La génomique, un outil de gestion prometteur pour la gestion des pêches : le cas du homard d'Amérique, *Homarus americanus*, dans l'Est du Canada

Thèse

Laura Benestan

Doctorat en biologie
Philosophiae doctor (Ph. D.)

Québec, Canada

La génomique, un outil prometteur pour la gestion des pêches : le cas du homard d'Amérique, Homarus americanus, dans l'est du Canada

Thèse

Laura Benestan

Sous la direction de :

Louis Bernatchez, directeur de recherche
Rémy Rochette, codirecteur de recherche

Résumé

Le homard d'Amérique, *Homarus americanus*, supporte la pêche commerciale la plus importante dans l'Est du Canada et est donc devenue une espèce prioritaire en terme de gestion et de conservation. Cette thèse vise à acquérir des connaissances importantes sur la reproduction et l'adaptation locale des populations de *H. americanus* à l'aide d'une approche pluridisplinaire alliant génomique des populations et écologie marine. Dans un premier temps, nous avons cherché à définir des unités génétiques et évaluer leur correspondance avec les 41 unités de gestion actuelles. Nos résultats ont révélé la présence de deux entités régionales (nord/sud) composées de 11 populations génétiquement distinctes à plus fine échelle. Nous avons aussi démontré qu'il était possible d'obtenir de fort succès d'assignation à l'échelle régionale, ce qui permet d'envisager un outil de traçabilité. Ensuite, nous avons évalué l'impact des facteurs environnementaux tels que la distribution spatiale, la circulation océanique et la température de surface de la mer sur la distribution des unités génétiques précédemment définies. Nous avons alors démontré que les courants océaniques avaient une plus forte influence sur la divergence neutre des populations que la distribution spatiale. D'autre part, nous avons découvert que la température minimale annuelle avait une influence significative sur la divergence adaptative, et que ce signal persistait même après avoir soustrait l'influence de la distribution spatiale à cette relation. Finalement, nous avons exploré l'influence du sexe ratio et des marqueurs sexuels sur les analyses de structuration génétique d'une espèce marine faiblement structurée, ici le homard d'Amérique. Grâce aux 12 marqueurs sexuels identifiés, nous avons pu révéler le système de détermination sexuelle présent chez cette espèce et caractériser les bases moléculaires de ce déterminisme. Dans l'ensemble, les résultats de cette thèse illustrent le potentiel des outils génomiques dans la mise en place d'une gestion durable du homard d'Amérique dans les eaux canadiennes.

Abstract

The American lobster, *Homarus americanus*, supports the largest commercial fishery in Eastern Canada and has therefore become a priority species in terms of conservation and management. This thesis aimed to gain important knowledge about the genetic structure and adaptive potential of *H. americanus* using a multidisciplinary approach, combining population genomics and marine ecology. Our first goal was to identify genetic units and assess their correspondence to the 41 management units presently in use. Our results revealed the presence of two regional entities (north/south), with at a finer scale, 11 genetically distinguishable populations. We also demonstrated that it was possible to identify the origin of individuals blindly, with an average of 90% individuals correctly reassigned to the regional genetic unit where they were sampled. This high assignment success, unexpected for a marine species, could be used as a relevant traceability tool. Next, we assessed the impacts of environmental factors such as spatial distribution, ocean circulation and sea surface temperature on the previously identified genetic structure. We showed that ocean currents had a greater effect on the putatively neutral genetic structure than spatial distribution. On the other hand, annual minimum temperature appeared to explain a significant portion of the putatively adaptive genetic variation, and this signal persisted even after subtracting the influence of the spatial distribution. Finally, we explored the influence of sex ratio and sex-linked markers on the analyses of genetic structure of high gene flow species, here the American lobster. We found 12 sex-linked markers from which we inferred a probable genetic mechanism of sex determination of the American lobster and characterized its molecular basis. Overall, the results of this thesis illustrate the potential of a genomic approach as a new tool for the sustainable management of American lobster in Canadian waters.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ADN : Acide Désoxyribonucléique (*Deoxyribonucleic Acid*)

AEM : *Asymetric Eigenvector Maps*

AMOVA : Analyse de la Variance Moléculaire (*Analysis of Molecular Variance*)

Db-MEM : *Distance-based Eigenvector Maps*

Fst : Indice de différenciation génétique (*Index of genetic differentiation*)

P-valeur : Valeur de probabilité associée à un test statistique (*Probability value associated to a statistical test, P-value*)

RDA : Analyse de Redondance (*Redundancy Analysis*)

RAD-sequencing : *Restricted Associated DNA sequencing*

SNP : Polymorphisme à un Seul Nucléotide (*Single Nucleotide Polymorphism*)

SST : Température de Surface de la Mer (*Sea Surface Temperature*)

SR : Sexe-Ratio (*Sex-Ratio*)

Remerciements

Tellement de personnes ont contribué à l'élaboration de cet ouvrage sans pour autant y figurer. Grâce à elles aujourd'hui, je réalise un rêve incroyable et j'espère pouvoir un jour les aider autant qu'elles ne l'ont fait pour moi. J'espère surtout ne manquer personne et avoir les mots justes pour exprimer la profonde gratitude et l'estime intemporelle que j'aurai toujours envers vous.

Je me rappellerai toujours de ce jour à la Réunion où un Skype avec toi Louis a changé ma vie. Je suis de ces personnes qui pensent que « la vie fait bien les choses » et cet évènement figure parmi les meilleurs exemples de ma vie pour illustrer cet adage. Je ne te remercierai jamais assez pour m'avoir confié ce projet fantastique, quasiment sur mesure, pour moi qui désirait travailler sur les problématiques touchant la gestion des pêches et la génétique de la conservation. Ta créativité, ton enthousiasme et ta passion font de toi un scientifique hors pair et un modèle à suivre pour nombreux d'entre nous. De plus, ta capacité à trouver des approches alternatives et des solutions à toutes épreuves est aussi un atout certain dont j'ai fortement bénéficié en te côtoyant et que j'espère continuer à appliquer par la suite. Au-delà du côté professionnel, ton humanisme est aussi exceptionnel car rempli d'une grande sensibilité et d'une attention très protectrice que tu portes à chacun de tes étudiants. Malgré les moments de fragilité, les doutes et les craintes, je te serrai à jamais redevable. Merci de m'avoir intégré dans cette belle grande famille du laboratoire Bernatchez, à qui je souhaite une longue vie !

À mon co-directeur, Rémy, qui a toujours été tellement enthousiaste et plein d'idées excellentes pour mon projet. Rémy, tu fais partie de ces scientifiques d'exception que j'admire beaucoup de part leur expertise, leur humilité, leur accessibilité et leur support à toutes épreuves envers leur étudiant. Tu as été un co-directeur en or et je regrette sincèrement de ne pas avoir passé plus de temps en ta compagnie, à apprendre davantage sur l'écologie de mon espèce. Tu es de ces scientifiques qui me donnent l'espoir qu'un jour il est possible que je trouve ma place au sein de cette communauté.

Je souhaiterais remercier les membres de mon comité de thèse : Nadia Aubin-Horth, Nicolas Derôme et Julie Turgeon. Plus particulièrement, merci à Nadia et Julie pour m'avoir suivi depuis le début de mon doctorat jusqu'à aujourd'hui. Vous êtes toutes les deux des modèles de femmes scientifiques inspirantes et j'espère un jour avoir une carrière aussi

épanouie que la vôtre. Vos réflexions, vos commentaires et vos critiques m'ont aidé à construire ce projet doctoral sur des bases solides et à trouver ma place au sein de la communauté scientifique. Je souhaiterai tout particulièrement remercier Julie Turgeon pour m'avoir offert de faire le mentorat du cours d'évolution, ce qui m'a permis de retourner aux les bases de biologie évolutive et de les enseigner. Par ailleurs, je me souviendrai toujours des conservations passionnantes et très diverses que j'ai eu la chance d'avoir avec toi Julie. Merci pour ces moments.

Quand j'ai décidé de suivre le cours de statistiques de Pierre Legendre à Montréal, je savais qu'au fond de moi ce serait beaucoup de travail mais en contre-partie, un énorme plus dans ma formation scientifique. Et je ne me suis pas trompée, Pierre Legendre est plus qu'une somité dans le domaine, c'est un excellent pédagogue qui arrive à vous faire comprendre l'importance de chaque analyse statistique et les limites de son utilisation. Avoir suivi son cours a été un des meilleurs choix que j'ai fait au cours de mon doctorat et sans lui mon troisième chapitre serait remplies d'erreurs statistiques que j'ai eu la chance d'éviter en le côtoyant. Les analyses incluant les vecteurs AEM et dbMEM sont brillamment construites et j'espère qu'elles auront un bel avenir dans les futures publications de génomique du paysage marin.

Que serait le laboratoire Bernatchez sans ces extraordinaires post-docs que j'ai eu la chance de rencontrer et même d'avoir comme collaborateurs sur mes papiers. Bien plus que de simples collègues de travail, certains sont devenus de précieux amis que j'espère garder pour toujours. Pour commencer, Thierry Gosselin, tu as sauvé mon doctorat en m'aidant dans un des moments les plus difficiles de ma vie, je t'en serai toujours redevable. Avoir passé des heures avec toi à écrire des lignes de commandes de folie pour analyser et visualiser les données sous R a été une excellente formation à la programmation scientifique et la bio-informatique en général. Ton soucis de la perfection t'amène à tester de nombreux programmes et à explorer de fond en comble les données, ce que j'ai aussi appris avec toi. Merci à Charles Perrier pour ces discussions et reflexions scientifiques toujours plus poussés qui m'ont fait tellement avancer ! Merci pour ton amitié sincère, ton intégrité et ta personnalité qui reste en place à toutes épreuves. J'ai eu de la chance d'avoir partagé ton bureau pendant environ trois ans. Et puis merci à Thierry et Charly pour les belles escapades à observer les chouettes laponnes, pour ces combats contre le froid sur les pistes de

snowboard à -30 degrés, pour l'escalade de glace aux chutes Montmorency et pour tous ces moments hors laboratoire qui font l'amitié que nous avons aujourd'hui. Merci également à Anne-Laure, pour être une amie et une mère incroyable que j'admire beaucoup. Vivre l'aventure *ConGen* avec toi et explorer la *Road of the Sun* dans la voiture d'un Mike Miller qui roule trop vite restera un moment mythique de mon doctorat! À nos séances de yoga chaud avec Andy et le fameux Trévor, à tes deux petits bouts de choux, Gabin et Marcelline que j'aime tant !!!

Comment ne pas parler aussi d'Anne-Marie, une scientifique exceptionnelle qui est tellement appliquée et investie dans ce qu'elle fait que tout le monde est d'accord pour dire qu'elle est brillante. Merci de m'avoir aidé jusqu'au bout en relisant ma thèse et surtout merci pour tous les moments d'amitié avec toi, Cécile et Karine. Quelle belle équipe! Tu m'as toujours épaté par ton implication, ton dévouement à aider les autres et ton excellence scientifique. Tu as été un excellent modèle pour moi et pour plein d'autres, restes comme tu es !!!

Quand Jean-Sébastien et Anne sont arrivés au laboratoire Bernatchez, une onde d'énergie positive s'est alors propagée et a contaminé tout le labo. Cette fine équipe tout droit venue de la Colombie Britannique a propulsé le laboratoire au rang de laboratoire à la plus cool attitude de tout l'IBIS. Merci à vous deux, vous avez été fantastiques, toujours prêts à m'aider, à me soutenir et à me donner confiance en moi. Connaitre des scientifiques comme vous, ça donne envie de rester dans le milieu ! Merci aussi à Martin Laporte (pour son humour du plus français des québecois), Clément Rougeux, Eric Normandeau, Bérénice Bougas, Alysse Pérault-Payette (pour son incroyable hotel à Portland), Jérémy Gaudin, Jérémy Leluyer, Vincent Bourret, Guillaume Côté, Cécilia Hernandez, Anne Carrier, Anaïs Lacoursière-Roussel, Quentin Rougemont, Maëlle Sevellec, Ben Sutherland et Lucie Papillon pour leur précieux conseils, aide et discussions qui m'on aidé à construire ma réflexion scientifique et à développer mes techniques de laboratoire.

Merci à mon amie la plus proche, qui est comme une sœur de cœur pour moi : Josiane Lamoureux. Te rencontrer a été la plus belle chose qui me soit arrivée au Québec. Il y a des gens qui, lorsque tu les côtoie te rendent meilleur et te font aimer la vie. Tu es de ces personnes là ma Josi. Tu es une personne que j'aime et qui est très importante dans ma vie. Merci aussi au cirque d'être dans ma vie et merci à toute cette belle famille de cirque

international avec qui j'ai eu la chance de faire des spectacles et de voyager : NataRaja, les Festiflam, les Fogo Rasto, les Pailles en feu, les Manda light, les Firebird, Linda Farkas, les Super Cho, les Flame'oz, la familia de Festa di Fuoco, la French mafia. Merci à l'équipe de l'escalade : Lucie, Ana, Basile, Olive, Rémy, Antoine et surtout Véro. Et puis à la belle gang de gumboots : Karine, Annie, Alex, Anne-Marie, Nathalie, Matthieu, Eric, Paul et Guillaume.

Le meilleur pour la fin, merci à ma famille. Mes parents, Tayeb et Francine, ont toujours soutenu chacun de nous vers le métier qu'il souhaitait faire et aujourd'hui nous représentons ce grand mélange qui retrace les rêves de chacun : un scaphandrier, une cinéaste, une infirmière et une biologiste marin. Sans ces parents en or, franchir tous les obstacles et se battre pour nos rêves aurait été comme une lutte acharnée sans armure. J'ai une chance incroyable et je le réalise encore plus au jour d'aujourd'hui. Merci à ma grande sœur Amandine et mon grand frère Samy qui ont toujours cru en moi et m'ont fait relativiser quand j'en avais besoin. Le plus grand trésor de ma vie c'est ma sœur jumelle Emma, que j'aime profondément et qu'il est de plus en plus douloureux de quitter à chaque fois que je rentre chez moi, dans le beau coin du sud de la France. Emma, cette thèse t'es dédiée comme le sont tes films à nos années de vie passées. Je t'aime.

Avant-propos

Cette thèse est organisée en six chapitres incluant un chapitre d'introduction et un chapitre de conclusion (Chapitre 1 et 6). Les quatre autres chapitres sont présents sous forme d'articles scientifiques dont trois ont été acceptés et publiés dans la revue *Molecular Ecology*, tandis que le dernier article a été récemment soumis à *Current Biology*.

Le chapitre 2 est publié sous la référence : **Benestan, L**., Gosselin, T., Perrier, C., Sainte Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular ecology*, *24*, 3299-3315.

Pour le chapitre 2, Laura Benestan, Louis Bernatchez et Rémy Rochette ont conçu le projet. L. Benestan a effectué le protocole de préparation des librairies de type *Restriction site-Associated DNA sequencing* (RAD-sequencing). L. Benestan a produit les données et les a analysées. L. Benestan a également bénéficié de l'aide bio-informatique de Thierry Gosselin, ce qui a grandement contribué au succès de ce deuxième chapitre. Compte tenu des discussions conduites avec Charles Perrier sur ce chapitre, ce dernier a été également inclus dans la liste des auteurs. Pour son temps investi à concevoir et coordonner l'échantillonnage avec L. Benestan et Rémy Rochette, Bernard Sainte Marie fait également partie des auteurs de ce chapitre.

Le chapitre 3 est publié sous la référence : **Benestan, L.M**., Ferchaud, A.L., Hohenlohe, P.A., Garner, B.A., Naylor, G.J., Baums, I.B., Schwartz, M.K., Kelley, J.L. and Luikart, G. (2016). Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular ecology, 25*, 2967-2977.

En raison de sa participation et de sa forte implication dans la rédaction et la conception des sections du chapitre 3, Anne Laure Ferchaud partage la qualité d'auteur principal pour ce dernier chapitre. Pour ce même chapitre, il est à noter que Gordon Luikart est l'auteur sénior qui a supervisé cette publication. Tous les auteurs de ce chapitre ont commenté la version finale du manuscrit avant publication.

Le chapitre 4 est publié sous la référence : **Benestan, L.**, Quinn, B., Laporte, M., Maaroufi H., Rochette, R. & Bernatchez, L. (2016). Seascape genomics provides evidence

for thermal adaptation and current-mediated population structure in American lobster (*Homarus americanus*). *Molecular ecology (in press)*.

Également pour le chapitre 4, notez que Brady. F Quinn a réalisé le modèle de dispersion larvaire du homard d'Amérique dans l'est du Canada. De plus, Halim Maaroufi et Martin Laporte ont respectivement apporté leurs expertises en protéomique et en analyses statistiques à ce chapitre. Tous les auteurs de ce chapitre ont commenté la version finale du manuscrit avant publication.

Le chapitre 5 a été soumis sous la référence : **Benestan L**., Normandeau E., Rougeux C., Rycroft N., Atema J., Rochette R. and Bernatchez L. (2016) Sex matters : gender information is critical for unbiased population structure inferred from high-density SNP genotyping. *Current Biology*.

Pour le chapitre 5, Jelle Atema et Nathan Rycroft ont récolté les échantillons et extrait les ADN. Clément Rougeux a préparé les librairies de RAD-*sequencing* et Eric Normandeau a apporté son expertise en bio-informatique. À ce titre, tous ces auteurs ont participé significativement à l'élaboration du présent manuscript.

Chapitre 1 - Introduction générale

1.1. Mare incognitum

Les océans couvrent 71.1 % de la surface de la Terre et abritent plus de 230 000 espèces qui occupent des niches écologiques très hétérogènes allant par exemple des récifs coralliens aux "forêts" de kelp (Costello *et al.* 2010). Tous ces paysages marins sont connectés par un fluide dense capable de transporter nutriments, oxygène, gamètes et individus sur plusieurs milliers de kilomètres. Influencées par la dynamique de ce fluide, de nombreuses espèces marines arborent des traits d'histoire de vie complexes et font face à des pressions de sélection très différentes au cours de leur cycle de vie (Dawson & Hamner 2008).

Les espèces bentho-pélagiques illustrent clairement cette complexité de traits de vie : les premiers stades larvaires effectuent leur croissance dans la zone pélagique (*i.e.* méroplancton) où elles se déplacent passivement au gré des courants marins, alors que le stade adulte effectue des mouvements actifs sur le *benthos* océanique (Dawson & Hamner 2008; Riginos & Liggins 2013). Ces fluctuations spatiales (*e.g. pelagos* versus *benthos*) et temporelles (*e.g*. larves versus adultes), conjointement avec la diversité des paysages marins qu'occupent ces espèces, sont des facteurs clés qui influencent leur évolution et leur structuration populationnelle. Aussi, définir à quelle ampleur et comment ces facteurs agissent sur l'histoire démographique et adaptative des espèces marines demeure un objectif difficile à atteindre. À ce sujet, nos connaissances sont encore limitées face à l'immensité du territoire que ces espèces occupent, ainsi que la diversité et complexité génétique, phénotypique et comportementale qui les caractérisent.

La documentation des déplacements des organismes marins et l'observation intensive de leur comportement à l'aide des premiers systèmes de type capture-marquage-recapture ont ouvert la porte à une meilleure compréhension de la dynamique des espèces marines. Ces approches pionnières ont révélé l'existence de comportements inattendus tels que le *homing* (*i.e*. fidélité au site de ponte), présents chez de nombreuses espèces de l'Atlantique Nord comme le saumon (Dittman & Quinn 1996; Crossin *et al.* 2007), la morue (Robichaud & Rose 2011) ou encore suggéré chez certains crustacés comme le homard (Chittleborough 1974; Pezzack & Duggan 1986). Plus récemment, la télémétrie, une technologie permettant de suivre les déplacements d'un animal dans le milieu océanique, a permis l'acquisition de nouvelles connaissances sur les migrations et les interactions entre les espèces marines et leur

écosystème (Perras & Nebel 2012). L'avènement de la télémétrie a également soulevé de nouvelles questions, notamment à propos de la documentation de la distribution spatiale des individus et de l'influence du bagage génétique et de l'environnement marin (*i.e.* l'adaptation locale) sur cette distribution (Lenormand 2002).

1.2. La génomique des populations

De l'*Origine des espèces* à aujourd'hui (Darwin 1872), nous établissons graduellement les fondements de notre compréhension de l'évolution du monde vivant, en envisageant une espèce comme un ensemble dynamique formé de plusieurs entités nommées populations (*i.e.* groupe d'individus se reproduisant et interagissant écologiquement avec les autres membres du même groupe sur un espace donné (Waples & Gagiotti 2006)). Cette perspective a fait naître une nouvelle discipline : la génétique des populations. La génétique des populations vise à documenter la distribution et les changements de fréquences alléliques et génotypiques dans les populations occupant des milieux variés de l'aire de répartition d'une espèce (Hedrick 2011). L'expansion de cette discipline a longtemps été limitée par les coûts et le temps associés au développement des marqueurs génétiques nécessaires à la quantification et à l'analyse des variations d'ADN au travers des populations (Luikart *et al.* 2003).

En 2010, l'arrivée des techniques de séquençage « massif en parallèle » ou *Massive Parallel Sequencing*, couplée à de nouvelles méthodes de développement de marqueurs génomiques à large échelle, a démontré qu'il était désormais possible d'identifier, de séquencer et de génotyper des milliers de polymorphismes mononucléotidiques (SNPs; *Single Nucleotide Polymorphism*) sur des centaines d'individus en une seule et unique étape (Davey *et al.* 2011). Bien que les SNPs aient une diversité limitée à quatre états alléliques, leur forte abondance dans le génome (un SNP toutes les centaines de paires de bases environ; Morin *et al.* 2004) a permis d'augmenter la précision et la résolution des analyses de génomique de populations (Allendorf *et al.* 2010; Hemmer-Hansen *et al.* 2014). De plus, leur présence, à la fois dans les régions codantes et non codantes du génome, a également permis de tester l'existence de patrons d'adaptation locale (Allendorf *et al.* 2010).

Cette révolution génomique a facilité la mise en lumière simultanée des patrons de différenciation génétique neutre et adaptatif à fine échelle (Willette *et al.* 2014; Hemmer-

Hansen *et al.* 2014) chez des espèces jusqu'alors considérées comme panmictiques (*i.e.* qui montre une unité génétique homogène). En effet, mesurer l'influence des forces évolutives neutres (*e.g.* flux génique, mutation, dérive) et non neutres (*e.g.* sélection naturelle) permet de mieux comprendre l'intéraction entre ces différentes forces et leur influence respective sur la composition génétique des populations. L'analyse combinée par les marqueurs potentiellement neutres et adaptatifs permet également de délimiter plus précisément les unités de gestion et de conservation adéquates (Allendorf *et al.* 2010; Funk *et al.* 2012).

