, le simple calcul du coefficient de corrélation ne peut pas être appliqué. On applique alors le coefficient de corrélation du produit des moments de Pearson (désigné par « *rho* » ou « *r* »). Il reflète le degré de relation linéaire entre deux variables et est compris entre -1 et +1. Un coefficient avec une valeur de +1 signifie qu'il y a une relation linéaire positive parfaite et de -1 une relation linéaire négative parfaite entre les deux variables. La valeur 0 signifie que les variables n'ont aucune relation linéaire.

AUTHOR:
NATACHA NIKOLIC

TITLE:
GENETIC DIVERSITY AND EFFECTIVE SIZE OF WILD FISH POPULATIONS: THE CASE OF ATLANTIC SALMON A DIADROMOUS THREATENED MIGRATORY FISH.

SUPERVISOR:
CLAUDE CHEVALET

LOCATION AND DATE:
Salle de Conférence INRA, July 15, 2009

ABSTRACT

This thesis was concerned mainly with the genetic diversity and the effective size (*Ne*) of wild fish stocks of diadromous migratory endangered Atlantic salmon (*Salmo salar*). For this, four populations were chosen in the northern (Scotland) and southern (France) part of Europe for their differences in their structure and management.
This work led to the development of two new models. *DemoDivMS* predicted genetic diversity of a population whose evolutionary scenario, given by the user, describes variations of *Ne* during the previous generations. *VarEff* tries to estimate the effective sizes, sampling time and pasts, based on direct analytical calculations to shorten the time of calculations.

KEY WORDS :
Genetic diversity, Effective size, Atlantic salmon, Migratory.

AUTEUR :
NATACHA NIKOLIC

TITRE :
DIVERSITÉ GÉNÉTIQUE ET TAILLE EFFICACE CHEZ LES POPULATIONS DE POISSONS SAUVAGES : LE CAS DU SAUMON ATLANTIQUE UN POISSON MIGRATEUR AMPHIHALIN MENACÉ.

DIRECTEUR DE THESE :
CLAUDE CHEVALET

LIEU ET DATE DE SOUTENANCE :
Salle de Conférence INRA, 15 Juillet 2009

RESUME en français

Ces travaux de thèse se sont intéressés principalement à la diversité génétique et à la taille efficace (*Ne*) de stocks de poissons sauvages amphihalins migrateurs en danger, le Saumon atlantique (*Salmo salar*). Pour cela, quatre populations ont été choisies en Europe du Nord (Ecosse) et en Europe du Sud (France) pour leurs différences de structure et de mode de gestion.
Ces travaux ont abouti au développement de deux nouveaux modèles. *DemoDivMS* prédit la diversité génétique d'une population dont le scénario évolutif, donné par l'utilisateur, décrit des variations de *Ne* au cours des générations antérieures. *VarEff* cherche à estimer les tailles efficaces, au moment de l'échantillonnage et passées, en s'appuyant sur des calculs analytiques directs raccourcissant les temps de calculs.

TITRE et résumé en anglais au recto de la dernière page

MOTS-CLES :
Diversité génétique, Taille efficace, Saumon atlantique, Migrateurs.

DISCIPLINE ADMINISTRATIVE (identique à celle de la page de titre)
Sciences Ecologiques, Vétérinaires, Agronomiques, et Bioingénéries

---

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :
UMR444, INRA, chemin de Borde Rouge BP 52627 Auzeville, 31326 Castanet-Tolosan, France.