1.3. La génomique de la conservation appliquée aux pêches

À l'heure de la surpêche et du changement climatique, les écosystèmes marins sont extrêmement vulnérables et nécessitent une gestion appropriée (McCauley *et al.* 2015). Plus particulièrement, l'effondrement de la majorité des stocks de poissons pêchés à l'échelle mondiale démontre qu'il existe un réel besoin de mettre en place une pêche durable (Pauly *et al.* 1998). Une pêche durable repose sur une gestion qui prend en compte la biologie de l'espèce, c'est à dire son histoire démographique et adaptative (Palumbi 2003; Reiss *et al.* 2009). La génomique des populations permet de mettre en évidence ces patrons démographiques et adaptatifs; par exemple en estimant l'ampleur du flux génique entre deux sites géographiques (Palumbi 1994) ou encore en évaluant le potentiel d'adaptation locale des populations d'une espèce, nécessaire à sa viabilité à long terme (Pinsky & Palumbi 2014). Une discordance entre les unités de reproduction et les unités de gestion peut ainsi entraîner la surexploitation ainsi que la disparition de nombreuses populations locales (Reiss *et al.* 2009; Valenzuela-Quiñonez 2016). En effet, si une population qui est démographiquement indépendante (*i.e.* la dynamique de la population dépend davantage des naissances et morts locales que de l'immigration; Funk *et al.* 2012) tend à décliner, l'absence ou le faible apport extérieur de migrants ne suffira pas à la sauver (*i.e.* pas d' « effet sauvetage »; Bowler & Benton 2005). Inversement, si les stocks (*i.e.* groupes d'individus exploités par une unité de gestion) de deux unités adjacentes ne sont pas démographiquement isolés, alors la gestion effectuée sur l'un de ces stocks pourrait avoir d'importantes conséquences sur la viabilité de l'autre stock. Par exemple, si un de ces deux stocks est surexploité et tend à disparaître alors l'autre stock risque de subir le même déclin. Ces réalités démographiques en lien avec la distribution des stocks sont à considérer dans l'établissement des politiques de gestion des

pêches et dans l'élaboration des plans de gestion (Waples *et al.* 2008; Reiss *et al.* 2009; Fu & Fanning 2011; Valenzuela-Quiñonez 2016). Pourtant, peu d'études mettent en concordance les plans de gestion des espèces exploitées avec la structure génétique des populations étudiées, bien que cette structure ait déjà été documentée chez de multiples espèces exploitées (Reiss *et al.* 2009). De plus, l'interprétation des patrons de structuration génétique observée chez ces espèces nécessite une compréhension détaillée de leurs histoires de vie et de l'environnement dans lequel elles évoluent (Hansen & Hemmer-Hansen 2007; Riginos & Liggins 2013).

1.4. De la génétique à la génomique du paysage marin

Investiguer l'influence spatiale et environnementale des paysages marins sur les processus micro-évolutifs (*i.e.* tout changement évolutif au-dessous du niveau de l'espèce, fait référence ici aux changements de fréquence des alléles au sein d'une population; Wilkins 2006) est un point clé pour interpréter les patrons démographiques et adaptatifs observés et ainsi révéler des unités biologiques sous-jacentes (Riginos & Liggins 2013). Cette discipline communément appelée génétique du paysage marin a premièrement axé ses recherches sur le rôle des courants océaniques dans les phénomènes de dispersion larvaire en testant leur lien statistique avec le flux génique (Riginos & Liggins 2013; Selkoe *et al.* 2016). Plusieurs études ont ainsi clairement démontré que, chez les espèces possédant une phase larvaire planctonique, deux sites adjacents peuvent ne pas être connectés lorsqu'ils sont situés sur des côtes bordées par des courants océaniques opposés (Gilg & Hilbish 2003). Réciproquement, deux sites très distants peuvent être connectés lorsqu'ils sont traversés par le même courant océanique (Iacchei *et al.* 2013). De façon similaire, les gyres océaniques peuvent prévenir la diffusion des larves résidentes, séparant ainsi le phénomène de dispersion larvaire des distances géographiques (Weersing & Toonen 2009). Les modèles de dispersion et de recrutement larvaire sont donc des caractéristiques pertinentes à considérer pour améliorer notre compréhension de la structure génétique des populations marines, difficilement résolue à l'aide d'une simple corrélation avec la distribution géographique (*e.g.* latitude et longitude; White *et al.* 2010).

Un deuxième axe de recherche s'est récemment ajouté aux analyses de génomique du paysage marin. Cet axe repose sur la quantification de l'influence des facteurs géographiques

et environnementaux sur la divergence adaptative des populations. Pour définir cette variation génétique adaptative, il est possible d'utiliser des méthodes basées sur de la différentiation populationnelle (*Population Differentiation* ou PD) ou encore de l'association environnementale (*Environmental Association* ou EA) (Rellstab *et al.* 2015; Francois *et al.* 2016). Ces méthodes servent à identifier des marqueurs potentiellement sous sélection sans *a priori,* dans le cas des PD, et avec *a priori* (*i.e.* valeurs de paramètres environnementaux), pour les méthodes de type EA. La combinaison de ces deux méthodes permet à la fois de limiter les erreurs de type I, en considérant uniquement l'ensemble commun de marqueurs génétiques détectés par les deux méthodes, ou encore s'affranchir des erreurs de type II, lorsque que l'ensemble de marqueurs génétiques détectés par chacune de ces méthodes est pris en compte. L'utilisation conjointe des méthodes de PD et EA est donc particulièrement pertinente car elle permet de maximiser nos chances de repérer toutes signatures potentielles de la sélection naturelle sur le génome des espèces marines. Quantifier et délimiter l'influence de la sélection naturelle sur cette variation génétique potentiellement adaptative constitue une étape clé pour ensuite prédire comment ces espèces vont réagir au changement climatique (Savolainen *et al.* 2013).

1.5. Le homard d'Amérique

1.5.1. La biologie de l'espèce

L'aire de répartition du homard d'Amérique, *Homarus americanus*, s'étend du Labrador à la Caroline du Nord, englobant des habitats très variés, notamment en termes de température (de 5°C à 20°C) et de salinité (de 27 ppt à 35 ppt). *H. americanus* est une espèce migratrice qui, au printemps, se déplace vers les eaux côtières pour se reproduire, incuber ou faire éclore ses œufs, et qui, à l'automne, retourne vers des eaux plus profondes, au large (Campbell & Stasko 1986). À l'éclosion des œufs, les larves rejoignent le plancton et sont dispersées par les courants pendant trois à 12 semaines, dépendamment de la température à laquelle les larves se effectuent leur développement (Ennis 1997). À la métamorphose (stade IV), la post-larve prend l'apparence d'un homard adulte et descend sur le fond pour y effectuer sa croissance et son passage au stade adulte. Au stade juvénile, les individus effectuent peu de mouvement en raison du fort risque de prédation (Morse & Rochette 2016). Des études de capture-marquage-recapture révèlent que les homards adultes ont une tendance

à effectuer des déplacements limités, de 15 à 70 kms dans le Golfe du Saint Laurent (Comeau & Savoie 2002) ou à Terre Neuve (Rowe 2011), et plus important dans le Golfe du Maine et la Baie de Fundy (Campbell & Stasko 1986) où ils peuvent se déplacer jusqu'à plus de 300 kms. Cependant, même dans le Golfe du Maine et la Baie de Fundy, la majorité des individus (> 75%) montre des déplacements inférieurs à 15 kms (Campbell & Stasko 1986). De plus, certaines études ont suggéré un comportement de fidélité au site de ponte chez cette espèce (Pezzack & Duggan 1986) et ont également observé que les individus, femelles et mâles, étaient capables de retourner à leur abri d'origine après avoir été déplacés (Karnofsky *et al.* 1989).

1.5.2. Les mâles et les femelles dans les populations naturelles

L'échantillonnage de nombreuses populations naturelles a démontré que la proportion de mâles par rapport à la proportion de femelles était équivalente dans la majorité des sites. Néanmoins, des différences saisonnières existent dans les assemblages d'individus selon les sites. Par exemple, les mâles sont plus souvent capturés dans des température de plus de 16°C par rapport aux femelles (Jury & Watson 2013). De plus, la compétition entre les mâles résulte régulièrement en des sexe-ratios (SR) débalancés et inversement des relations de dominance hiérarchique s'établissent seulement si le SR est débalancé en faveur des femelles (*i.e.* plus de femelles par mâle). Le SR est aussi influencé par la salinité, paramètre environnemental auquel les femelles seraient plus sensibles que les mâles (Howell, Watson & Jury. 1999). Malgré ces différences physiologiques importantes, les bases génétiques du déterminisme sexuel n'ont jamais encore été référencées pour cette espèce. En effet, chez la majorité des crustacés, une grande diversité de systèmes du déterminisme du sexe a été observée. De plus, la présence d'un grand nombre de chromosomes chez ces espèces (pour le homard en moyenne 110 chromosomes; Hughes 2014) complique et limite l'inférence de leurs systèmes de déterminisme sexuel.

1.5.3. La génétique des populations de homard d'Amérique

Les premières études de génétique des populations, basées sur l'analyse des alloenzymes et de l'ADN mitochondrial, ont rapporté l'absence de structuration génétique chez *H. americanus* considérant des sites d'échantillonnage pourtant éloignés de plusieurs milliers de kilomètres (Tracey *et al.* 1975; Harding *et al.* 1997). Ces études attribuaient alors ce résultat au fort potentiel de dispersion de la larve planctonique (Hedgecock 1986). Ce n'est

qu'en 2009 qu'une étude utilisant 13 marqueurs de type microsatellite démontre l'existence d'une faible structuration génétique des populations de *H. americanus* (Kenchington *et al.* 2009). Couvrant une large échelle géographique, un total de huit unités génétiquement différenciées appartenant à deux sous-ensembles régionaux est identifié : le Golfe du Saint-Laurent et Terre Neuve (région nord) d'une part et la côte de Nouvelle-Écosse avec le Golfe du Maine d'autre part (région sud). Les auteurs expliquent cette dichotomie nord/sud en soumettant l'hypothèse de la présence d'un refuge glaciaire au sud et à partir duquel un groupe d'individus aurait colonisé la partie nord suite à la fin de la période glaciaire. Cet effet fondateur aurait conduit à une structuration génétique réduite au sein de la région du nord comparativement à la région sud (Kenchington *et al.* 2009).

1.5.4. La pêche au homard d'Amérique

Suite au déclin de plusieurs espèces marines exploitées ces dernières décennies, la durabilité des pêches est devenue un enjeu global (Pauly *et al.* 2002). Un des défis de la pêche au homard d'Amérique est donc de s'assurer d'une durabilité des captures afin d'éviter l'effondrement des stocks qui pourraient avoir d'importantes conséquences écologiques, économiques et humaines. En effet, la pêche au homard d'Amérique (*Homarus americanus*) est la pêche commerciale la plus importante de tout le Canada Atlantique, représentant plus de 25% de la valeur des débarquements canadiens, toutes espèces et produits confondus. Cette pêche génère plus de 30,000 emplois directs et des milliers d'emplois connexes dans la transformation et l'approvisionnement de biens et services, ce qui amène ainsi le homard d'Amérique à être le principal moteur de l'économie des pêches au Canada (Rochette *et al*. *submitted*). Malgré l'importance socioéconomique considérable de cette espèce, son plan de gestion a été élaboré sur la base de considérations géo-administratives et non définies en fonction des unités biologiques de l'espèce. À l'heure actuelle, nos connaissances sur l'état et la distribution de ces unités biologiques restent limitées alors que pour mettre en place une pêche durable, les mesures de gestion doivent concorder avec la biologie et l'écologie de l'espèce exploitée. Définir exactement ces unités biologiques, à l'aide des outils génétiques actuels, va contribuer à informer et à conseiller les gestionnaires et les pêcheurs sur des pratiques de gestion durable.

1.6. Objectifs de la thèse

Cette thèse s'inscrit dans le cadre d'un vaste programme de recherche intitulé « Structure et connectivité des stocks du homard d'Amérique, *Homarus americanus,* dans l'est du Canada » qui fait partie du Réseau Canadien de Recherche sur la Pêche (*Canadian Fisheries Research Network* ou CFRN). Ce réseau bénéficie d'une subvention stratégique du Conseil de Recherches en Sciences Naturelles et en Génie du Canada. Le programme de recherche se décline en cinq axes de recherche interdépendants et complémentaires qui sont :

(1) la distribution des femelles œuvées, leur productivité et la qualité de leurs œufs;

(2) la modélisation de la dispersion larvaire;

(3) l'étude des facteurs affectant le recrutement des larves sur le *benthos*;

(4) l'étude des mouvements des homards juvéniles et adultes sur le *benthos*;

(5) l'étude de génomique des populations.

Le volet génomique, ici présenté, est couplé aux autres disciplines afin de documenter, en synergie et de façon pluridisciplinaire, la structuration génétique des populations du homard d'Amérique. Par ailleurs, ce projet de recherche implique des personnes venant du gouvernement, de l'industrie des pêches et des milieux universitaires. En travaillant au sein d'un tel réseau pluridisciplinaire, où se mêlent sciences sociales et sciences naturelles (see Turgeon *et al.* 2016), ce projet a aussi pris en considération les enjeux économiques et éthiques que suscite la pêche au homard d'Amérique.

Les objectifs généraux de la thèse s'articulent donc autour des principaux enjeux de gestion et de conservation du homard d'Amérique. En utilisant une récente approche de génomique des populations permettant la caractérisation de la divergence génétique neutre et adaptative à grande échelle, cette thèse vise à :

(1) identifier les unités génétiques du homard d'Amérique présentes sur la majorité de son aire de répartition à l'aide d'un plan d'échantillonnage à large échelle;
(2) développer et définir un cadre d'analyse pertinent et utile aux études en génomique de la conservation;

(3) délimiter l'impact de la distribution spatiale, des courants océaniques et de la température de surface de la mer sur la structuration génétique potentiellement neutre et adaptative;

(4) définir le type de déterminisme sexuel présent chez cette espèce et établir des recommandations quant à l'inclusion des marqueurs sexuels dans une analyse classique de génomique des populations sur une espèce marine faiblement structurée.

Ces objectifs ont été réalisés dans un contexte méthodologique offrant de nouvelles ressources génomiques, telles que celles obtenues à l'aide du développement d'un protocole de préparation de librairies de séquençage de type RAD-*sequencing* (*Restriction Associated DNA sequencing*). Avec le RAD-*sequencing*, nous avons été en mesure de génotyper plus de 10 000 marqueurs génétiques de type SNP (Polymorphisme Nucléotidique Simple, *Single Nucleotide Polymorphism*) sur des centaines d'individus provenant de 13 à 19 sites d'échantillonnage au total.

Le premier chapitre avait pour but d'identifier des unités génétiques distinctes à fine échelle et d'évaluer leur correspondance avec les 41 unités de gestion actuelles. Pour ce chapitre, nous souhaitions savoir si la structuration génétique mise en évidence par l'analyse de milliers de marqueurs SNPs était la même que celle détecté par les 13 microsatellites décrite par Kenchington *et* al. (2009) et/ou si la plus forte résolution obtenue à l'aide de ces marqueurs nous permettait de détecter une structuration à plus fine échelle.

Suite à la publication de ces analyses et à l'expertise acquise lors de l'atelier de génomique évolutive nommé *ConGen*, nous avons défini un cadre méthodologique approprié aux analyses génomiques appliquées à des contextes de gestion et de conservation. Ce cadre a été décrit en détail dans le deuxième chapitre.

Le troisième chapitre visait à délimiter et quantifier l'impact de la distribution spatiale, des courants océaniques et de la température sur la structuration neutre et adaptative des populations. La génomique du paysage marin étant un domaine récent et en pleine expansion, nous avons été les premiers à utiliser des *distance-based Moran's eigenvector map* (db-MEM), représentant la distribution spatiale, et des *Asymmetric Eigenvector Maps* (AEM), représentant les courants océaniques via le modèle de dispersion larvaire, dans des analyses de redondance (RDA) qui incluait des données génomiques. Nous avons ainsi été en mesure de développer une méthode statistique facilitant l'intégration des données d'écologie marine avec celles d'écologie moléculaire.

Finalement, le quatrième chapitre s'intéressait à délimiter l'impact des marqueurs

sexuels sur les analyses classiques de génomique des populations (*e.g.* analyses multivariées, indice de différenciation génétique) ainsi qu'à caractériser le type de déterminisme sexuel présent chez le homard d'Amérique. Cette approche nous a permis de souligner l'importance de collecter les informations sur le sexe des individus pour effectuer une analyse de génomique des populations chez des espèces à faible différenciation génétique.

Chapitre 2 - RAD genotyping reveals fine scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*).

Publié sous : Benestan L, Gosselin T, Perrier C, Sainte Marie B, Rochette R. & Bernatchez L. (2015). *Molecular ecology*, **24**, 3299-3315.

2.1. Résumé

Décrypter la structuration génétique des populations marines est une tâche difficile à accomplir en raison de leur faible niveau de différenciation génétique et de la résolution limitée donnée par les méthodes traditionnelles de génotypage. À l'aide de récents outils génomiques de type RAD-*sequencing*, nous avons identifié 10,156 SNPs dans la perspective de déterminer la structuration génétique du homard d'Amérique et d'effectuer des tests d'assignation populationnelle. Pour cela, nous avons collecté 586 homards américains dans 17 sites répartis sur la majorité de l'aire de répartition naturelle de l'espèce. Dans un premier temps, nos résultats ont révélé l'existence d'une structuration génétique hiérarchique, séparant les homards de la partie nord de l'aire de distribution de ceux de la partie sud (F_{CT} = 0.0011; P-valeur = 0.0002). Puis une strcuture locale à fine échelle a été mise en évidence avec l'identification de 11 populations génétiquement différentes (moyenne F_{ST} = 0.00185 ; CI: 0.0007 à 0.0021, P-valeur < 0.0002). Une procédure de rééchantillonnage a montré que le succès d'assignation populationnel atteignait un optimum en utilisant un sous-ensemble de 3000 SNPs montrant les plus fort F_{ST}. En appliquant la méthode d'Anderson (2010) pour éviter les « biais de surclassement », 94.2% et 80.8% des individus ont été correctement assigné à leur région et leur unité génétique d'origine, respectivement. Enfin, nous avons montré que le succès d'assignation était positivement associé à la taille d'échantillon utilisé. Ces résultats démontrent la pertinence de génotyper un grand nombre de SNPs pour améliorer la délimitation de la structuration génétique à fine échelle et le succès d'assignation populationnel dans un contexte de faible structuration génétique. Ici, nous discutons de l'implication de ces résultats en terme de conservation et de gestion des espèces marines, plus particulièrement en ce qui concerne l'échelle géographique de l'indépendance démographique.

2.2. Abstract

Deciphering genetic structure and inferring connectivity in marine species has been challenging due to weak genetic differentiation and limited resolution offered by traditional genotypic methods. The main goal of this study was to assess how a population genomics framework could help delineate the genetic structure of the American lobster (*Homarus americanus*) throughout much of the species' range, and increase the assignment success of individuals to their location of origin. We genotyped 10,156 filtered SNPs using RAD-sequencing to delineate genetic structure and perform population assignment for 586 American lobsters collected in 17 locations distributed across a large portion of the species' natural distribution range. Our results revealed the existence of a hierarchical genetic structure, first separating lobsters from the northern and southern part of the range (F_{CT} = 0.0011; P-value = 0.0002), and then revealing a total of 11 genetically distinguishable populations (mean F_{ST} = 0.00185; CI: 0.0007-0.0021, P-value < 0.0002), providing strong evidence for weak, albeit fine-scale population structuring within each region. A resampling procedure showed that assignment success was highest with a subset of 3000 SNPs having the highest F_{ST}. Applying Anderson's (2010) method to avoid "high-grading bias", 94.2% and 80.8% of individuals were correctly assigned to their region and location of origin, respectively. Lastly, we showed that assignment success was positively associated with sample size. These results demonstrate that using a large number of SNPs improves fine scale population structure delineation and population assignment success in a context of weak genetic structure. We discuss the implications of these findings for the conservation and management of highly connected marine species, particularly regarding the geographic scale of demographic independence.

2.3. Introduction

Determining genetically distinct populations and establishing appropriate management units are primary goals of modern conservation biology and population management (Palsboll *et al.* 2007). Towards that end, assignment tests are very useful and versatile tools (Manel *et al.* 2005; Schwartz *et al.* 2007), encompassing a wide array of applications, ranging from population structure inferences to the "real-time" detection of migrants (reviewed by Manel *et al.* 2005). However, highly connected and/or recently diverged populations with large effective population sizes often show very weak genetic differentiation, thus decreasing the power of genetic tools for defining management units and assigning individuals to their origin (Allendorf *et al.* 2010). The advent of Next Generation Sequencing (NGS) genotyping methods (Davey *et al.* 2011) promises an increase in the usefulness of genomics markers to finely define weakly structured populations (Hess *et al.* 2013; Ogden *et al.* 2013; Wilette *et al.* 2014) and more accurately assign individuals (Nielsen *et al.* 2012; Larson *et al.* 2014; Candy *et al.* 2015).

Elucidating the genetic structure of populations for conservation and management purposes is particularly challenging in marine species (Allendorf *et al.* 2010). Over the last several decades, numerous studies have attempted to interpret the very weak genetic differentiation (typically $F_{ST} < 0.01$) found in most marine species and determine how to link this genetic information to management plans (Palumbi 2003; Waples & Gagiotti 2006; Waples *et al.* 2008). Here, a major issue concerns the biological meaning of such weak genetic differentiation in terms of levels of demographic independence between populations (Waples & Gagiotti 2006; Waples *et al.* 2008). In many marine species characterized by large effective population size (N_e), weak genetic structure generally translates into pronounced genetic connectivity ($N_e m$) but it is unclear how this relates to demographic connectivity (m), which matters most for short-term population management (Cano *et al.* 2008). Indeed, the transition from demographic dependence to independence in populations with large N_e occurs within the asymptotic region of the hyperbolic relationship between F_{ST} and $N_e m$, where genetic data have typically been insufficiently precise to discriminate between migration rates that are meaningful or not to demographic independence.

Working on a larger genomic scale by very substantially increasing the number of markers could overcome previous methodological limitations by (i) improving the accuracy

of population genetic estimates, (ii) allowing the use of assignment tests for inferring "real-time" migration, and (iii) providing new insights from previously unexplored genomic regions (Kohn *et al.* 2006). Recent studies on sturgeon (Ogden *et al.* 2013) and sea anemone (Reitzel *et al.* 2013) are examples in non-model marine species, where NGS highlighted previously undetected demographic and evolutionary patterns. Even though the number of NGS-based genotyping studies has increased exponentially over the last few years, there has been little investigation into the possible gains that such data offer for deciphering fine scale population structure in non-model marine species (Lamichhaney *et al*. 2012; Nielsen *et al*. 2012; Pujolar *et al*. 2013; Hess *et al.* 2013) and particularly in invertebrate species (Reitzel *et al.* 2013).

Performance of assignment methods depends mainly on the degree of population differentiation among candidate source populations, sample sizes of individuals and the number of markers used (Cornuet *et al.* 1999; Bernatchez & Duchesne 2000; Banks *et al.* 2003). In principle, genotyping thousands of SNP markers in a large number of individuals should help circumvent these constraints. However, we are not aware of any study that specifically investigates the improvement of assignment methods through the use of large sets of NGS markers in situations of weak genetic differentiation (typically $F_{ST} < 0.01$).

The main goal of this study was to assess how NGS could help delineate the genetic structure of the American lobster (*Homarus americanus*) throughout much of the species' range, and increase the assignment success of individuals to their location of origin. The American lobster (henceforth lobster) supports one of the most valuable fisheries in North America. Its distribution ranges from Cape Hatteras (North Carolina, USA) in the South to the Strait of Belle Isle (Labrador, Canada) in the North. Typically inhabiting coastal waters less than 50 m deep, lobster can be found offshore in some localities at depths reaching 700 m (Cooper & Uzmann 1971). The carapace length at which 50% of females are sexually mature decreases with increasing temperature and varies from about 70 to 108 mm depending on locality (Watson *et al.* 2013). Mating and spawning occur during summer, usually one or more years apart, and larvae are hatched after an incubation period of 11 to 12 months on the abdomen of the female (Templeman 1940; Waddy *et al.* 1995). The planktonic/pelagic larval phase lasts on average 3-6 weeks and its duration is inversely related to temperature (Ennis 1986; Quinn *et al.* 2013).

Early studies based on allozymes and random amplified polymorphic DNA (RAPD) revealed virtually no genetic differentiation in lobster from geographically separate regions (Tracey *et al.* 1975; Harding *et al.* 1997). More recently, Kenchington *et al*. (2009) conducted a detailed study of lobster along the northeast coast of North America with 13 microsatellite markers. A north-south genetic discontinuity centered on southwest Nova Scotia was detected, and a weaker, smaller-scale substructure was revealed in the southern region but not in the northern region. Weak genetic structure in American lobster might reflect potential for extensive dispersal (Incze & Naimie 2000; Xue *et al.* 2008) via ocean currents during the long pelagic larval period (Ennis 1986). Adult lobsters have also been shown to undertake extensive seasonal migrations over distances of up to 100 km in some regions (Campbell 1986), but also exhibit homing behavior (Pezzack & Duggan 1986). Moreover, as mating and larval release may be separated in time by about 2 years (Waddy *et al.* 1995), these events may occur in different locations, and mating rather than larval release could determine genetic patterns. Therefore, the contribution of adult movements to gene flow and population structure remains unclear.

In this study, we genotyped 586 adult American lobsters collected in 17 locations using 10,156 SNPs discovered by RAD-sequencing. We first document the regional and finer-scale population genetic structure among the sampled locations and then quantify the efficiency of assignment tests as a function of number of SNPs used and sample size per location. Finally, we discuss the benefits of genotyping a large number of SNP markers for the study, conservation and management of the American lobster as well as other marine species that experience high levels of gene flow.

2.4. Results

Genotyping results

The average number of sequence reads among the 16 libraries was 169 million (range: 112-189 million) and the average number of quality filtered reads per library was 130 million (range: 87-156 million), providing an average depth of coverage per individual over all SNPs of 43x and a mean depth per nucleotide position ranging from 18x to 448x. Thirty-eight individuals (~6.0%) had an insufficient mean coverage (<10x) and were removed from

subsequent analyses. After applying the different filtering steps, 10,156 SNPs were retained for subsequent analyses (Table 2.2).

Selecting candidate SNPs for demographic inference

From the 10,156 SNPs retained, a genome scan using ARLEQUIN detected 8645 SNPs seemingly not under selection (~85.1%), 406 SNPs (~4.0%) under divergent selection and 1105 SNPs (~10.9%) potentially under balancing selection. BAYESCAN identified 8324 SNPs (~82.0%) seemingly not affected by selection, 32 SNPs (~0.3%) potentially under divergent selection and 1800 SNPs (~17.7%) potentially under balancing selection (Figure S2.1). Here, we used the most conservative neutral model available in BAYESCAN (pr_odds = 10,000) in order to minimize false positives detected as being under positive or balancing selection (Lotterhos & Whitlock 2014). The finding of a high number of SNPs potentially under balancing selection may also support several studies suggesting or showing that balancing selection is more prevalent in the genome than previously expected (Nielsen 2005; Charlesworth 2006; Shimada *et al.* 2011). In addition, SNPs detected as being under balancing selection could also be defined as being nearly all monomorphic, which is a general feature of samples from natural populations (Roesti *et al*. 2012). Subsequent inferences of genetic structure were carried out using the 8144 SNPs (~80.1%) candidate markers that were concluded not to be under selection by both BAYESCAN and ARLEQUIN.

F-statistics

Our results showed that the majority of sampling locations were genetically differentiated. Average F_{ST} was 0.00185 across all 8144 SNPs and all pairwise comparisons of the 17 sampling sites ranged from 0.00002 (BRO vs. OFF) to 0.00374 (BON vs. BRO) (Table S 2.1). Overall, 129 out of the 136 pairwise comparisons of genetic differentiation between sampling locations were significant (P-value < 0.05), which resolved 11 genetically distinguishable populations among the 17 sampling sites. Eight out of these 11 putative populations corresponded to unique sampling locations (BON, BOO, BRA, CAR, CAN, SEA, RHO, TRI) and three (hereafter South Gulf of Saint Lawrence = SGL, Southwest Nova Scotia = SNS and Cape Cod = CCO) clustered together neighboring sampling locations (SGL: GAS, SID, MAG, MAL; SNS: BRO, LOB and OFF; CCO: MAR and BUZ). Average F_{ST} was 0.00199 across all SNPs and the 11 putative populations, and ranged from 0.001 (SNS vs. SEA) to 0.00374 (BOO vs. SNS). Significant P-values for most of the pairwise

comparisons of genetic differentiation were consistent with the very narrow 95% confidence intervals around F_{ST} estimates, which averaged ±0.0006, and never encompassed zero for all the significant comparisons (Table S 2.1).

Both the heatmap and the dendrogram based on F_{ST} values separated samples belonging to the north region from those belonging to south region of the sampled lobster distribution range (Figure 2.2). The heatmap illustrated the dichotomic nature of the F_{ST} values, with lower F_{ST} values generally observed between sampling locations within each of the two large geographic regions (north or south), and higher F_{ST} values between locations belonging to the different geographic regions (Figure 2.2). The AMOVA showed a modest yet highly significant net genetic differentiation between samples from the north and the south regions (F_{CT} = 0.0011, P-value = 0.0002; Table 2.3). The variation between sampling locations within each region was also significant (F_{ST} = 0.0010, P-value < 0.0002) and equal to the one found between regions (Table 2.3). We detected a strong and highly significant positive association between genetic and coastal geographic distances (r^2 = 0.56, P-value < 9.999e-05) when considering all pairwise comparisons (Figure 2.3). This association was still significant, albeit weaker, when considering samples only within the north region (r^2 = 0.41 and P-value = 0.046) or the south region (r^2 = 0.20 and P-value = 0.049).

Clustering of individuals and populations

The genetic split between north and south regions was also discerned by both the DPCA and *K-means* analyses but not by STRUCTURE and ADMIXTURE. Thus, all lobsters analyzed were grouped into a single cluster according to STRUCTURE and ADMIXTURE when using 8144 potentially neutral SNPs. The same result was obtained when we included all 10,156 SNPs (results not shown). In contrast, the DAPC revealed two clusters, according to the lowest BIC, separated along the first discriminant function (PC1), which explained 33.62% of the total genetic variation among individuals (Figure 2.4). Discriminant functions 2 (PC2), 3 and 4 accounted for 6.27%, 3.84%, and 2.28% of the variance, respectively, and did not reveal any particular clustering (results not shown). Although there was some overlap between the two groups, the first cluster resolved by discriminant function 1 corresponded mainly to individuals from the north region, whereas the second cluster contained mainly individuals from the south region (Figure 2.4). Moreover, an optimal *K* of 2 clusters, corresponding to the north-south separation, was found when performing the analysis at the

population level using the pseudo-*F*-statistics (Figure 2.4).

Individual assignment analysis

The assignment success of individuals to their respective sampling locations was strongly affected by the number of SNPs used that were ranked based on their average F_{ST} value across all sampling locations (Figure 2.5). Thus, the average assignment success to sampling location increased with the number of SNPs from 60.2% when using the top 500 most differentiated SNPs to a maximum of 80.8% using the top 3000 most differentiated SNPs and then decreased to only 8.9% using all 10,156 SNPs. Regarding the effect of individuals sampled per location, increasing the number of individuals from 10 to the maximum average of 34 increased the proportion of individuals (using the top 3000 SNPs) correctly assigned to their location of origin from 13.7% (range: 0 - 50.0%) to 80.8% (range: 56.6 – 95.6%) on average (Figure 2.5). Visual inspection of this relationship indicates that sampling a greater number of individuals than were sampled in this study would have generated additional gains in assignment success.

At the regional scale, GENODIVE assigned lobsters to their region of origin with very high success. Lobsters sampled in the north and south regions were re-assigned correctly at 93.6% and 94.8% respectively, using the top 3000 most differentiated SNPs. At the population level, that is considering the 11 putative genetically distinct populations as defined above, assignment success was lower than between the north and south regions but still high with an average of 80.8%. However, assignment success was highly variable depending on population, ranging between 55.5% (CAN) and 95.6% (SGL) (Figure 2.6). Interestingly, the lowest assignment success is for a site (CAN) along the Scotia shelf where there might be a discontinuity in structure between the north and south regions. We also estimated assignment success for sampling sites that were pooled together as representing a same putative population based on F_{ST} values and alpha = 0.05. Assignment success was still high for these sites, averaging 77% for sites within SGL (GAS, MAL, MAG, SID), 78% within SNS (LOB, OFF and BRO) and 83% within CCO (MAR and BUZ) (Figure S2.2). As expected, miss-assigned individuals were generally assigned to other sites within each of these three putative populations. This indicates that despite the lack of statistically significant genetic differences between sites that were pooled as representing a same putative population, individuals from a given site were genetically more similar among themselves than they were to lobsters from

other sites.

We found only 140 pairs of loci with an r^2 value > 0.5 in all sampling locations. Indeed, non-independence of markers was expected to be low since the lobster genome is several times larger than that of many marine fish (~ n=69 chromosomes: Coluccia *et al.* 2001; C=4.75: Genome size database). We randomly removed one of the linked SNPs and we assessed assignment success again using the remaining 2 860 SNPs. Assignment success obtained in this case was very similar to assignment success using all 3000 SNPs, with on average 93.7% (instead of 94.2%) individuals correctly assigned to their region of origin and 79.6% (instead of 80.8%) individuals correctly assigned to their population of origin.

When using the randomized dataset, less than 5% of individuals were correctly assigned to their location of origin, clearly indicating the rejection of the null hypothesis of random assignment based on the empirical data set. In contrast, the bootstrapped dataset (using the top 3000 SNPs) gave a high assignment success of 80.3% on average, which is similar to the primary dataset, further validating results of the assignment tests. However, assignment tests could not confidently tell apart migrant individuals from incorrect assignments, since no individuals were outside the 95% likelihood limits of their respective population. When sampling locations were considered separately, GENECLASS2 and GENODIVE gave a similar assignment success (average 81.5% and 80.8% respectively, Student's t-test, P-value = 0.82). Correlation between assignment successes obtained by both methods for a given population was also high (Rho = 0.84), indicating largely consistent conclusions between the two programs.

2.5. Discussion

The main goal of this study was to assess how using thousands of SNPs could help to better delineate fine-scale genetic structure and increase the assignment success of individuals to their site, putative population and region of origin in weakly genetically structured marine species using the American lobster as a case study. Results revealed the existence of a hierarchical genetic structure, first separating populations from the north and the south regions of the sampled range and then separating populations within each of these regions. Thus, 11 putative populations were resolved out of the 17 sampling locations,

revealing population genetic structuring at finer spatial scale than previously revealed for this species. On the other hand, whereas F_{ST} values were often highly statistically significant, they were always small and comparable to values frequently reported for other species of marine vertebrates and invertebrates. These small F_{ST} values suggest pronounced genetic connectivity among sites and putative populations or recent separation and slow approach to equilibrium in very large populations (Marko & Hart 2011). However, contrary to earlier studies on this species, confidence intervals around F_{ST} estimates were very narrow and excluded zero, as a consequence of the very large number of markers used. Results from the assignment tests provided further evidence for this general pattern of population structuring since 94.2% of individuals were correctly assigned to their region of origin and 80.8% were correctly assigned to their putative population of origin within each region. In addition, assignment success remained high when assigning individuals to sampling locations that were not significantly differentiated based on F_{ST}, indicating that lobsters from the same location are genetically more similar among themselves than they are with individuals from other locations. Overall, these results confirm the resolution gained by using a large number of SNP markers to delineate fine scale population structuring and to perform assignment tests in highly genetically connected marine species (Waples & Gagiotti 2006). Below, we discuss the implications of these findings for the study, conservation and management of American lobster and other highly connected marine species.

Fine-scale population structuring

The small yet significant genetic differentiation found among 94.8% of the pairwise sites comparisons, along with generally high site, population or regional assignment success, contributes to a growing literature finding that many marine organisms are subdivided into genetically separated units, sometimes at small spatial scales (e.g. Atlantic cod, *Gadus morhua*: Ruzzante *et al.* 1999, Knutsen *et al.* 2003); flathead mullet, *Mugil cephalus*: Krück *et al.* 2013); Atlantic herring; Pacific lamprey, *Entosphenus tridentatus*: Hess *et al.* 2013), which has changed the general perception that most marine species are panmictic across broad geographic scales (Swearer *et al.* 1999; Mora & Sale 2002; Banks *et al.* 2007; Iacchei *et al.* 2013). In the particular case of American lobster, earlier studies did indeed suggest that the species was panmictic over large geographic areas (Tracey *et al.* 1975; Harding *et al.* 1997). However, Kenchington *et al*. (2009) provided evidence of a north-south discontinuity

in genetic structure that is corroborated by the genetic structure observed with SNPs reported herein. Kenchington's study also showed a fine-scale genetic structure in the southern region, but not in the northern region where panmixia was proposed. In contrast, our results suggested the existence of 6 populations among the 9 sampling sites from the northern region. Although genetic differences were small and variable depending on sampling sites comparison, they were accompanied by a relatively high assignment success. This outcome is most likely due to the increased accuracy and statistical power provided by screening thousands of SNPs across the lobster genome, as anticipated by Allendorf *et al.* (2010). Our results show that the use of thousands of SNPs returned very narrow (±0.0006) confidence intervals even around weak estimates of differentiation, therefore substantially increasing the accuracy of F_{ST} estimates. Willing *et al.* (2012) recently demonstrated via computer simulations that a large number of screened markers could be used to detect genetic differentiation as small as $F_{ST} = 0.001$, assuming there is a real genetic structure. This increased accuracy of genetic estimates may enhance our ability to relate indirect measures of gene flow and migration to demographic connectivity (that is *m,* the proportion of migrants among populations per generation), which matters more than genetic connectivity for short-term population management (Waples & Gagiotti 2006; Cano *et al.* 2008). Here, our results of population assignment suggest that at least some of the lobster putative populations might be "demographically independent", meaning that their dynamics is driven more by local birth and death than immigration and emigration (Hanski 1998). For instance, more than 89% of individual lobsters were correctly assigned for 6 of the 11 proposed populations, suggesting on average for these a maximum proportion of migrants (m) of about 0.11, that is considering that a proportion of that 0.11 most likely corresponds to spurious miss-assignment errors. Interestingly, and although this must be interpreted cautiously (Lowe & Allendorf 2010), Hastings (1993) proposed a value of $m = 0.1$ as the threshold below which populations may be considered demographically independent. Admittedly however, interpretations regarding demographic independence must be done cautiously because our study was based on egg-carrying females, which is likely to have increased detectable genetic population differentiation. Whereas this strategy was used to standardize our sampling design, it may have biased the estimates of demographic independence, to which males and juveniles may also contribute. Therefore, future studies on this species should also compare patterns of connectivity in males and juveniles.

Our findings set the stage for future research into the demographic processes that are relevant to fine-scale genetic structuring in American lobster and other weakly differentiated marine species. For American lobster, bio-physical larval dispersal models have shown that lobster post-larvae may disperse up to 300-400 km from where they hatch (Incze & Naimie 2000; Xue *et al.* 2008; Chassé & Miller 2010), but it is not known what proportion of these individuals will successfully settle and survive to recruit into the "local" reproductive adult population. Similarly, although some adults have a resident behavior year-round, most undergo seasonal movements or long-range migrations to search for overwintering habitat that protects against harsh coastal winter conditions (*e.g.*, ice scour or storms) and/or dampens seasonal thermal variability (Campbell 1986; Bowlby *et al*. 2007; Cowan *et al*. 2007). Despite the observation of long distance movements by some individuals, migrating adult lobsters tagged within the northern and southern regions defined here, including egg-bearing females, are generally recaptured within 5-10 km of their original tagging location, even after a number of years at liberty (Campbell 1986, Pezzack & Duggan 1986, Comeau & Savoie 2002). This would be congruent with the low migration rate suggested by our assignment tests. There is also evidence that adult American lobsters display homing behavior (Comeau & Savoie 2002), as reported in palinurid lobsters (*Panulirus cygnus*: Chittleborough 1974); *Panulirus argus*: Herrnkind *et al.* 1975); *Jasus edwardsii*: Kelly & MacDiarmid 2003); *Panulirus versicolor*: Frisch 2007); *Palinurus elephas*: Follesa *et al.* 2009). Homing behavior could result in large groups of adults belonging to a same population segregating to their coastal areas for reproduction, independent of other such groups, thereby potentially reducing genetic connectivity even if the adults undergo long-range migrations at certain times of the year (Lawton & Lavalli, 1995).

Our results indicated that isolation by distance does play a role in the observed pattern of genetic structure and that this was not only driven by the hierarchical separation between the south and north regions, since significant isolation by distance existed within as well as between regions. Clearly, this underlines the need for a more comprehensive study investigating the impact of factors other than geography in determining the genetic structure of American lobster. This, however, was beyond the scope of this paper and will be treated elsewhere (see Chapter 4). Namely, integrating larval dispersal and consideration of additional environmental factors (*e.g.*, ocean temperature, salinity, bottom topography, coastline) into a seascape genetics framework could help better understand the ecological

determinants underlying the observed pattern of genetic structure in lobster, similar to previous works on highly connected marine species (Banks *et al.* 2007; White *et al.* 2010; Selkoe *et al.* 2010).

Hierarchical structure between south and north regions

The genetic distinctiveness of the north and south regional groups of populations was previously interpreted as the result of a range expansion from South to North following the end of the last glacial period, approximately 10,000 years BP (Kenchington *et al.* 2009). An additional explanation could lie in oceanographic features that promote larval exchange and retention within each of these two regions (Urrego-Blanco & Sheng 2014). Moreover, the direction of larval dispersal between the two regions is likely constrained by the dominant southwesterly current outflow from the Gulf of St. Lawrence to the Gulf of Maine via the Atlantic coast of Nova Scotia, and not the other way around. At the mid-Scotian Shelf, off Mahone Bay, the surface currents disperse larvae away from the coast (Hannah *et al.* 2001), and this could act as a barrier to gene flow, assuming the larvae do not survive. This hypothesis is in agreement with previous studies showing difference in productivity between southern and northern populations along the Nova Scotia (reviewed by Miller 1997). At the same geographic area than our study, a recent genetic study also revealed the existence of a north/south dichotomy in northern shrimp (*Pandalus borealis*) that could be explained by oceanic circulation and temperature variation (Jorde *et al.* 2015). That being said, the net genetic differentiation between the north and south regions identified here was weak, which is also consistent with physical oceanographic studies suggesting that a proportion of larvae may drift through the strong Scotian Shelf current every generation and translate into long-term and pronounced genetic connectivity (Hannah *et al.* 2001). As discussed above, however, we cannot exclude the possibility that the weak differentiation between lobsters from the two regions may also reflect their very recent divergence along with presumably large effective population sizes. On the other hand, the assignment tests indicated again that the proportion of migrants between the two regions is very low. Thus, the 94.2% assignment success within each region suggests a short-term demographic independence between the two regional groups. This is also consistent with results of all the tagging studies involving adult lobsters, which report no long distance movements between the Gulf of Maine and Gulf of St. Lawrence lobsters (Lawton & Lavalli, 1995).

The use of clustering software (DAPC, pseudo *F*-statistics, AMOVA, STRUCTURE and ADMIXTURE) with different sensitivities to uncover subtle population structure resulted in contrasting findings. STRUCTURE and ADMIXTURE did not reveal any genetic structure (either regional or local) whereas DAPC, pseudo *F*-statistics and AMOVA showed a significant division between the south and north regions. This is congruent with simulations studies (Waples & Gagiotti 2006; Kalinowski 2010) showing that Bayesian clustering methods fail to detect any genetic structure when genetic divergence is very low ($F_{ST} < 0.01$). Apparently, this still holds true even when using thousands of markers, as suggested in this study. Also, Kanno *et al*. (2011) and Jombart *et al*. (2010) showed the efficiency of DAPC to discern significant genetic clusters where STRUCTURE failed to detect any signs of clustering in the system. Thus, DAPC appears more efficient than STRUCTURE at detecting population clustering in systems of weakly ($F_{ST} < 0.01$) differentiated populations.

Assignment success as a function of number of markers and sample size

Several simulation-based studies and analytical models previously demonstrated that correct assignment varies as a function of the number of markers and individuals used (e.g. Cornuet *et al.* 1999; Bernatchez & Duchesne 2000; Paetkau *et al.* 2004). Here, our results empirically illustrate how the potential of using a large number of SNP markers may enhance the resolution of assignment methods for weakly differentiated populations. However, while we showed how increasing the number of markers genotyped up to a maximum of 3000 top ranked markers improved assignment success, beyond that number the assignment success decreased gradually, indicating that more weakly differentiated markers added noise and contributed to blurring rather than improving assignment. We believe that this is most likely due to a sampling error (arising from too few individuals being analyzed), which is stronger on weakly differentiated markers with only modest allele frequency differences between populations relative to more differentiated markers (Roques *et al.* 1999). It would be important in future studies to assess whether this pattern of decreasing assignment success beyond a given number of top rank markers will be generalized in other marine species with similarly weak population structure. As for the effect of sample size, our results showed that our maximum number of individuals per sampling location in total (n=34 on average) was not sufficient to reach the highest assignment success possibly attainable in this system with the top 3000 markers. Clearly, further improvement in assignment success could have

potentially been gained by substantially increasing the number of individuals genotyped per sampling location.

Management implications

Our sampling design was largely based on obtaining samples belonging to different spatial units currently used for lobster management in the Northwest Atlantic (*e.g.,* Lobster Fishing Areas [LFAs] in Canada; Figure 2.1). Interestingly, the pattern of structuring we observed generally fitted these LFAs in the sense that most sampling sites representing different LFAs were genetically differentiated and lobsters belonging to different LFAs were often reassigned with high success. In some cases, however, such as the South Gulf of St. Lawrence, samples from different LFAs were not different based on F_{ST} values and assignment success was reduced, albeit remaining markedly more important than expected by chance alone. This suggests that there is a geographic distance below which demographic dependence may occur. Therefore, future studies should aim to refine the geographic scale of structuring by applying a sampling design including different geographic scales, many samples from the same LFA, different lobster life stages from larvae to adults, and both genders. Moreover, the temporal and seasonal stability of population structure should be addressed in order to properly document the match between population structure and management units. Finally, and although sample sizes should be increased, the promising results of individual assignment to their population of origin indicates that a lobster SNP database covering most, if not all populations, could also provide new informative tools in the context of commercialization and marketing of American lobster. For instance, in the context of eco-certification and increased consumer awareness, such a database could provide a means for local managers and fishermen to define territorial branding. Moreover, the application of population assignment based on such a database could improve the traceability from fishers to consumers (*e.g.*, FisPopTrace Consortium; Nielsen *et al.* 2012). We envision a bright future for the use of high-density genotyping facilitated by NGS-based genotyping protocols, both for improving our basic knowledge of population genetic structure of highly connected marine species and for using that knowledge to improve management and conservation practices of exploited species.

2.6. Methods

Sampling

We collaborated with commercial fishermen to sample lobsters from 17 locations throughout much of the species' range, 15 that were inshore and two that were offshore (Figure 2.1). Sampling was done between May and August 2012. We only sampled adult females bearing late-stage eggs that would hatch in the coming weeks (n = 624 total), to standardize the sampling design and to estimate the genetic structure of individuals that had survived to reproduce. We reasoned that this sampling design would perhaps be most likely to reveal genetic structure, particularly if females displayed homing behavior related to spawning and hatching (Pezzack & Duggan 1986). The second walking leg of each individual was removed and preserved in 95% EtOH until DNA extraction. A total of 36 individuals were sampled in all but one study location (n = 48 for MAG).

Molecular techniques

Genomic DNA was extracted using a salt-extraction protocol (Aljanabi & Martinez 1997) with additional RNase A treatment (Qiagen) following the manufacturer's recommended protocols. DNA integrity (i.e., presence of degradation or smears) was inspected on a 1% agarose gel. Samples with degraded DNA were excluded. Extracted genomic DNA (gDNA) was quantified using Quantit Picogreen dsDNA assay kits (Invitrogen). RAD-sequencing libraries were prepared following a protocol modified from Miller *et al.* (2007) (see Supplementary materials). Each library contained 48 individuals barcoded with a unique six-nucleotide sequence. Real-time PCR was used to quantify libraries. Single read, 100 bp target length, sequencing on Illumina HiSeq2000 platform was conducted at the Genome Quebec Innovation Centre (McGill University, Montreal, Canada).

Bioinformatics and genotyping

The libraries were demultiplexed using the *process_radtags* program in STACKS v.1.09 (Catchen *et al.* 2013). Polymorphic SNPs were identified on reads truncated to 90 bp and filtered for overall quality and presence of barcodes. The formation of RAD loci was allowed with a maximum of two nucleotide mismatches ($M = 2$) - identified as an optimum threshold according to the method developed by Ilut *et al.* (2014) - and a minimum stack depth of three ($m = 3$) among reads with potentially variable sequences (*ustacks* module in

STACKS, with default parameters*)*. Then, reads were aligned *de novo* with each other to create a catalogue of putative RAD tags (*cstacks* module in STACKS, with default parameters). In the *populations* module of STACKS and following consecutive filtering steps, we first retained RAD tags with a minimum stacks depth (*m*) of 10 to a maximum stacks depth of 100. This step removed SNPs genotyped with too low coverage ($m < 10$) to be accurately called as well as SNPs genotyped with too high coverage ($m > 100$), which could be located on highly overrepresented sites due to repeats in the lobster genome. Then, we retained SNPs genotyped in at least 70% of the individuals and 70% of the sampling locations. Potential homeologs were excluded by removing markers showing heterozygosity > 0.50 within samples (Hohenhole *et al.* 2011). We also removed markers out of Hardy Weinberg equilibrium (P-value = 0.01) at more than 60% of the locations. Individuals and SNPs with more than 30% of missing data were also eliminated. To avoid bias in the estimation of the baseline differentiation and eliminate any sequencing and PCR error from the SNP dataset, polymorphisms with a minor allele frequency (MAF) > 0.1 in at least one location (i.e. minor allele occurring at least 4 times in one location) and polymorphisms with MAF > 0.05 on average across sampling locations were kept. It has been shown that very low frequency SNPs (MAF < 0.05) create biases in quantifying genetic connectivity, and should therefore be removed when inferring demographic processes (Roesti *et al.* 2012). Details of the number of SNPs kept after each filtering step are provided in Table 2. The resulting filtered VCF file was converted into the file formats necessary for the following analyses using PGDSPIDER v.2.0.5.0 (Lischer & Excoffier 2012).

Detecting SNPs under selection

SNPs potentially under balancing and divergent selection should also be removed when assessing genetic connectivity between populations (Beaumont & Nichols, 1996; Luikart *et al.* 2003). This was achieved using BAYESCAN v.2.1 (Foll & Gaggiotti 2008) as well as the Fdist approach (Beaumont & Nichols, 1996) implemented in ARLEQUIN v.3.5 (Excoffier, 2010). BAYESCAN estimates population-specific F_{ST} coefficients by the Bayesian method described in (Beaumont & Balding 2004) and uses a cut-off based on the mode of the posterior distribution to detect SNPs under selection (Foll & Gaggiotti 2008). SNPs with a posterior probability over 0.95 were considered as outliers, after running 100,000 iterations on all samples together (i.e., not pairwise, with remaining default parameters). We specified a

'prior' odd of 10,000, which set the neutral model being 10,000 times more likely than the model with selection in order to minimize false positives (Lotterhos & Whitlock 2014). ARLEQUIN was executed with 200,000 simulations and 100 demes simulated as recommended by the authors, and SNPs were considered as outliers based on their F_{ST} and P-value.

Individual and population clustering

We first inferred population structure by using two Bayesian clustering methods that are implemented in the programs STRUCTURE v2.3.4 (Falush *et al.* 2003) and ADMIXTURE v1.23 (Alexander *et al.* 2009). Both programs provide a means of identifying the best value for *K*, the number of putative populations. With STRUCTURE, we used 10,000 burn-in iterations followed by another 10,000 Markov chain Monte Carlo (MCMC) steps assuming an admixture model based on individuals and including no prior information on sampling location. We ran ADMIXTURE using 20,000 bootstraps. For both programs, we varied the number of groups (*K*) from 1 to 17 with 5 iterations for each value and stabilization of parameters was checked for this length of burn-in and MCMC. We then performed a Discriminant Analysis of Principal Components (DAPC) in the R package *adegenet* (Jombart *et al.* 2010), without prior information on group individual populations, and we used the function *find.clusters* to assess the optimal number of groups with the BIC (Bayesian Information Criterion) method. The DAPC is a non-model-based method, which maximizes differences between groups while minimizing variation within groups. Therefore, retaining too many discriminant functions with respect to the number of populations can lead to over-fitting the discriminant functions, which results in spurious discrimination of any set of clusters. To avoid this bias, we evaluated the optimal number of discriminant functions (n=100) to retain according to the optimal α-score obtained from our data (Jombart *et al.* 2010). In addition, a *K-means* clustering analysis was performed on sampling locations with the GENODIVE v.2.0b25 program (Meirmans & Van Tienderen 2004), using simulated annealing and testing for *K* clusters from 1 to 10, for 5000 permutations. This analysis provides the Calinski-Harabasz pseudo-*F*-statistic for determining the number of clusters (Caliński & Harabasz 1974).

Population differentiation

The extent of pairwise population differentiation was quantified using the unbiased

F_{ST} estimator θ (Weir & Cockerham 1984) and 95% confidence intervals were calculated for each pairwise comparison based on 5000 permutations using GENODIVE. Significance of the observed F_{ST} values was determined by running 10,000 permutations and assessed against a FDR-adjusted P-value to account for multiple testing (Benjamini & Hochberg, 1994). We used the function *hclust* available in the R package *ggdendro* to create a UPGMA dendrogram based on the F_{ST} values. A heatmap was produced in order to illustrate the F_{ST} matrix considering four different F_{ST} groups delimited from the distribution of pairwise F_{ST} values (see Results). A hierarchical Analysis of Molecular Variance (AMOVA) (Excoffier *et al.* 1992) based on north vs. south regional groupings (see Results) was performed. In addition, we conducted three standard Mantel tests to correlate genetic distances (F_{ST}) and natural logarithm of geographical distances. Geographic distances between each pairs of sampling locations were calculated considering the contour of the coast, using ArcGIS software. The first Mantel test included all pairwise comparison whereas the two others were based only on pairwise comparison of samples belonging to the same region (either south or north) in order to take the spatial dependence in the data into account (Meirmans 2012). The Mantel tests were performed with the library *adegenet* (Jombart *et al*. 2010) and significances of the tests were assessed using 10,000 permutations.

Population assignment

Pairwise genetic differentiation (F_{ST}) between the 17 sampling locations were calculated for each SNP using *hierfstat* library in R (Goudet 2005). All of the 10,156 SNPs were ranked according to their F_{ST}, from the highest to the lowest. As recommended by Anderson (2010), the calculation of F_{ST} and the ranking of the SNPs were based on a *training-set* of individuals (50% of the individuals for each sampling location), and the assignment success was assessed using the other, *holdout-set*, of individuals. As such, pools of individuals to select markers (*training-set*) and used to assess assignment success (*holdout-set*) were totally independent, thus circumventing the problem of *high-grading bias* (Anderson 2010). To assess the impact of the number of SNPs on the assignment test results, we performed assignment tests on subsets of SNPs (500, 1000, 2000, 3000, 4000, 5000, 6000, 7000 and 10,156 SNPs) selected according to their ranking using the *training-set* of individuals, and these subsets were tested for local assignment on the *holdout-set* of individuals. Linkage disequilibrium among markers could introduce bias when we estimated

assignment success for the different subsets of markers (Manel *et al.* 2005). We therefore tested for linkage disequilibrium between each pair of loci for the 3000 most differentiated SNPs using VCFTOOLS in order to minimize bias of linkage disequilibrium on assignment success.

To assess the impact of the number of individuals per sampling location on assignment success, we created five random datasets of 10, 15, 20, 25 and 30 individuals, which were randomly chosen without varying the number of SNPs used (using the optimal number of 3000 SNPs, see Results), and this procedure was repeated three times. Then, we performed a standard leave-one-out assignment test on these five datasets (Peatkau *et al.* 2004). In order to further test the null hypothesis that assignment estimates obtained from our empirical data set were not due to some stochastic process, we performed assignment tests on a randomized dataset with populations of identical size and randomly chosen individuals shuffled among populations. To obtain confidence intervals (CI) on estimates, we ran each assignment test on 10 generated bootstrapped datasets using repeated resampling of individuals with replacement.

Assignment tests were performed on the *holdout* set of individuals for each population both at the regional (north / south) and local (i.e. putative population) scales using GENODIVE with the *frequentist* method of Paetkau *et al.* (1995). In a given genotype, when the observed frequency of any allele was zero (a missing allele), the frequency of this allele was replaced by a fixed value of 0.005 as recommended by Paetkau *et al.* (2004), in order to avoid the calculation of a multilocus likelihood of zero. A null-distribution of likelihood values was generated using a Monte Carlo Chain (Cornuet *et al.* 1999) for 5000 permutations. In an attempt to distinguish migrants from miss-assignments, we used Cornuet's *et al.* (1999) algorithm with a statistical threshold calculated separately for every population based on an α value of 0.05 (Berry *et al.* 2004). Individuals with likelihood values of originating from their sampling location (L_H) inferior to this threshold are thus defined as putative migrants. Since the GENECLASS2 program (Piry *et al.* 2004) has been more commonly used for population assignment in previous studies (Paetkau *et al.* 2004; Berry *et al.* 2004; Castric & Bernatchez 2004), we also compared the local assignment test results obtained from GENODIVE to those given by GENECLASS2, using the same parameters (0.005 for missing alleles, alpha value of 0.05, and L_H criterion).

2.7. Acknowledgements

This research is part of the "Lobster Node" of the NSERC Canadian Fisheries Research Network (CFRN). We first and foremost wish to thank the fishermen of the Lobster Node without whom this project would have been impossible. Project design and work plan were done in collaboration with scientists from the Department of Fisheries and Oceans (M. Comeau, J. Tremblay), representatives of fishermen associations and the facilitator of the Lobster Node, M. Allain. We would like to thank A.Boudreau, V.Brzeski, Y.Carignan, Clearwater, B.Comeau, M.Comeau, JP.Allard, M.Deraspe, N.Davis, S.Delorey, C.Denton, R.Doucelle, J.Grignon, M.Haarr, R.MacMillan, G.Paulin and M.Thériault who helped to collect the samples. We are very grateful to J. Gaudin and E. Normandeau for their help in bioinformatic analyses. This manuscript was improved by comments from B. Sutherland, J. S. Moore, A. Dalziel, G. Parent, A. M. Dion-Côté and Q. Rougemont. The NSERC CFRN funded this research. We also acknowledge Paul Hohenlohe, Lorenz Hauser and the two anonymous reviewers for their comments and advices, which greatly improved the quality of the manuscript. L. Benestan was supported by a doctoral fellowship from NSERC CFRN and Réseau Aquaculture Québec (RAQ).

2.8. Tables

Table 2.1 Regional groupings of lobster sampling locations and information on locations and samples: latitude and longitude, sampling date and number of individuals successfully genotyped (N_{GEN}).

Region	Sampling location	Code	Latitude	Longitude	Sampling date	N_{GEN}
North	Malpeque Bay, PEI	MAL	46.5290	-63.6874	May-12	31
	Caraquet, QC	CAR	48.8990	-64.9289	May-12	36
	Magdalene Islands, QC	MAG	47.3790	-61.8530	Jun-12	38
	Gaspé, QC	GAS	48.7313	-64.3065	May-12	32
	Triton, NF	TRI	49.5218	-55.6107	Jun-12	35
	Bonavista, NF	BON	47.6113	-53.0088	Jun-12	32
	Dingwall, NS	DIN	46.9139	-60.4285	Jun-12	35
	Bras d'Or Lake, NS	BRA	45.7516	-60.8170	Jul-12	32
	Canso, NS	CAN	45.3362	-60.9944	Jul-12	35
South	Lobster Bay, NS	LOB	43.6792	-65.8784	Jul-12	36
	Seal Cove, NB	SEA	44.6403	-66.7199	Jul-12	33
	Boothbay Harbour, US	BOO	43.8165	-69.6897	Jul-12	35
	Marblehead, US	MAR	42.4999	-70.8578	Jul-12	34
	Buzzard's Bay, US	BUZ	41.5292	-70.8357	Jul-12	36
	Browns Bank	BRO	42.4588	-65.2083	Jul-12	35
	Georges Basin	OFF	42.1538	-66.0143	Jul-12	36
	Rhode Island, US	RHO	41.5800	-71.4774	Aug-12	35

Table 2.2. Number of putative SNPs retained following each filtering step

FROM READS TO SNPS	**SNP count**
STACKS CATALOG	200,313
POPULATION FILTERS	
Genotyped	
> 70% of the samples	74,229
> 70% of the populations	
MAF FILTERS	
Global MAF > 0.05	15,552
Local MAF > 0.1	
COVERAGE FILTER	
From 10 to 100x	15,505
HWE FILTERS	
Hardy-Weinberg equilibrium (P-value 0.05)	10,324
$H_{OBS} < 0.5$	10,156
GENOME SCAN FILTER	
Putatively neutral	8,144
Putatively under divergent selection	32

Table 2.3. Analysis of molecular variance (AMOVA) among 17 sampling locations distributed in the north and south regions of the sampled distribution range of lobster.

Source of variation	*Percentage of variation*	*Variance*	*P-value*
Between regions	0.11	0.001	0.0002
Among locations within regions	0.10	0.001	0.0002
Among individuals within locations	99.79	0.363	--

2.9. Figures

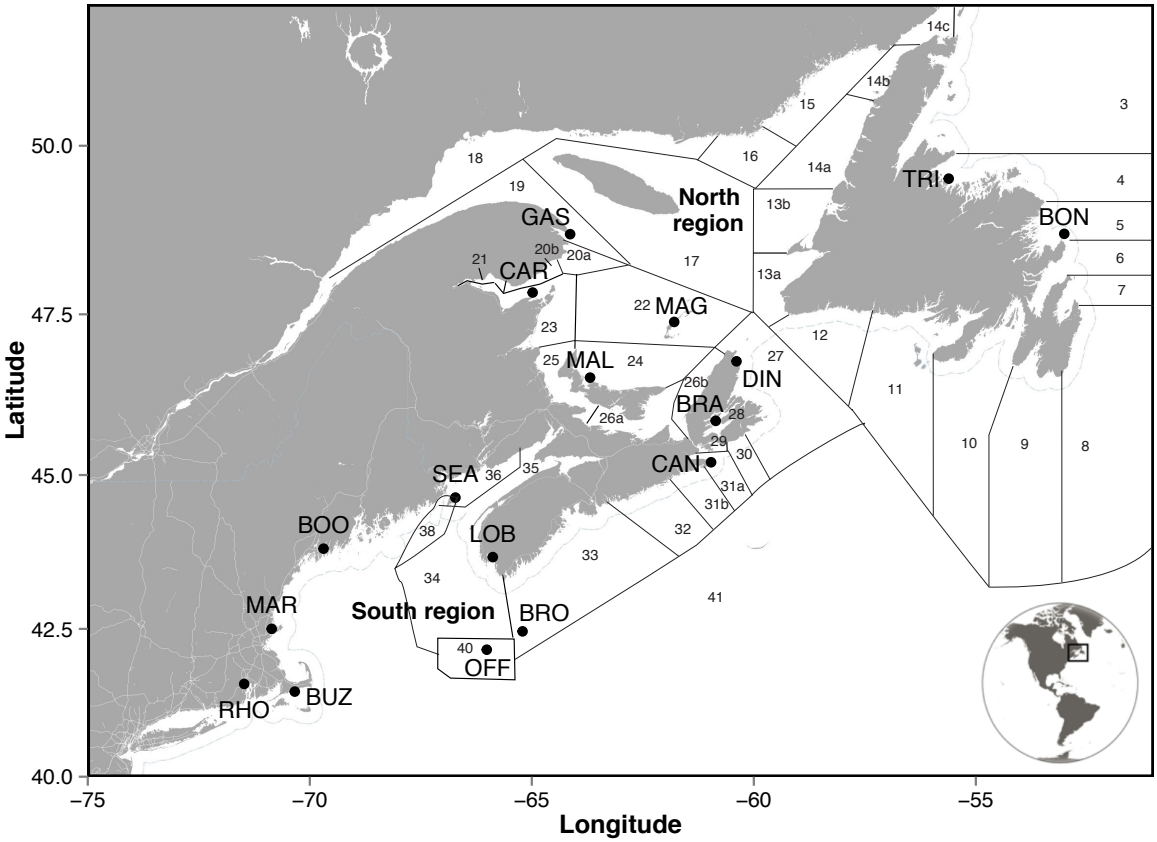


Figure 2.1. Map of lobster sampling locations.

South region including Gulf of Maine: BOO: Boothbay Harbour, BRO: Browns Bank, BUZ: Buzzard's Bay, LOB: Lobster Bay, MAR: Marblehead, OFF: Georges Basin, RHO: Rhode Island, SEA: Seal Cove. North region including Gulf of St. Lawrence: BON: Bonavista, BRA: Bras d'Or Lake, CAN: Canso, CAR: Caraquet, GAS: Gaspé, MAG: Magdalen Islands, MAL: Malpeque Bay, DIN: Dingwall, TRI: Triton. The figure also illustrates the limits of the 41 current management units in Canada (designated by a number), called Lobster Fisheries Areas (LFA).

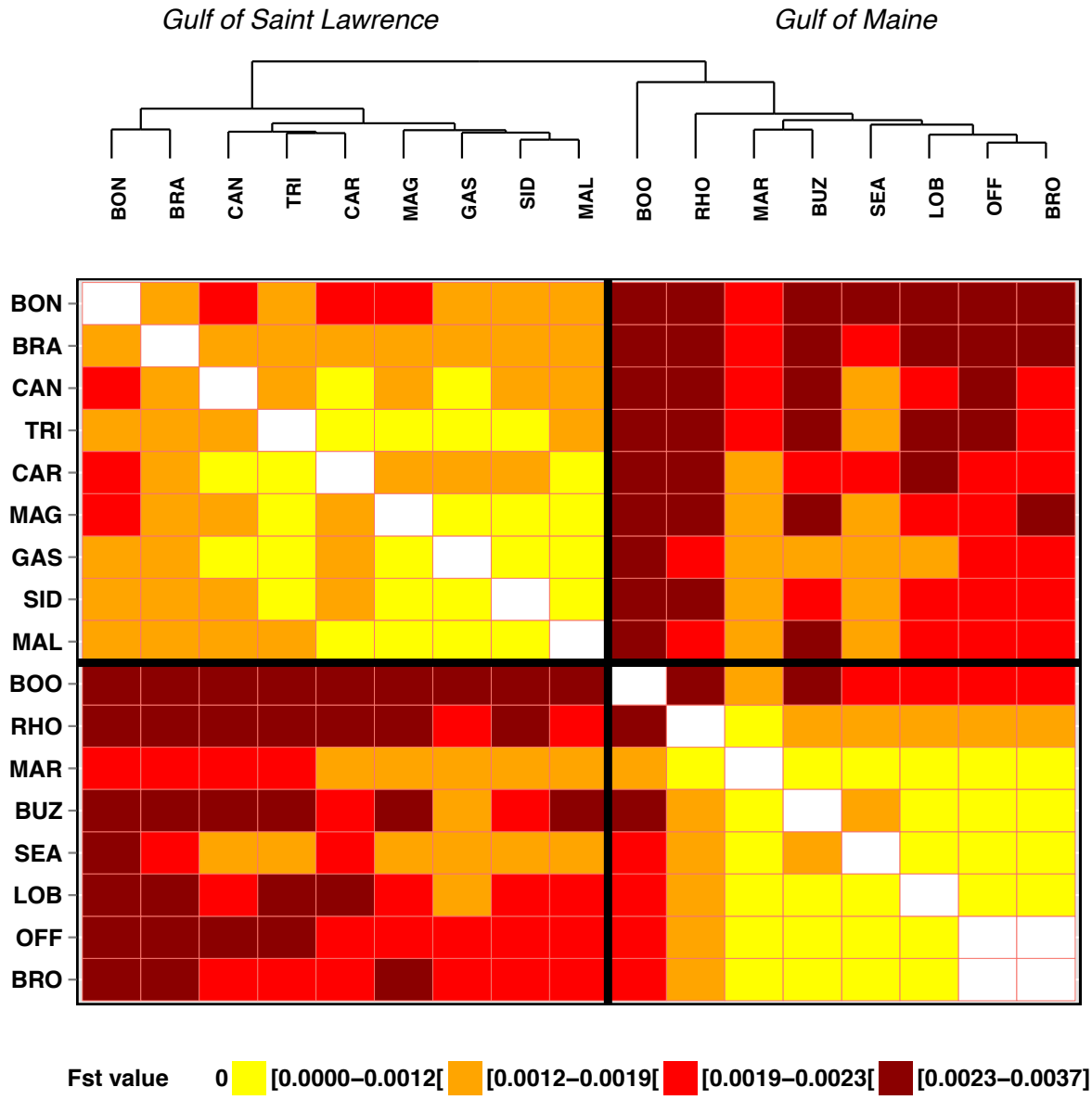


Figure 2.2. Fst population dendrogram and heatmap based on Fst values among 17 lobster sampling locations.

The heatmap color code illustrates the F_{ST} matrix considering four different F_{ST} groups delimited from the pairwise F_{ST} distribution: low F_{ST} (below 5th percentile, $F_{ST} < 0.0012$), low-intermediate F_{ST} (5th to 25th percentile, $0.0012 \leq F_{ST} < 0.0019$), intermediate (25th to 75th percentile, $0.0019 \leq F_{ST} < 0.0023$), high F_{ST} (above 75th percentile, $F_{ST} \geq 0.0023$).

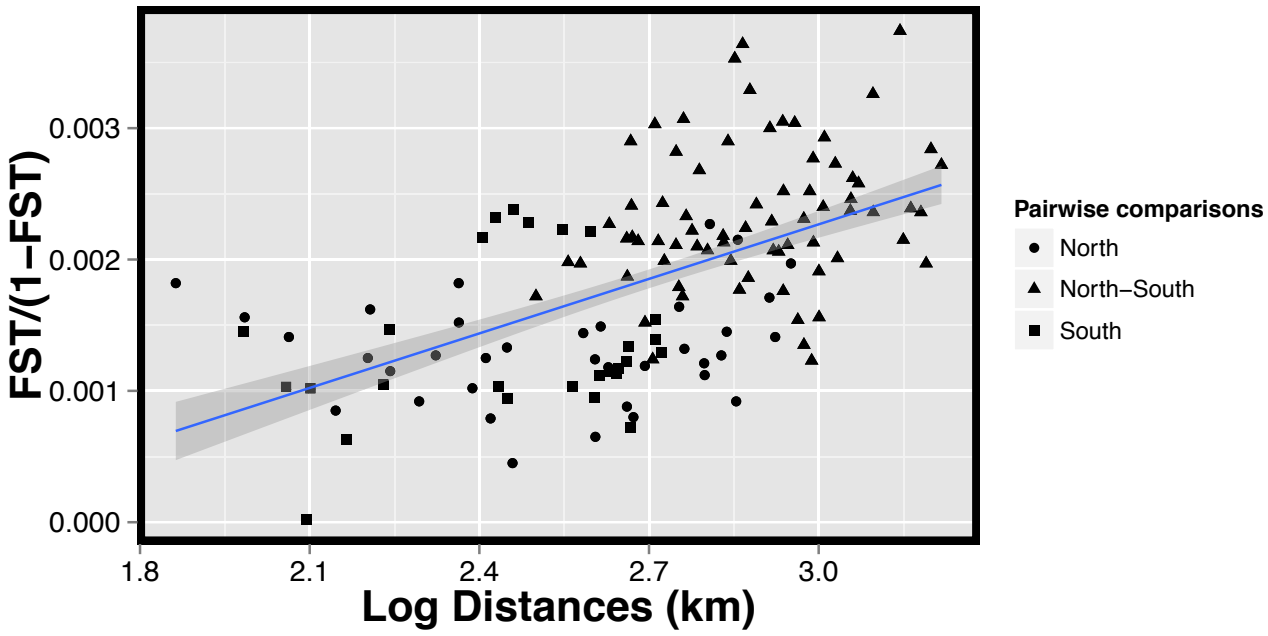


Figure 2.3. Pairwise genetic distances (F_{ST}) in relation to geographic distances (Log (km)).

Pairwise genetic distances (F_{ST}) in relation to geographic distances (Log (km)) between lobster sampling locations, with a linear regression line (in blue) fitted with 95% confidence limits (in grey). Pairwise comparisons within and between the south (SR) and north (NR) regions are represented by circles (within SR), triangles (within NR) or squares (SR vs. NR).

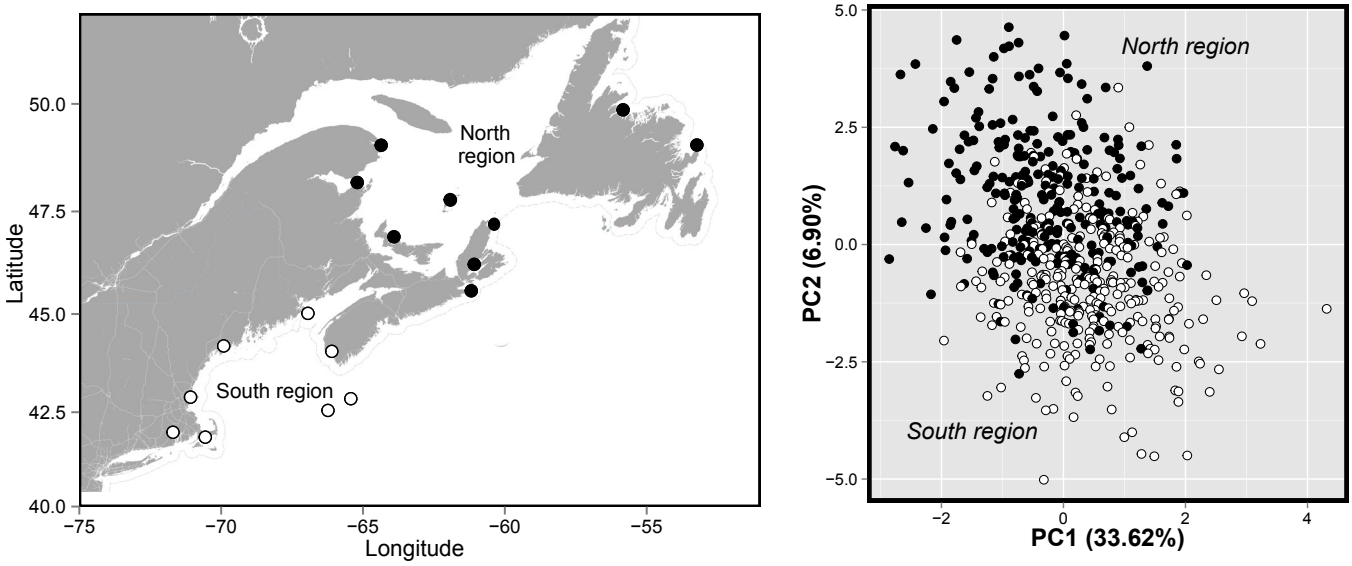


Figure 2.4. Discriminant analysis of components (DAPC) of genetic differentiation.

Right panel: Discriminant analysis of principal components (DAPC) of genetic differentiation among the 586 genotyped lobsters based on 8144 single nucleotide polymorphism markers (each point represents one individual) with principal component 1 (PC1: 33.62% of variance) against principal component 2 (PC2: 6.90% of variance); Left panel: Pseudo-*F*-statistics analysis assigning each sampling location to either the south or the north region. The individuals (left panel) and sampling locations (right panel) from the south and north regions are represented by white and black symbols, respectively.

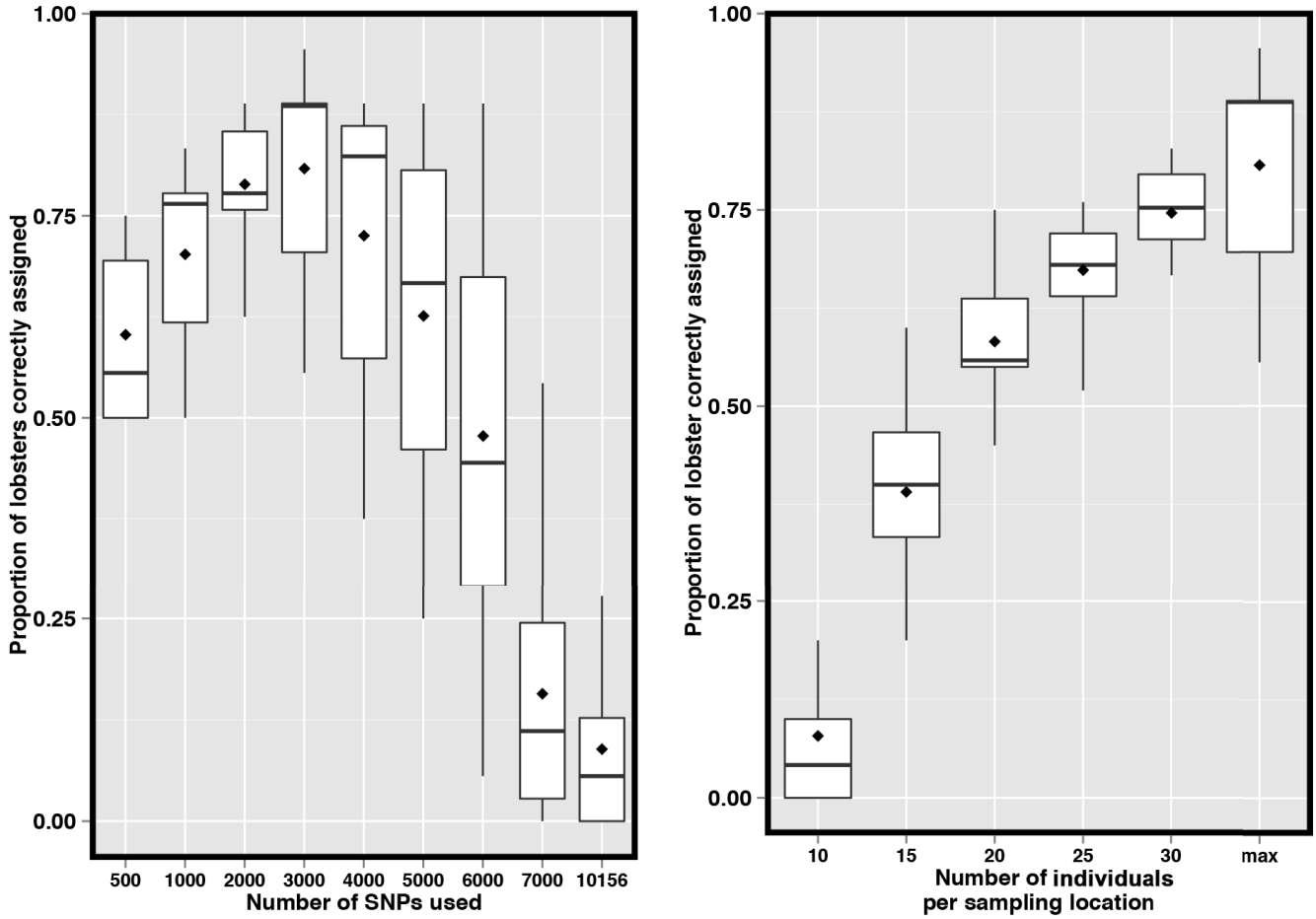


Figure 2.5. Boxplot of the assignment tests results.

Left panel: Boxplot of the proportion of lobsters correctly assigned to their sampling location (y-axis) as a function of number of SNPs ranked by decreasing order of average F_{ST} values (x-axis) following the Anderson (2010) method. Right panel: Boxplot of the proportion of lobsters correctly assigned to their sampling location (y-axis) as a function of number of individuals per sampling location and according to a standard leave-one-out procedure. The "max" label refers to the maximum number of individuals per sampling location, which varies from 31 to 38 (average 34) individuals (see Table 1). In both panels, the horizontal limits of the box represent one standard deviation around the mean (black diamond), the horizontal line within the box is the median, and the whiskers extend from the box to the 25th and 75th percentiles.

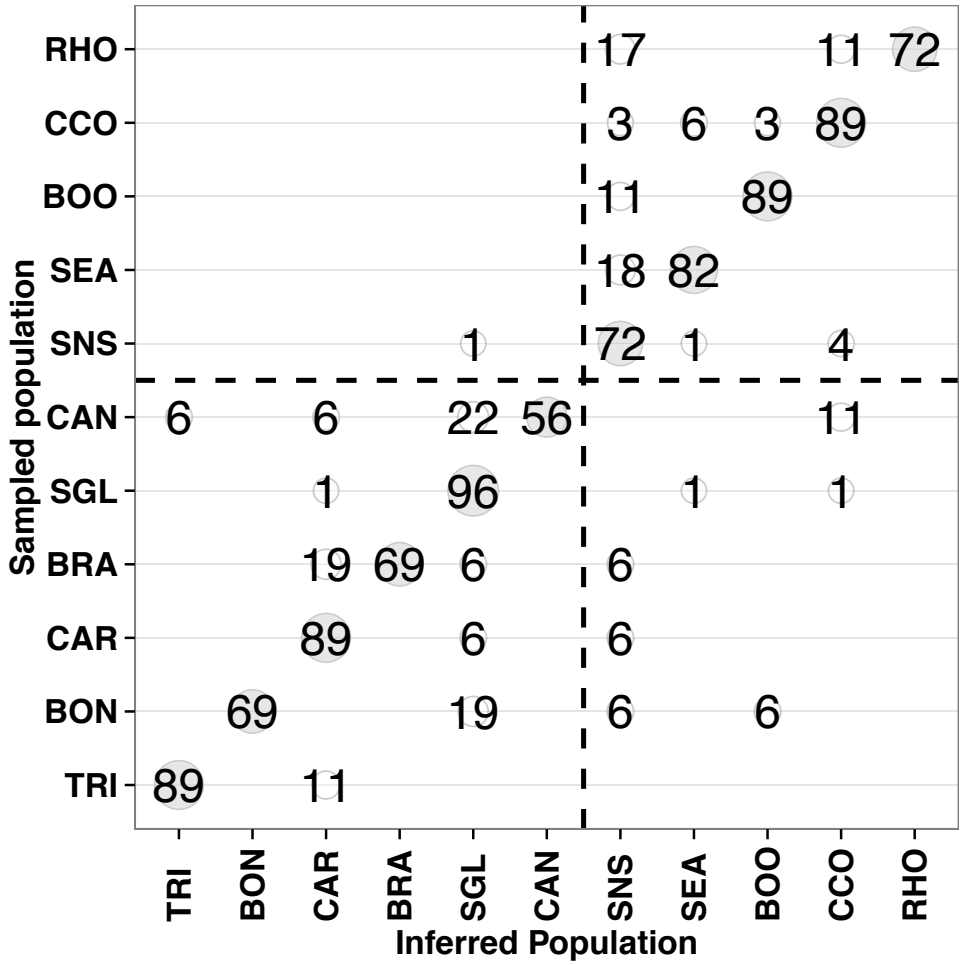


Figure 2.6. Assignment test results.

Blind assignment success expressed as the percentage of lobsters sampled from one putative genetic population that are classified into their population of origin (grey-shaded numbers on diagonal) or inferred to belong to another putative population (non-shaded numbers). Eleven putative populations were identified (see text), of which 8 were single sampling locations (BON, BOO, BRA, CAR, CAN, SEA, RHO, TRI; ordered from North-East to South-West) and three were clusters of neighboring sampling locations (South Gulf of St. Lawrence, SGL, grouping GAS, DIN, MAG and MAL; Southwest Nova Scotia, SNS, grouping BRO, LOB and OFF; Cape Cod, CCO, grouping MAR and BUZ). Dashed lines represent a higher-level genetic discontinuity separating putative populations in the south (above horizontal line on y-axis) from those in the north (below horizontal line on y-axis) of the sampled distribution range.

2.10. Supplementary materials

Protocol for RAD Library preparation

1) Dilution *(2 hours)*

   Start with approximately 500ng of genomic DNA from each sample of 96 samples. Bring the volume of each sample to 40µl with H20 (RNAse free water) and transfer into a 96-well plate. Perform short spin.

2) Digestion with *Sbf1* enzyme *(20 minutes)*

   a. Mix preparation:

	1X	50X	100X
H2O	4.5µl	225µl	450µl
NE Buffer 4 (10X)	5.0µl	250µl	500µl
Sbf1-HF (NEB R3642L)	0.5µl	25µl	50µl

   b. Add 10µl to each well

   c. Quick spin, then vortex and spin again.

   d. Incubate the plate at 37°C for 60 minutes

   e. Incubate the plate at 65°C for 20 minutes to inactivate the enzyme.

3) Barcoding *(10 minutes)*

   Add 2µl of the appropriate barcoded *Sbf1 P1 RAD adapter* (50nM) to each well in the sample plate using the multi-channel pipette -Each sample has a different adapter.

4) Ligation *(20 minutes)*

	1X	50X	100X
H20	5.9µl	295µl	590µl
NE Buffer 4 (10X)	1.0µl	50µl	100µl
rATP (100 mM, Fermentas R0441)	0.6µl	30µl	60µl
T4 DNA Ligase (NEB M0202M)	0.5µl	25µl	50µl

   a. Add 8µl to each well

   b. Quick spin, then vortex and spin again.

   c. Incubate the well at 20°C for 60minutes

   d. Incubate at 65°C for 20mn to inactivate the enzyme.

5) Pooling *(10 minutes)*

   Multiplex the 12 samples that are to be sequenced together in the same library. Perform a quick vortex.

   You will get a 720µl final volume (60µl x 12 samples).

   Take only 300µl for sonication. Store the remaining samples (420µl) at -20°C.

6) Sonication *(15 minutes each sample)*

Sonicate the multiplexed sample to produce an average fragment size of 500 bp.

Sonicator Setting:

- Power: 20%
- Time process on: 5 minutes (10*30secondes)
- Time process off: 10minutes (10*1minute)

7) Drying *(2 hours)*

Put the samples in the Speedvac for 2 hours until there is no water.

Speed Vac Setting:

- Set
- Concentrate

  NB: Place the samples inside the vacuum chamber of speed vac before you start the speed vac.
- Press Concentrate
- Check the amount of sample left in the tube. It should look translucent. Don't over dry as it will create problem in re-suspension in EB buffer

8) Add 100µl Elution Buffer to dried samples (Qiagen or self prepared (10mM Tris-HCl, pH 8.5 ideally). Wait for 2 minutes. *(3 minutes)*

9) Fragments size selection (400-600pb) *(2 hours)*

Careful: Put the magnetic particles at room temperature 30 minutes before using it

a. Add 54µl of beads. Quick vortex
b. Incubate 15 minutes at room temperature
c. Put the tubes on the magnetic plate. Transfer and retain 154µl of the supernatant into a new tube for further processing in 9d. Throw the tubes with magnetic particles.
d. Add 76µl EB buffer and 70µl of beads to the supernatant collected in step 9c. Quick vortex and spin.
e. Incubate 15 minutes at room temperature
f. Put on the magnetic plate. Throw the supernatant in a tube (or store it at -20°C with suitable labeling)
g. Wash twice the beads with 300µl of 75% Ethanol: let the Ethanol during 30 second and remove it.

   NB: In this step plate should be on the magnetic plate.

   Be sure that there is no more ethanol is left in the tube when you finish this step
h. Dry the beads for 4-5 minutes.

Avoid over-drying the magnetic beads. It will reduce the efficiency of elution significantly

i. Take the tubes out of the magnetic plate. Elute in 22 µl EB buffer. Quick vortex and spin and wait for 3 minutes.

j. Put the tubes on the magnetic plate and transfer 22µl of supernatant in a new tube

Use 1µL for nanodrop quantification.

Use 1µl for DNA Chip in order to check the fragments size.

10) Cut and blunt the fragment *(10 minutes)* 1X

Blunting buffer (10X)	2.5µl
dNTP mix (1mM)	2.5µl
Blunting Enzyme Mix (NEB E1201L)	1.0µl
Sample (from step J)	20.0 µl

Incubate at 20°C (Room temperature) for 60min

NB: Put the magnetic particles at room temperature 30 minutes before using it

11) Add 25µl of Elution Buffer to the above reaction products *(3 minutes)*

12) Purification with beads *(40 minutes)*

a. Add 50µl of beads to above reaction product. Quick vortex

b. Incubate 15 minutes at room temperature

c. Put on the magnetic plate. Throw 100µl of supernatant (or store at -20 c).

d. Wash twice the beads with 300µl of 75% Ethanol: let the ethanol for 30 seconds and remove it

Be sure that there is no more ethanol is left in the tube when you finish this step

e. Dry the beads for 4-5 minutes

Avoid over-drying the magnetic beads. It will reduce the efficiency of elution significantly

f. Elute in 42µl of EB buffer. Quick vortex and spin. Wait for 3 minutes

g. Put the tubes on the magnetic plate and transfer 42µl of supernatant in a new tube

13) Add A-overhangs to the fragments *(10 minutes)*

NE Buffer 2 (10X)	5.0µl
dATP (10mM)	1.0µl
Klenow Fragment (NEB M0212L)	2.0µl
Sample (from step 12g)	42.0µl

Incubate at 37°C for 60 minutes.

Keep the magnetic beads outside in room temperature at least 30 min before next step 14.

14) Purification with beads *(40 minutes)*

a. Add 50µl of beads

a. Incubate 15 minutes at room temperature

b. Put on the magnetic plate. Throw 100µl of supernatant ((or store at -20 c).

c. Wash twice the beads with 500µl of 75% Ethanol: let the ethanol for 30 seconds and remove it

Take care there is no more ethanol is left in the tube when you finish this step.

d. Dry the beads for 5 minutes

Avoid over-drying the magnetic beads.

e. Elute with 43µl of EB buffer Quick vortex and spin. Wait for 3 minutes

f. Put the tubes on the magnetic plate and transfer 43µl of supernatant in a new tube

15) Ligation of the P2 adapter to fragments *(10 minutes or 45 minutes)*

If you have a P2 solution already prepared, follow the step B. If not, prepare P2 adapter in following the step A).

A) P2 Adapters preparation for 20µl: + 98°C during 2min, from 98°C to 10°C during 20min (decrease of 7% each second)

H2O	6µl
Solution Tris (20mM), NaCl (100mM)	10µl
P2 adapter top (100µM)	2µl
P2 adapter bottom (100µM)	2µl

B) P2 adapter ligation *(10 minutes)*

NE Buffer 2 (10X)	5.0µl
P2 RAD adaptateur (10µM)	1.0µl
rATP (100 mM, Fermentas R0441)	0.5µl
T4 DNA Ligase (NEB M0202M)	0.5µl
Sample (from step 14f)	43.0µl

Incubate at 20°C (Room temperature) for 30min.

Keep the Ampure magnetic beads outside in room temperature at least 30 min before next

step 14.

16) Purification with beads *(40 minutes)*

a. Add 50µl of magnetic beads.

b. Incubate 15 minutes at room temperature

c. Put on the magnetic plate. Throw the supernatant (or store it at -20°C)

d. Wash twice the beads with 500µl of 75% Ethanol: let the ethanol for 30 seconds and remove it

Take care there is no more ethanol is left in the tube when you finish this step

g. Dry the beads for 4- 5 minutes

Avoid over-drying the magnetic beads

h. Elute in 32µl of EB buffer. Quick vortex and spin. Wait for 3 minutes

Use 1µL for nanodrop quantification.

17) Make PCR mix *(1 hour 20 minutes)*

H20	34.0µl
P1 Adapter Primer (10µM)	4.0µl
P2 Adapter Primer (10µM)	4.0µl
2X Phusion Master Mix (NEB F-531L)	50.0µl
Sample (from the step 16h)	8.0µl

Store remaining 24 µl sample from step 16 at -20°C

Cycling conditions: Step 1: 98°C for 30 sec

Step 2: {98°C 10 sec, 65°C for 30 sec, 72°C for 30 sec} 17X

Step 3: 72°C for 5 minutes; Hold at 10°C.

18) Purification with beads *(40 minutes)*

a. Add 100µl of beads

a. Incubate 15 minutes at room temperature

b. Put on the magnetic plate. Throw the supernatant (or store it at -20°C)

c. Wash twice the magnetic particles with 500µl of 75% Ethanol: let the ethanol for 30 seconds and remove it

Take care there is no more ethanol is left in the tube when you finish this step.

d. Dry the magnetic particles for 5 minutes

e. Elute in 21µl of EB buffer. Quick vortex and spin. Wait for 3 minutes

f. Put the tubes on the magnetic plate and transfer 23µl of supernatant in a new tube

Use 1µL for nanodrop quantification

19) Fragments size selection (400-600pb) *(2 hours)*

Careful: Put the magnetic particles at room temperature 30 minutes before using it

a. Add 80µl of EB buffer

b. Add 54µl of beads. Quick vortex

c. Incubate 15 minutes at room temperature

d. Put the tubes on the magnetic plate. Transfer and retain 154µl of the supernatant into a new tube for further processing in 9d. Throw the tubes with magnetic particles.

e. Add 76µl EB buffer and 70µl of beads to the supernatant collected in step 9c. Quick vortex and spin.

f. Incubate 15 minutes at room temperature

g. Put on the magnetic plate. Throw the supernatant in a tube (or store it at -20°C with suitable labeling)

h. Wash twice the beads with 300µl of 75% Ethanol: let the Ethanol during 30 second and remove it.

NB: In this step plate should be on the magnetic plate.

Be sure that there is no more ethanol is left in the tube when you finish this step

i. Dry the beads for 4-5 minutes.

Avoid over-drying the magnetic beads. It will reduce the efficiency of elution significantly

j. Take the tubes out of the magnetic plate. Elute in 22 µl EB buffer. Quick vortex and spin and wait for 3 minutes.

k. Put the tubes on the magnetic plate and transfer 22µl of supernatant in a new tube

Use 1µL for nanodrop quantification.

Use 1µl for DNA Chip in order to check the fragments size.

Use 1µL for nanodrop quantification.

20) Use 1µl for DNA Chip in order to check the fragments size.

21) Use 1µl for Picogreen quantification

22) Dilute your DNA to achieve a final concentration of 10ng/µl for 20µ

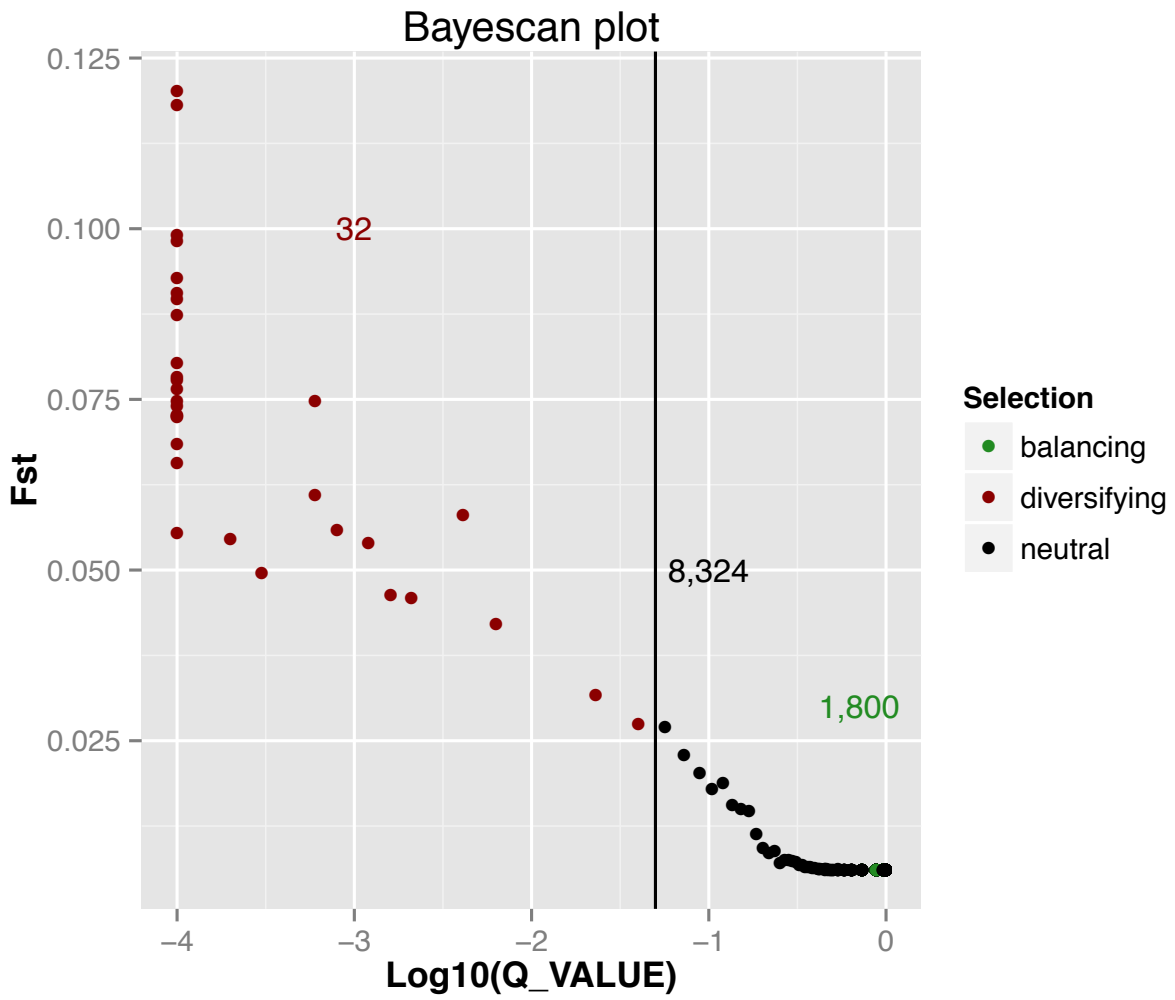


Figure S2.1. Bayescan test for selection.

Bayescan test for selection on individual SNPs among 17 lobster sampling locations, implemented in BAYESCAN program. Red symbols represent SNPs potentially under divergent selection whereas black symbols represent SNPs potentially neutral and green symbols represent SNPs potentially under balancing selection.

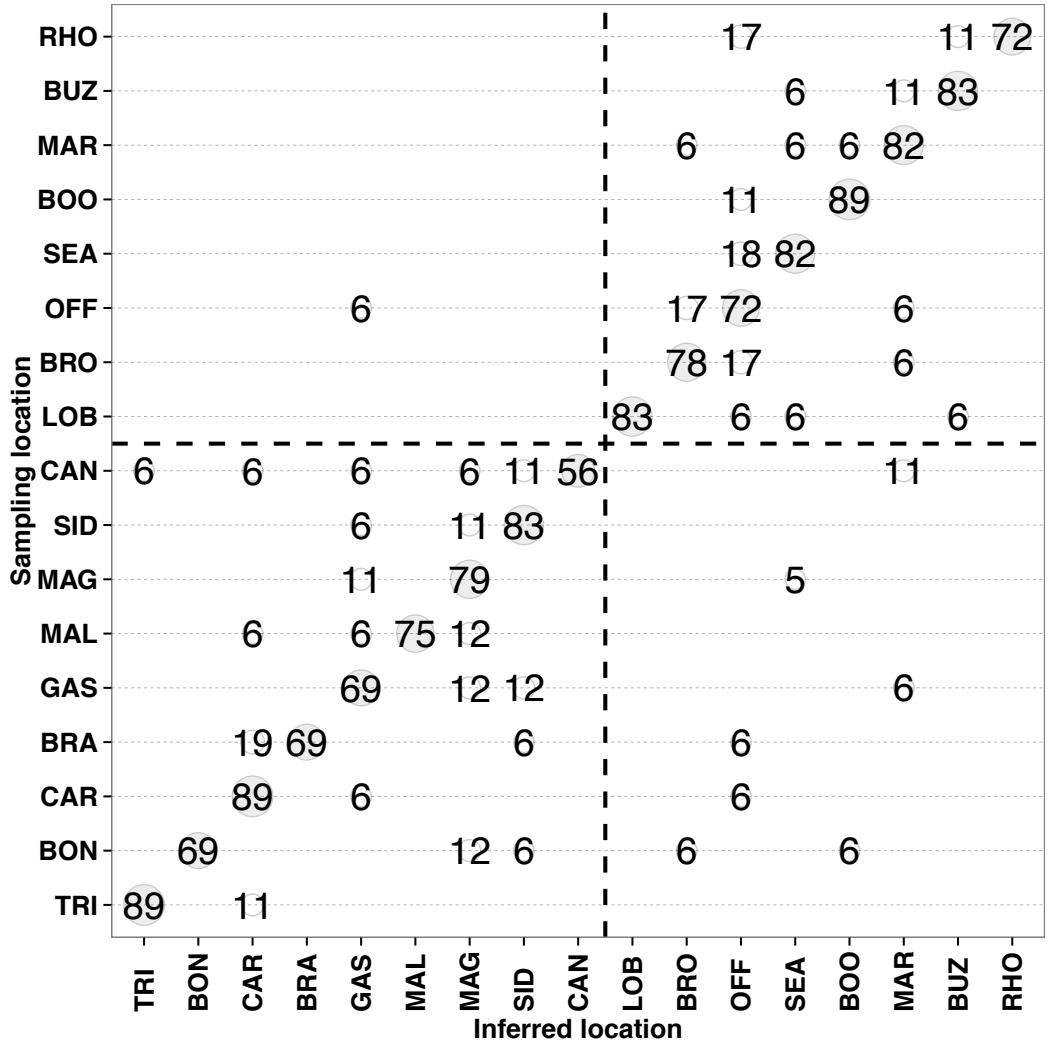


Figure S2.2. Assignment test results.

Blind assignment success expressed as the percentage of lobsters sampled from one sampling location that are classified into their sampling location of origin (grey-shaded numbers on diagonal) or inferred to belong to another sampling location (non-shaded numbers).

2.11. Erratum

In an article we recently published in Molecular Ecology (Benestan *et al.* 2015), we documented fine scale population structure and performed population assignment in the American lobster (*Homarus americanus*). Results first revealed the existence of a weak, albeit highly significant hierarchical genetic structure separating lobsters from the northern and southern part of the studied range (FCT=0.0011, P-value<0.001). At a finer scale within region, we resolved 11 genetically distinct populations differing by F_{ST} values averaging 0.00185 (P-value<0.001). We then performed population assignment, which showed that at the regional scale we could reach an allocation success of 94.2%. At the population level, success was lower but still high (80.8%). We assessed the potential for population assignment using the method of Anderson (2010) designed to avoid a "high grading" bias, which consists of choosing a panel of markers using a *training* data set and then using this panel to perform assignment tests on a completely separate *hold-out* dataset.

Unfortunately, while the Anderson method (2010) was applied properly at the regional level, whereby very high assignment success was achieved without suffering from high grading bias, the method was not applied correctly at the population level. That is, the ranking of markers was made by mistake using all individuals followed by assignment tests on the *hold-out* data set. Upon correcting this mistake, our assignment success at the population level was dramatically reduced to reach on average a maximum of 31,1% over all populations with all the 10156 SNPs (Figure 2.7). Thus, these re-analyses indicate that the high assignment success of 80.8% at the population level reported in our original paper was overly optimistic because it was caused by high-grading bias.

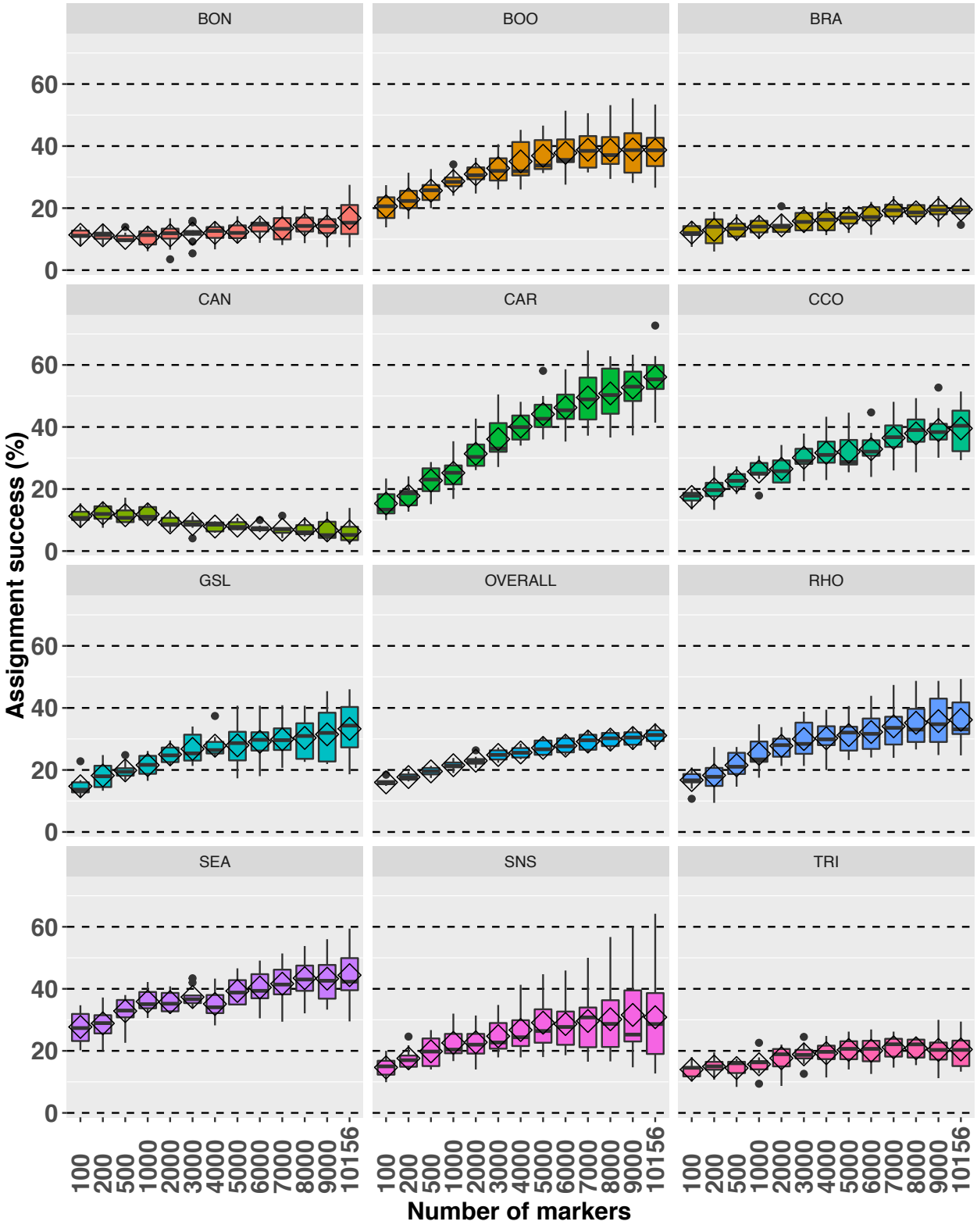


Figure 2.7. Assignment test success in relation to the number of markers. Proportion of assignment success in relation to the number of markers used and ranked following the THL method (50% of the individuals were used to rank the SNPs) considering each of the 11

populations that were defined in the study as well as considering all the populations. The analysis was run ten times for each incremental number of ranked markers based on their F_{ST} values and the results were averaged for each population (BON, BOO, BRA, CAN, CAR, CCO, GSL, RHO, SEA, SNS, TRI) and then over all populations (OVERALL).

Here, we also consider the possibility that the relatively small number of samples for each location (n=36 at maximum) and then dividing the samples into a *training* and a *hold-out* or *test* set (on average n=18 each) may also have contributed to the low assignment by resulting in imprecisely estimated F_{ST} values when ranking the markers and imprecisely estimated allele frequencies for population assignment due to sampling errors associated with small sample size. Namely, we wanted to clarify whether the differences in assignment success obtained previously and when properly applying the Anderson method (2010) was caused either (i) by the sole problem of high-grading bias, and/or (ii) potentially due to a down-grading effect associated with increased sampling error due to low sample size. This was also motivated by the fact that high assignment success was achieved at the regional level for which sample sizes were much bigger than at the population level. We thus applied a Leave-one-out (LOO) procedure described in Anderson (2010), which requires that each individual, in turn, be left out, while the entire process of locus selection and allele frequency estimation is carried out without that individual, and then that individual is assigned back to a population. This procedure improved the assignment success only slightly and still resulted in a much lower assignment than what we reported in Benestan *et al*. (2015), with a maximum of 32,4% on average considering the 11 populations and using all the 10156 SNPs.

Assessing the effect of high-grading bias vs. sample size at the regional level

To further investigate the possible effect of high-grading vs. small sample size and to ensure the reproducibility of our work, we developed and created an easy-to-use workflow, which allows one to carry out *holdout* and *test* set construction and population assignment quickly and efficiently. The workflow is implemented in one function called GBS_Assignment available in the assigner package accessible though Thierry Gosselin's Github page (https://github.com/thierrygosselin). This function works with gsi_sim, a program written in C, which can be used to assess the accuracy expected of genetic stock identification given a genetic baseline (Anderson *et al.* 2008).

Using *assigner*, we first used data from the two regional genetic clusters for which sample size is large (n= 306 and 280 for the North and the South regions, respectively). We then applied the Anderson method properly as was done in Benestan *et al.* (2015) for the regional level and compared the assignment success with that obtained when we voluntarily created a full high-grading effect by using the same individuals for ranking and testing. While the success obtained was indeed higher when creating the high-grading scenario, the difference between both results was more modest than at the population level, the mean success increasing from 95.0% on average to 97.0%. Thus, if not properly eliminated using the Anderson method (2010), the high-grading effect would exist but not be as greatly pronounced at the regional level. However, we note that neither is the high-grading effect at the regional level completely insignificant as it does represent a 40% reduction of the expected rate of misassignment.

Since regional assignment success was still high after correcting for high-grading bias and samples size were high for both regions, we used this system as a positive control to test the impact of the sample size on assignment success. In this case, we expected that regional assignment success, corrected for high-grading bias and performed on the same number of individuals that we sampled per location (which was about 30 individuals), would decline as the sample size decreased. To test this hypothesis and delineate the gradual effect of the number of samples on assignment success, we selected randomly from 20 to 100 samples from each region, ranked SNPs based on half of these samples (*training* set), and then calculated the assignment success on the *hold-out* set (THL method: 50% of the individuals were used to rank the SNPs). Overall, when using fewer than 50 individuals, assignment success obtained with the 10156 markers was reduced to about 60%, close to the 50% success expected randomly between the two regions (Figure 2.8). The assignment success became higher than 80% only when >100 individuals were used for each region. Thus, in our system at least, sample size has a strong effect on the assignment success being reached. These results at the regional level thus suggest that increased assignment success at the population level would require considerably larger sample sizes. Given that sample size used in standard GBS or RADseq studies are often comparable to those used in this study (say 30-40 per sampling location), our results also show that it may be difficult in many such studies to accurately perform assignment tests in weakly differentiated species.

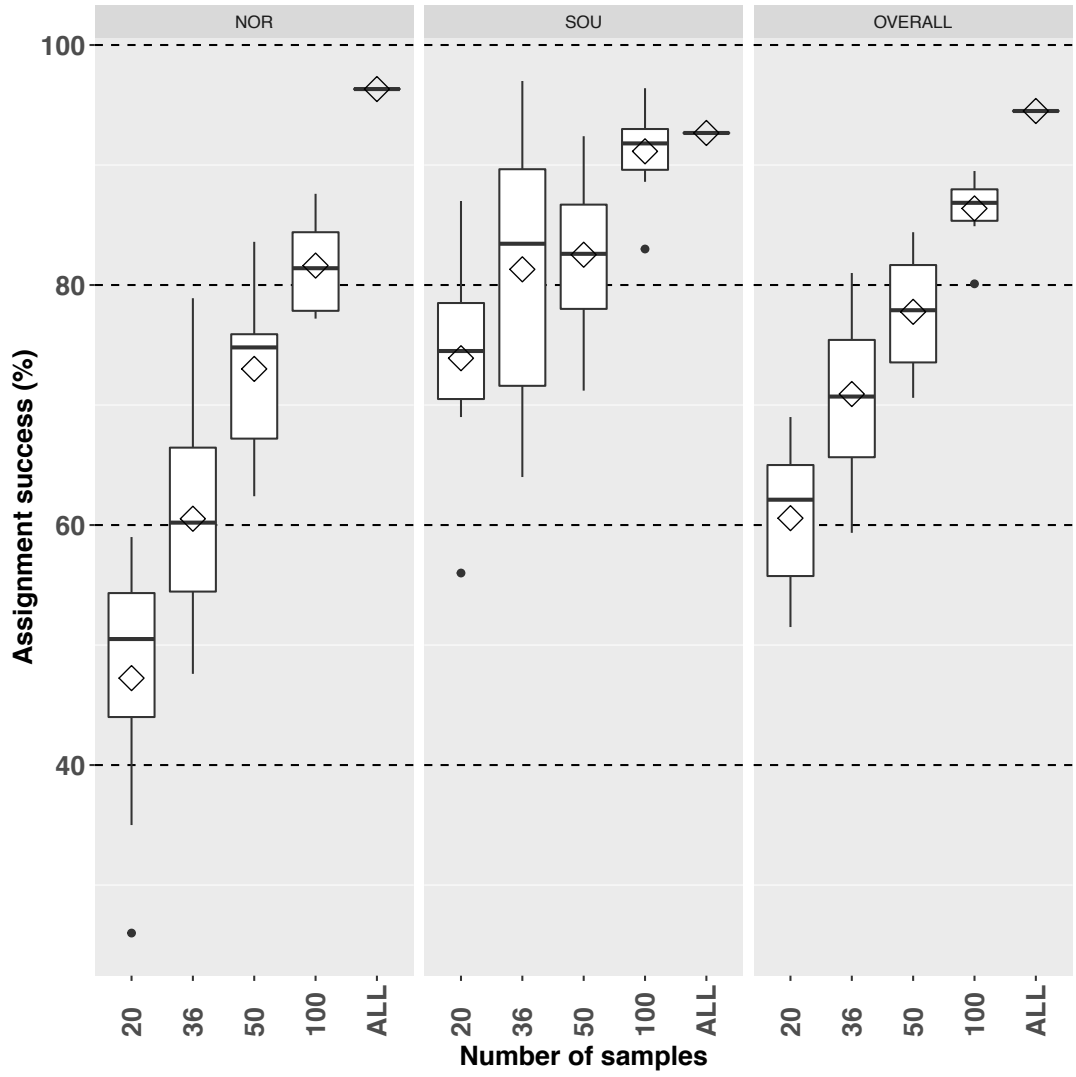


Figure 2.8. Assignment success in relation to the number of samples. Assignment success (using the 10156 SNPs) in relation to the number of individuals sampled considering the North (panel on the left) and the South (panel in the center) as reference populations and following the THL method. Both results were summarised in the OVERALL panel (on the right). The analysis was run ten times with different subsets of samples in order to obtain confidence interval. In the panels, the vertical limits of the box represent one standard deviation around the mean (white diamond), the horizontal line within the box is the median, and the whiskers extend from the box to the 25th and 75th percentiles.

Using top-ranked SNPs vs all SNPs

By applying Anderson method (2010), we observed that population assignment

success is higher when using all the markers available than with a subset of the most discriminant markers (Figure 2.7). This pattern is very different from the one we presented in Benestan *et al.* (2015), where assignment success increased up to the 3000 top-ranked markers used and then decreased substantially beyond that number. Therefore, this difference was clearly an artefact of the high-grading bias generated by not applying Anderson method correctly in our paper. From this, we recommend that future studies use all the markers available since assignment success is not generally expected to be better with fewer markers. Additionally, assessing the power of assignment using all the available markers circumvents entirely the problem of high-grading bias and does not require breaking the sample into a separate *holdout* set for marker panel selection and a *test* set.

Conclusion

These additional analyses demonstrate that high assignment success can be achieved in situations of weak absolute genetic differentiation as long as the sample size is relatively high, as was the case in our analysis at the regional level. Nevertheless, we do not have sufficient data to conclude at this time that we will be able to assign lobsters to their sampling locales with greater than about 25%-30% accuracy on average. Our results at the population level also indicate that a strong high grading bias will result from using a subset of pre-selected markers if the Anderson method (2010) is not applied. In fact, the results of this *erratum* also illustrate that selecting a subset of SNPs does not lead to better assignment results than using all the markers for performing assignment tests. Furthermore, both properly avoiding high-grading bias and achieving accurate population assignment with weak population differentiation ($F_{ST} < 0.01$) requires large samples. We recommend that any further studies aiming at quantifying assignment success and/or correcting for high-grading bias in weakly differentiated populations should plan to analyse large sample sizes (at least n =50, and ideally >= 100) per location. Additionally, assignment success should be initially assessed using all the available markers. Subsequently, if a selected panel of markers appears to offer much more accurate assignment than all the markers, high-grading bias should be suspected, and an exhaustive search for an error in the methodology should be conducted.

Finally, we are grateful to Eric Anderson and Kelly Barr, for detecting and pointing out to us the erroneous assignment analyses performed at the population level in Benestan *et*

al. (2015). Anderson and Barr's analyses are available on GitHub at https://github.com/eriqande/lobster_checkin.

Chapitre 3. Conservation genomics of natural and managed populations: building a conceptual and practical framework.

Publié sous **:** Benestan LM, Ferchaud AL, Hohenlohe PA, Garner BA, Naylor GJ, Baums IB, Schwartz MK, Kelley JL and Luikart G. (2016). *Molecular ecology,* **25**, 2967-2977.

Conservation and evolutionary genetics are rapidly shifting from a genetic to a genomic perspective, where studies assess thousands of in hundreds of individuals (Allendorf *et al.* 2010; Ouborg *et al.* 2010; Narum *et al.* 2013). The field has benefited from previous developments in population genomic studies of model organisms, especially in humans (see examples reviewed in Allendorf *et al*. 2010). A practical and conceptual framework for effective study design and analytical approaches is needed to help guide the new generation of population geneticists in using large-scale genomic dataset. Indeed, integrating knowledge about many of the new molecular and computational tools available for analyzing genomic datasets is crucial to answering questions in evolutionary and conservation biology. With knowledge of the tools available, researchers should use the underlying scientific question to guide all aspects of a conservation or evolutionary genomic study, from experimental design through data analysis (see Figure 3.1).

To help educate population genomics researchers, 15 experts in the field of conservation genomics directed a one-week workshop called "ConGen 2015" (abbreviated from Conservation Genetics) at the University of Montana Flathead Lake Biological Station. This meeting review was written for everyone interested in population genomics, from graduate students to professors and resource managers. Here, we highlight the key topics and important take home messages discussed during the workshop, with an emphasis on the recent pertinent literature in this field. More particularly, we described (1) how to design a massively parallel sequencing (MPS; also called next generation sequencing) study for a model or non-model species, (2) how to filter DNA sequence data from MPS data (i.e., extracting loci and/or SNPs on the basis of criteria), and (3) to analyze MPS data using classic (e.g., clustering algorithms) and recent approaches (e.g., likelihood algorithms), within traditional or new pipelines (e.g. Galaxy). This overview will allow researchers to better understand some of the strengths and limits of recent molecular and computational approaches.

3.1. **Designing a MPS study: Keeping in mind your biological question**

One of the biggest differences in using MPS data versus classical genetic data (e.g., microsatellites) is the amount of time spent on data analysis, with data production greatly outpacing our ability to analyze it. As stated by ConGen instructor Paul Hohenlohe, it is not

just about generating data. Conservation genomics offers an unprecedented genomic perspective by using large numbers of markers to simultaneously genotype putatively neutral and adaptive loci, thus offering glimpses into adaptive potential (Allendorf *et al.* 2010). Then, designing a MPS study requires the consideration of a large number of factors represented by Figure 3.1 and recently reviewed by Andrews *et al.* (2016).

The most important starting point remains, "**What is your scientific/biological question?**" This should determine how a researcher navigates all subsequent questions such as "What is your sampling design and how should you allocate your budget among samples, populations, individuals, loci and depth of sequence coverage?" Question-driven rather than method-driven research allows researchers to not be limited by the methodological tools available, thus offering the flexibility and openness required to find the appropriate method that answer their question. For instance, recent simulation studies showed that a sampling design with geographically close populations (with recent gene flow and thus low genome-wide F_{ST}) across a selection gradient (environmentally distinct locations) had more power to detect local adaptation (Lotterhos & Whitlock 2015). When do you need to sequence the entire genome versus only genotype hundreds or thousands of loci to answer your question? For equivalent budgets, a large number of individuals can be genotyped at lower coverage, if you are interested in accurate estimates of population parameters (e.g., gene flow, F_{ST} outlier loci), whereas few samples could be genotyped at a higher coverage when you need to genotype individuals accurately (e.g., to assess individual inbreeding level). Do you have an a priori hypothesis about the features of your biological model (e.g., the colonization history, the generation time, small and isolated versus large and panmictic populations, dispersal capabilities, heterogeneous versus homogeneous habitats) that could help you to predict the level of genetic diversity, the effect of genetic drift and the extent of the selective pressures?

3.2. **Existing methods for MPS data analysis**

3.2.1. Low-coverage genotyping methods and genotype likelihoods (Mike Miller)

Novel Bayesian methods that aim to analyze efficiently low-coverage genomic data are blooming (Le & Durbin 2011; Yu & Sun 2013; Cantarel *et al.* 2014). Understanding theory behind the application of Bayesian models to low-coverage genomics data is crucial and begins with learning how to calculate genotype likelihoods from DNA sequences. Thus, Mike Miller instructed students how to calculate genotype likelihood based on sequencing

errors, coverage and priors probabilities (*i.e.,* uniform or Hardy-Weinberg Equilibrium model). He showed how these key factors could significantly affect genotype likelihood results and then many downstream analyses (Sims *et al.* 2014). These analyses may then suffer from SNP calling and genotype uncertainty, which lead to inaccurate demographic inferences (Nielsen *et al.* 2011). One way to overcome this bias could be to sample larger numbers of individuals at the expense of coverage depth in order to gather more information about population parameters, as suggested by Buerkle & Gompert (2013).

The importance of removing PCR duplicates (reads resulting from PCR clonal amplification of the same original DNA strand) was also underscored, because of their potentially distorting influence on the calculation of genotype likelihoods (overconfidence in a genotype called only based on PCR duplicates) as suggested by Puritz *et al.* (2014b). PCR duplicates can easily be removed from paired-end restriction site associated DNA (RAD) sequencing datasets by identifying paired-end reads starting at identical position (Davey *et al.* 2013) and from genotyping-by-sequencing (GBS) datasets by using degenerate-base adaptors (Tin *et al.* 2015). Similarly, paralogs should be excluded from the analysis by detecting reads with high coverage, although genomic datasets often have high variance in coverage across loci (see Box 1; Malhis & Jones 2010). Finally, M. Miller also presented new computational approaches to detect genotyping errors, along with a new genotyping approach that combines RAD-seq with DNA capture arrays for low cost genotyping (Norgaard *et al.* in press; Ali *et al*. 2016). Calling genotypes based on their likelihoods can be easily performed with ANGSD (Korneliussen *et al.* 2014) and GATK (DePristo *et al.* 2011) programs.

3.2.2. Mapping reads to a reference genome (Paul Hohenlohe)

Aligning anonymous sequence reads against a reference genome assembly provides many advantages for filtering data (e.g., removing erroneous or clonal PCR duplicate reads) and identifying loci (Hand *et al.* 2015). If a reference genome is unavailable for the focal species, P. Hohenlohe advised using well-assembled genomes from related taxa. Efforts such as the Genome 10K project (https://genome10k.soe.ucsc.edu/) and the i5k Insect Genome project (https://arthropodgenomes.org/wiki/i5K) are rapidly growing the number of taxa for which this is possible. The issue of how closely related is "closely related enough" to be useful for alignment depends on details of the dataset, such as the sequence read length and whether more conserved regions such as genes are targeted for sequencing. A poorly

assembled reference genome can still be useful for assigning reads to loci and finding functional genes linked to candidate markers, even if it does not provide a complete physical map of the genome (e.g., Hand *et al.* 2015).

Techniques like paired-end RAD sequencing (or exon capture) can also be used to build a set of contig sequences for non-model species, which then provide a reference for further population-level sequence data (Hohenlohe *et al.* 2013; Jones & Good 2016). When faced with limited resources, P. Hohenlohe cautioned against pool-sequencing (i.e., pooled sequencing of many individuals without barcode), because of the pitfalls associated with estimating allele frequencies (missing rare variants), identifying paralogs, distinguishing true alleles from sequencing error, and hidden population structure. Whereas pooling showed promising results for accurate allele frequencies estimates (Futschik & Schlötterer 2010; Ferretti *et al.* 2013; Lynch *et al.* 2014), this approach is often less desirable than individual sequencing for a wide range of applications such as Structure analysis, parental assignment and genome scans (review in Cutler & Jensen 2010).

3.2.3. Stacks Workflow tutorial, stackr package and Galaxy (Laura Benestan and Tiago Antao)

There is a need for standardization and documentation of the many filtering and processing steps (Box 1) required to clean and use MPS data (e.g., by multiple researchers within a research group or the larger scientific community). Laura Benestan also emphasized that standardization helps ensure repeatability. The Stacks workflow tutorial created by Éric Normandeau for Louis Bernatchez's research group at Laval University was designed to facilitate, standardize, and document (in a log file) each of many filtering and analysis steps in discovery and genotyping of putative SNP markers from GBS/RAD sequencing data using the Stacks program (Catchen *et al.* 2013). Stacks is a widely used pipeline for analysis RAD-seq data but other pipelines such as pyRAD (Eaton 2014), RADtools (Baxter *et al.* 2011), GATK (McKenna *et al.* 2010), dDocent (Puritz *et al.* 2014) could also be used for calling SNPs. More particularly, pyRAD, dDocent and more recently Stacks are promising workflow programs that can handle insertion-deletion polymorphism into the alignment of the reads.

The Stacks workflow uses universal tools, including custom scripts, to standardize and make repeatable all aspects of the pipeline, while also highlighting areas where the researcher should exercise caution in the choice of parameter values. The workflow is freely

available on GitHub (https://github.com/enormandeau/stacks_workflow). The included manual describes each step required for performing MPS analyses in Stacks from downloading and installing Stacks to filtering the results. Raw single-end data produced by Illumina or Ion Proton technology are supported.

Post-Stacks analyses and data filtering (Box 1) can be conducted with the R package stackr (Gosselin & Bernatchez 2016). This package is freely available on Github (https://github.com/thierrygosselin/stackr). Stackr contains several R functions that allow users to: (1) read and modify outputs from Stacks, (2) filter markers based on coverage, genotype likelihood, number of individuals, number of populations, minor allele frequency, observed heterozygosity, and inbreeding coefficient (F_{IS}), (3) explore distributions of summary statistics and create publication-ready ggplot2 figures, (4) impute missing data using a Random Forest algorithm and (5) export datasets in vcf, genepop, fstat files or as genind objects to be easily integrated into other R packages for population genomics analyses.

Tiago Antao also demonstrated web-based Galaxy software platform (https://galaxyproject.org), which could help with standardization of filtering and genotyping. Galaxy produces flow-chart diagrams (of filtering steps) and log files to help researchers reproducing and sharing complete "pipeline" analysis with others. Galaxy is an interesting tool for data visualization as it could efficiently draw graphics (i.e., graphics of the distributions of quality scores) that allow users to explore and navigate their data. Running Stacks and related filtering approaches could be also done easily from this web-platform.

3.2.4. The "F-word": Filtering (Jim Seeb)

Genomics involves the genotyping of thousands of loci, genome-wide, to bring unprecedented resolution to problems of conservation planning (Allendorf *et al.* 2010; Shafer *et al.* 2015). However, this bright future for genomics hides the numerous filtering issues inherent to MPS datasets, which Jim Seeb referred to as the "F word". For instance, merging datasets filtered using different parameters could create spurious results such as strong and significant (but false) F_{ST}-outlier values between differently filtered population samples. The lack of necessary details on the filtering steps in many of today's publications using MPS data would affect the transparency and reproducibility of the results. This would contribute to the trend that most of the MPS studies cannot be accurately verified (Nekrutenko & Taylor

2012). To encourage scientist to publish and understand these important filtering steps, Box 1 reports some of the main filtering issues (associated with sequencing and assembling errors) that should be addressed in a MPS project. A complete and exhaustive publication will bring detailed recommendations and pipeline to conduct accurate filtering steps on MPS data in a forthcoming Population Genomics in R – Molecular Ecology Resources special issue. Identifying markers of interest through filtering steps could be done according to single SNP or haplotype approach, the latter being a possible alternative to overcome issues regarding linkage disequilibrium (LD; Box 1). Nevertheless, it is important to keep in mind that the appropriate level of filtering will always depend on the scientific question and the available dataset (Andrews *et al.* 2016).

In addition, RAD locus discovery and genotyping is often inhibited by the existence of duplicated genes and genomic regions (Allendorf *et al.* 2015; Andrews *et al.* 2016). Gene duplication occurs because of segmental duplication (unequal crossing over) or whole genome duplication (Amores *et al.* 2011). Loci can be assayed in duplicated regions by constructing linkage maps and genotyping with SNP chips or potentially with very deep GBS coverage (Waples *et al.* 2015). Distinguishing and including paralogous loci on the linkage map will allow researchers to circumvent the issues of producing an incomplete picture of the genome (Brieuc & Waters 2014; Kodama *et al.* 2014) and introducing bias into genetic estimates parameters (Meirmans & Van Tienderen 2013).

3.2.5. Structure program insights and tips (Jonathan Pritchard)

Jonathan Pritchard provided an overview and practical advice about the application of the program Structure (Pritchard *et al.* 2000). J. Pritchard explained that it is often unrealistic to expect that there is one "true" K that is best for modeling a particular data set. Through simulations, Kalinowski (2010) showed that sometimes Structure clustered individuals in unpredictable ways, which is because the Structure model is a cartoon (simplification) of more complicated natural population. Therefore, viewing and reporting plots for multiple K values is an important step (Gilbert *et al.* 2012) because different values of K can give insights into different levels of structure. Similarly, the selection of the optimal K is not an exhaustive procedure and has to be done with regard to the biology and the history of populations studied (Kalinowski 2010). For instance, the optimal number of clusters (K) found by Structure or subsequent analysis (e.g., Evanno *et al.* 2005) may have no biological

reality and could result from a context of isolation by distance, where Structure tends to overestimate genetic structure (Frantz *et al.* 2009). In the same vein, a recent simulation study showed that unbalanced sample size lead to wrong demographic inferences where smaller samples tend to be merged together (Puechmaille 2016). To overcome these issue, alternative methods such as principal component analysis or evolutionary trees could be tested in regards to Structure analysis (Jombart *et al.* 2010; Kalinowski 2010; Kanno *et al.* 2011; Benestan *et al.* 2015).

Reviewers often request extremely long Structure runs (millions of iterations). J. Pritchard claimed that is generally unnecessary and wasteful of researcher time (and carbon footprint). For most data sets he would recommend to do about 10 000 steps, but multiple times to assess robustness and convergence of the results. Structure tends to converge fairly quickly, but the program does not do a great job of exploring between local peaks in parameter space of the posterior distribution. Therefore, for an exploratory analysis, it would be more efficient to spend the computation time on independent runs (which have a good chance of finding distinct modes) than doing extremely long runs where the algorithm will be simply wandering around within one mode. Nevertheless, a certain minimum burn-in and run length helps overcome the stochasticity of the Monte Carlo approach, as recommended by Gilbert et al. (2012).

3.2.6. Improving our detection of local adaptation (Lisa Seeb)

Understanding genetic basis of local adaptation is one of the most exciting potential contributions of genomics to conservation biology (Allendorf *et al.* 2010). The most widely used methods for detecting evidence of selection are genome scan approaches based on differentiation (F_{ST}) outlier tests originally developed by Lewontin and Krakauer (1973) then refined more recently by Beaumont and Nichols (1996) and others (Beaumont & Balding (2004), Foll & Gaggiotti (2008)). Several programs were designed to perform genome wide outlier scan analysis such as LOSITAN, which is based on a stochastic F_{ST} null distribution (Antao *et al.* 2008), Arlequin (Excoffier & Lischer 2010) which includes a hierarchical model, and BayeScan (Foll & Gaggiotti 2008) which is based on a Bayesian F_{ST} distribution. Since each method has its drawbacks, requiring outliers (candidate adaptive genes) to be identified in multiple methods can help to reduce the incidence of false positives (Villemereuil *et al.* 2014; Francois *et al.* 2016) because these different methods may tend to

agree more on true positives than on false positives. Lisa Seeb described the work of Mita *et al.* (2013) who investigated the robustness of eight methods to detect loci potentially under selection according to eight demographic scenarios along an environmental gradient. Their work showed that whereas genotype-environment correlation methods have more power to detect signal of selection than genome scans, these methods were more prone to false positives when assessing these associations.

The importance of incorporating neutral genetic structure into genotype-environment correlation methods has led to the emergence of two recent software packages: BayeScEnv (Villemereuil *et al.* 2014) and LFMM (Frichot *et al.* 2013). However, as well as using suitable methods, L. Seeb emphasized that an appropriate sampling design is crucial to test for evidence of local adaptation. For instance, analyzing sets of independent populations ("replicates") across similar environmental gradients helped Larson et al. (2014) to find signals of selection in Chinook salmon. In addition, mapping outliers to find chromosomal islands of divergence can help to identify functional genes involved in local adaptation. We advise scientists interested in the utilization of environmental association analysis in genomics to read Hand et al. (2015b), Rellstad et al. (2015) or van Heerwaarden *et al.* (2015). Researchers should be aware that new and improved tests as well as evaluations of tests are published frequently (e.g., see (Foll *et al.* 2014; Whitlock & Lotterhos 2015).

3.3. **The use of genomics for management decisions**

3.3.1. Effective population size (N_e) estimation (Robin Waples)

Robin Waples taught concepts of the effective population size by using an analogy of a lottery. Imagine the ability of parents to produce viable offspring for the next generation depends on a lottery system. In a Wright-Fisher (ideal) population, everyone has the same number of tickets, and sampling is with replacement. In real populations, different individuals have different numbers of lottery tickets, because some of them will reproduce more than others, and hence they have different probabilities of being parents, thus reducing N_e compared to census size. He enumerated the different methods that can be used to estimate contemporary N_e: temporal methods, LD methods, approximate Bayesian computation (ABC) methods, and other single estimators based on heterozygote excess (Pudovkin *et al.* 1996), molecular coancestry (Nomura 2008), and sibship analysis (Wang 2009). ConGen

participants were also reminded that these methods make several important assumptions: no migration, no selection, mutation is unimportant, discrete generations, random sampling of an entire generation, and loci not physically linked.

R. Waples mentioned that genetic estimates of either contemporary or long-term N_e benefit from the proliferation of the number and types of markers, but this also introduces challenges, largely because of a) LD, which is unavoidable when large numbers of markers have to be packaged into a small number of chromosomes, and b) pseudo-replication, because of linkage, markers are not independent, so adding more and more loci does not increase precision as fast as it would under complete independence. LD is predicted to be the next big issue in dealing with genomics data since multilocus sampling improves whereas classic analyses such as N_e estimation, genome scan and clustering algorithms treated the loci as independent (Baird 2015). Kemppainen et al. (Kemppainen *et al.* 2015) present a useful exploratory tool (named LDna) able to give a global overview of LD associated with diverse evolutionary phenomena and identify potentially related loci. Based on simulations, Waples et al. (in review) showed that more loci do not increase the fraction that is physically linked, since most random pairs of loci are not linked. If linked loci downwardly bias N_e estimates (Larson *et al.* 2014), the bias from ignoring linkage is less severe when the number of chromosomes is large. Finally, strategies that filter out a locus in outlier pairs of loci are only partially effective and a bias correction factor based on the number of chromosomes is likely more effective (Waples et al. in review). Videos recording R. Waples' N_e lecture can be viewed at https://www.youtube.com/watch?v=ErhACWXRLss and https://www.youtube.com/watch?v=N3JbKZbKO5w

3.3.2. Defining conservation units: ESUs and MUs (Robin Waples)

Integrating genomic data into management can be challenging in practice. For instance, there is no single best or correct way to answer the questions "what is a population" and "how to identify the suitable conservation unit" (e.g., ESU, MU, etc.) because the definitions of these terms can be vague, not quantitative, and depend on the management objective (Waples & Gagiotti 2006). Since several "population" concepts can be found in literature (Fraser & Bernatchez 2001), R. Waples suggested choosing the population concept (ESU, MU, etc.) that is appropriate to the objective(s) of each study. One way to detect the number of populations is to test for a statistically significant genetic differentiation. Statistical

power is influenced by (1) population differences (effect size) and (2) data richness (numbers of individuals, number of samples, number of loci, and alleles). Then, important biological differences might be missed if data are limited (low power). On the other hand, statistical significance does not guarantee biological significance, especially when large amounts of data are available (i.e., high power detects even trivial differences, see Palsboll *et al.* (2007). This failing should be a major concern in the age of genomics. Also, standard statistical tests usually do not properly answer the question "Is it different enough?" because they reject only the null hypothesis of no differentiation (panmixia).

Another way to detect population structure and identify population units is to use Bayesian clustering methods such as Structure (Pritchard *et al.* 2000), BAPS (Corander *et al.* 2004) and ADMIXTURE (Alexander *et al.* 2009), but these methods may have reduced power with high gene flow species (Jombart *et al.* 2010; Kanno *et al.* 2011; Benestan *et al.* 2015). Nevertheless, absence of genetic differentiation at neutral markers does not mean absence of adaptive differences (Allendorf *et al.* 2010). Therefore, using markers influenced by selection could be a promising research avenue for delineating important conservation units (see study conducted on herring by Limborg *et al.* (2012) for an example), particularly in high gene flow species (Gagnaire *et al.* 2015). However, a pattern of adaptive divergence may not necessarily match the neutral pattern (e.g., when one adaptive group overlaps two neutral ones) as the processes affecting adaptive and neutral genetic markers are different. Then, combining neutral and adaptive markers in a hierarchical approach to define conservation units, as suggested by Funk et al. (2012) may encounter practical issues in delineating conservation units. Yet, few studies already used information on adaptive differentiation to improve conservation decisions (Limborg *et al.* 2012; Bourret *et al.* 2013; Larson *et al.* 2014). Nevertheless, given the considerable proportion of false positive in outliers detection (Mita *et al.* 2013; Francois *et al.* 2016), it is crucial to complement the pattern of adaptive divergence arising from genomics data with ecological, phenotypic and environmental data. Further research is needed to assess this issue in the future.

3.3.3. Adaptive genomics as a first step (Michael Schwartz)

Michael Schwartz, Director of the National Genomics Laboratory for Wildlife and Fish Conservation (in Montana), led a discussion that focused on the extent of direct use of genomic data in conservation and natural resource management (Shafer *et al.* 2015; Garner *et*

al. 2016). One side of the debate suggests that genomics has advanced fish and wildlife conservation by increasing the number of markers assayed, but has failed to live up to its promise to elucidate the genetic basis for adaptation in a way that can be used by managers (Shafer *et al.* 2015). The other side notes that genomics is currently being used by management agencies in a variety of taxa, but that the non-academic nature of some labs applying genomics to conservation can lead to a lag in publishing in academic journals. Participants and instructors suggested reasons for a potential gap between genomics and direct management application, most noticeably, that of cost and a lack of familiarity (e.g., some managers are more comfortable with the vocabulary or concepts surrounding microsatellite data (and data analysis) than with novel genomic techniques in decision-making).

The group then discussed how to avoid false positives when identifying outliers by applying statistical correction for multiple testing such as Bonferonni or false discovery rate (FDR) correction (Narum 2006). Power to detect true outliers seems to be highly dependent on sampling and statistic test used, whether it controls or not for population structure (Lotterhos & Whitlock 2015). There was an overall recognition by those using genomic approaches that careful identification of outlier loci was a first step. Then, additional empirical evidence showing the functional importance of the outlier in a relevant ecological context is a mandatory step to confirming that these genes are target of selection. For that purpose, common garden and transplant experiments, thought difficult to perform in most of the non-model species, would be required (Barrett & Hoekstra 2011). When such experiments are not possible, the observation of the same outlier-loci in multiple independent population sets can help confirm local adaptation signatures (Bradbury *et al.* 2010; Laporte *et al.* 2016).

3.3.4. RNA-sequencing for management decisions (Joanna Kelley)

Studying gene expression differences among individuals and populations can provide insight into (i) the molecular basis of phenotypic differentiation, (ii) variation in response to environmental conditions, disease, etc., and (iii) management decisions regarding how and where to manage or transplant populations. For example, Barshis et al. (2013) compared transcriptome-wide gene expression (via RNA-sequencing (RNA-seq) using Illumina sequencing technology) among conspecific thermally resilient corals to identify the

molecular pathways contributing to coral resilience. RNA-seq can be also used directly in management decisions. Narum & Campbell (2015) detected differential transcriptomic response to heat stress among ecologically divergent populations of redband trout, which will likely influence future conservation including avoiding translocations between the divergent populations.

The approaches to measuring gene expression including limited gene studies (qPCR and Northern blots) and transcriptome level studies (microarrays and RNA-seq, see (Zhao *et al.* 2014). There are two RNA enrichment techniques, polyA+ selection and ribosomal depletion, to remove the highly abundant ribosomal RNAs from the pool of total RNA, prior to library preparation (Cui *et al.* 2010). Both methods are efficient and their use depends largely on financial resources and whether researchers are interested in coding transcripts or transcripts that may be regulatory (for example, long non-coding RNAs). Directional RNA-seq libraries are recommended to find sense and anti-sense transcripts, which may be relevant for regulatory processes. Additionally, reference bias was briefly discussed. In that context, combining all datasets and generating de novo transcriptome assemblies carefully would be very useful in any comparative analysis. She discussed the pipeline and analyses described in Kelley *et al.* (2012) . Finally, Joanna Kelley referred to the Simple Fool's Guide from Stephen Palumbi's lab (Wit *et al.* 2012) for calling single nucleotide polymorphisms (SNPs) based on RNA-seq data.

3.3.5. General advice from instructors

The common advice given by each instructor was to keep the scientific question of the study in mind at each step from the initial study design to publication. There is no single pipeline for analyzing all (or even any two) MPS datasets, and thus the analysis of MPS data requires an investment in scripting and writing computer code (http://korflab.ucdavis.edu/Unix_and_Perl/; Antao 2015). In addition, students and professionals alike can gain a competitive edge in an increasingly competitive job market by understanding new computational methods and being comfortable operating in some kind of programming language. These skills are particularly desirable now as the sheer size of genomic data sets alone demands computational and scripting or coding prowess.

Robin Waples mentioned the importance of understanding all steps in the process from data production to genotype analysis (by filtering data) to avoid conducting analyses

that are not adequate and could lead to data misinterpretation. Instructor Tiago Antao disagreed somewhat by suggesting that one single person cannot expertly understand every single step of a genomics project; however, instructor and ConGen coordinator Gordon Luikart addressed these concerns by recommending close collaboration with people who are experts in some of the different steps of the process.

As a career advice, Jonathan Pritchard recommended early-career researchers to submit manuscripts online at the ArXiv or bioRxiv web page (e.g., Ali *et al.* 2015) so they can show them on their CV when applying for jobs and to perhaps get early feedback (edits) from the scientific community. Submission to bioRxiv could also advance the field of conservation genomics and ecology faster than by waiting until the paper is actually accepted by a traditional journal. Many journals no longer have an embargo and allow early online publication.

In summary, the growing potential for current application of genetic and genomics approaches to conservation is exciting. However, it also requires increasing the development of next generation approaches and great caution when using massive parallel sequencing. Along with this meeting review, Figure 3.1 and Figure 3.2 provide a conservation genomics framework and highlights important issues arising from the massive scale datasets.

3.4. **Acknowledgements**

LB and ALF received travel grants from RAQ (Réseau Aquaculture du Quebec) to attend the workshop and helpful supported by Louis Bernatchez and the Canadian Research Chair in Genomics and Conservation of Aquatic Resources. We are grateful to Robin Waples and Jonathan Pritchard for helpful comments regarding the manuscript as well as four Anonymous reviewers for their useful comments. We thank the ConGen instructors: Mike Miller, Fred Allendorf, Jim and Lisa Seeb, Tiago Antao, Jeff Good, Brian Hand, Tabitha Graves, Ryan Kovach, Brice Sarver. GL and ConGen were supported in part by grants from the US National Science Foundation (DEB-1258203), NASA (NNX14AB84G), and time and advice from Michelle Quinn at the University of Montana Montana's School of Extended & Lifelong Learning.

3.5. Figures

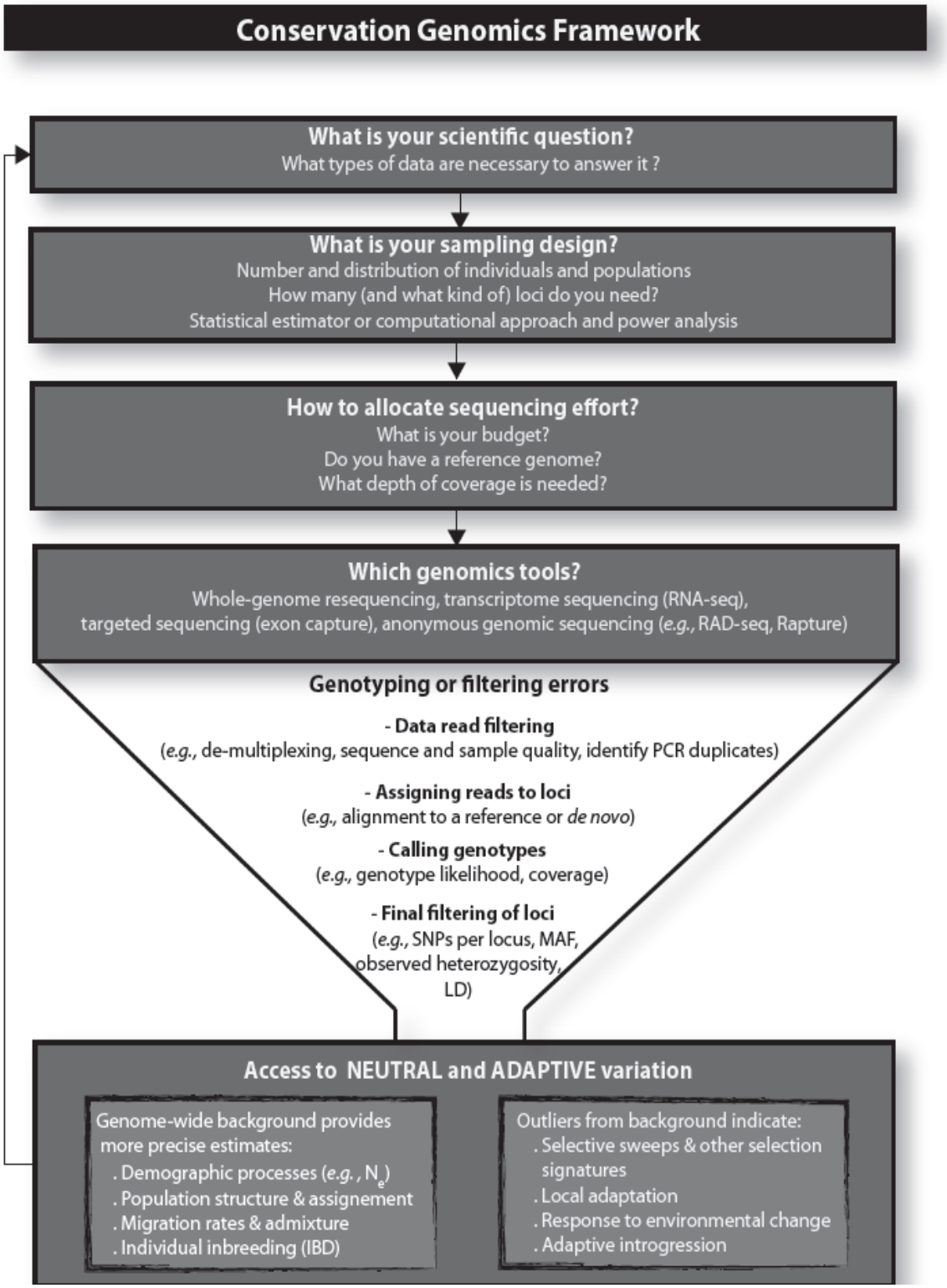
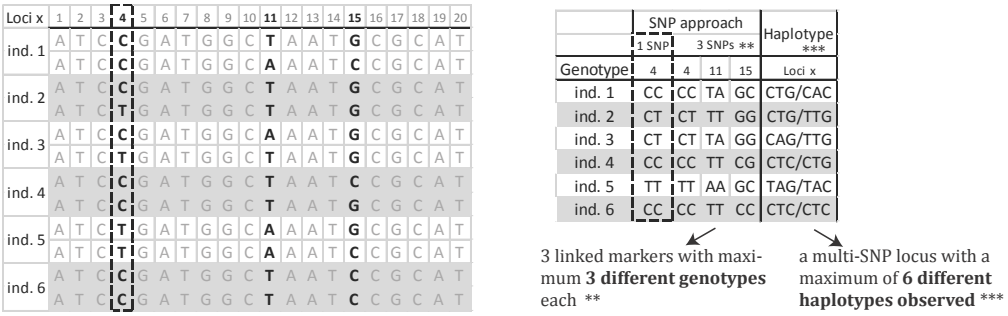


Figure 3.1. Practical framework with steps for designing a MPS study.

Practical framework with steps for designing a MPS study. All along the process researchers involved in MPS projects are faced with logistical trade-offs in order to accurately and efficiently answer their scientific question. The process is not straightforward and unidirectional but feedbacks and/or interactions are possible and common among all steps. "What kind" of loci refers to characteristics such as loci in genes, linked loci or haplotypes (for genealogical information), mapped loci often required for QTL studies or runs of homozygosity, or long loci (e.g., long RAD contigs from paired end reads). The "distribution of populations" refers to the need to sample populations from different landscape locations or across environmental gradients when conducting landscape genetic or genomic studies. "SNPs per locus" refers to the fact that researchers might use only one SNP per RAD locus (to ensure independent SNPs). Rapture, MAF, LD and IBD are acronyms for RAD-capture (Ali et al. 2015), Minor Allele Frequency, Linkage Disequilibrium, and Identity By Descent, respectively. Note that SNP chips are an alternative genomic tool (not in this figure) often used MPS for SNP discovery.

Box 1. Roadmap for filtering reads from massively parallel sequencing (MPS).

Primary problem	*Possible filtering solution*	*References*
Sequencing errors	Ensuring accurate SNP calling: keeping SNPs with sufficient coverage, quality scores and genotype likelihood *	Davey *et al.* (2011); Kim *et al. (* 2011); Nielsen *et al.* (2011), Catchen *et al.* (2013) ; Marinier *et al.* (2015); Mastretta-Yanes *et al.* (2015); Andrews *et al.* (2016); Laehnemann *et al.* (2016)
	Removing singletons	
	Correcting substitution errors to improve the quality of assemblies	
Missing data	Keeping SNPs genotyped in at least a certain percent of individuals and populations. This threshold will largely be influenced by the number of samples initially genotyped and the quality of data required for the research question.	Hohenhole *et al.* (2010); Benestan *et al.* (2015)
Duplicated loci	Keeping biallelic SNPs by individual for diploid species	Gayral *et al.* (2013); Pujolar *et al.* (2013); Mandeville *et al.* (2015) Ferchaud & Hansen (2016);
	Removing loci with too high coverage (*e.g.,* the mean plus 2*standard deviation)	Bianco *et al.* (2014); Ferchaud *et al.* (2014);
	Keeping SNPs with heterozygotity inferior to 0.5	Hohenhole *et al.* (2010)
Linkage desequilibrium (LD)	Keeping only independent loci (required for many approaches), *e.g.,* keeping only one SNP per loci, or using a cut-off of r^2 if a reference genome is available and physical position of loci is known	Larson *et al.* (2014); Baird *et al.* (2015); Waples *et al.* (in review)
Hardy-Weinberg	Keeping SNPs in Hardy Weinberg Proportions (HWP) in most of the populations (some populations could have sampling error that create spurious HWP). Nevertheless, SNPs out of HWP should not be removed if the main goal of the study is to detect outliers potentially under selection.	Hess *et al.* (2012); Miller *et al.* (2012); Lexer *et al.* (2014); Lozier (2014); Benestan *et al.* (2015); Waples (2015)
Polymorphism	Keeping informative SNPs based on a Minor Allele Frequency (MAF) threshold (*e.g.,* MAF > 0.05 at the population level if only informative SNPs are necessary to revel population structure or MAF > 0.0001 at global level for removing sequencing errors)	Roesti *et al.* (2012)



Here is an example of 6 diploid individuals (ind.) genotyped at loci x, 20 bp long. Among this subset of individuals, 3 SNPs were discovered and accurately called (*), at nucleotide positions 4, 11 and 15. These 3 SNPs could be treated as three different markers (**). Several classic analysis would treat these 3 markers as independent whereas they are physically linked. To counteract this problem, researchers often retain only one SNP, for example the first one, here SNP 4 (see dashed line). However, in order to make use of all the 3 SNPs, the haplotype approach (combining the 3 SNPs in a single haplotype) could be used (***) when filtering and genotyping.

Figure 3.2. Roadmap for filtering reads.

Chapitre 4. Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (Homarus americanus).

Publié sous : Benestan L, Quinn BK, Maaroufi H, Laporte M, Fraser K, Greenwood SJ, Rochette R & Bernatchez L (2016). *Molecular Ecology*.

4.1. Résumé

Étudier comment les caractéristiques environnementales façonnent la structuration génétique des populations est cruciale pour comprendre comment ces dernières interagissent avec leur habitat et sont potentiellement adaptées à celui-ci, ce qui permet ensuite de les gérer en conséquence. En utilisant des approches de différenciation de la population (PD) combinées avec des analyses d'association environnementale (EA), nous avons évalué l'importance relative de la distribution spatiale, des courants océaniques et de la température de surface de la mer (SST) sur les patrons de variation génétique potentiellement neutre et adaptatif des populations de homards d'Amérique provenant de 19 sites d'échantillonnage. Dans un premier temps, les approches PD (en utilisant BAYESCAN, ARLEQUIN et OUTFLANK) ont trouvé en commun 28 SNPs potentiellement sous sélection divergente et 9770 SNPs potentiellement neutres. L'analyse de redondance (RDA) a révélé que la distribution spatiale, les courants océaniques (représenté par la connectivité larvaire) et les valeurs de SST expliquent 31,7% de la différenciation génétique potentiellement neutre, où les courants océaniques sont responsables de la majeure partie de cette relation (21,0%). Après avoir retiré l'influence de la distribution spatiale, aucune valeur de SST n'était significative pour la variation génétique potentiellement neutre alors que pour la variation génétique potentiellement adaptative, la valeur de SST minimale annuelle avait encore un impact significatif et expliquait 8,1% de la variation. Deuxièmement, les analyses EA (en utilisant des tests de corrélation de Pearson, BAYESCENV et LFMM) ont identifiés conjointement sept SNPs comme candidats potentiels à l'adaptation thermique. La co-variation de ces SNPs a été évaluée à l'aide d'une analyse spatiale multivariée (sPCA) qui a mis en évidence une association significative avec la température minimale annuelle, même après avoir tenu compte de l'influence de la distribution spatiale. Parmi les 505 SNPs candidats détectés par au moins une de ces approches, nous avons découvert trois polymorphismes situés dans les gènes précédemment déjà connu pour jouer un rôle dans l'adaptation thermique. Nos résultats ont des implications pour la gestion du homard d'Amérique et vise à fournir une base sur laquelle prédire comment cette espèce fera face aux changements climatiques.

4.2. Abstract

Investigating how environmental features shape the genetic structure of populations is crucial for understanding how they are potentially adapted to their habitats, as well as for sound management. In this study, we assessed the relative importance of spatial distribution, ocean currents and sea surface temperature (SST) on patterns of putatively neutral and adaptive variation among American lobster from 19 locations by using population differentiation (PD) approaches combined with environmental association (EA) analyses. Firstly, PD approaches (using BAYESCAN, ARLEQUIN and OUTFLANK) found 28 outlier SNPs putatively under divergent selection and 9,770 neutral SNPs in common. Redundancy Analysis (RDA) revealed that spatial distribution, ocean current-mediated larval connectivity, and SST explained 31.7% of the neutral genetic differentiation, with ocean currents driving the majority of this relationship (21.0%). After removing the influence of spatial distribution, no SST were significant for putatively neutral genetic variation whereas minimum annual SST still had a significant impact and explained 8.1% of the putatively adaptive genetic variation. Secondly, EA analyses (using Pearson correlation tests, BAYESCENV, and LFMM) jointly identified seven SNPs as candidates for thermal adaptation. Co-variation at these SNPs was assessed with a spatial multivariate analysis (sPCA) that highlighted a significant temperature association, after accounting for the influence of spatial distribution. Among the 505 candidate SNPs detected by at least one of these approaches, we discovered three polymorphisms located in genes previously shown to play a role in thermal adaptation. Our results have implications for the management of the American lobster and provide a foundation on which to predict how this species will cope with climate change.

4.3. Introduction

Incorporating environmental information into a population genetics framework is essential to identify the proximal factors that modulate the strength and interactions of evolutionary forces, which ultimately determine the extent and scale of local adaptation of living organisms (Manel & Segelbacher 2009). Towards this end, the field of landscape genetics aims to assess how environmental parameters influence the extent of genetic variation within and among populations (Manel *et al.* 2003). While landscape genetic studies of terrestrial species have been flourishing over the last decade (Manel & Holderegger 2013), the number of studies investigating marine species in a "seascape genetics" framework has been more limited (Storfer *et al.* 2006; Riginos & Liggins 2013; Kershaw & Rosenbaum 2014).

Marine species are typically characterized by the absence of visible physical barriers to gene flow over large geographic distances (Palumbi 1994). However, dispersal potential may vary across a fragmented seascape due to patterns and gradients of environmental factors such as ocean currents, temperature, and salinity. In particular, over the past five years, seascape genetic studies have shown that complex patterns of genetic connectivity are related to larval connectivity estimates based on ocean currents in a wide range of marine species (reviewed in Selkoe *et al*. 2016), including mussels (*Mytilus sp.*: Gilg & Hilbish 2003), urchins (*Centrostephanus rodgersii*: Banks *et al.* 2007), corals (*Acropora palmata*: Baums *et al.* 2006), barnacles (*Balanus glandula*: Galindo *et al.* 2006), snails (*Kelletia kelletii*: White *et al.* 2010), California spiny lobster (*Panulirus interruptus*: Iacchei *et al.* 2013), New Zealand rock lobster (*Jasus edwardsii:* Thomas & Bell 2013), crabs (*Carcinus aestuarii*: Schiavina *et al.* 2014), reef fish (*Elacatinus lori*: D'Aloia *et al.* 2013) and shrimp (*Pandalus borealis*: Jorde *et al.* 2015). However, most of these studies did not consider the potential impacts of environmental factors on adaptive genetic variation (but see Pujolar *et al.* 2014; Tepolt & Palumbi 2015). An "adaptive" perspective is desirable, given that the key questions of how and where gene flow is constrained are tightly linked to the fitness of individuals in their environment (Lenormand 2002). Therefore, elucidating the environmental determinants of population structure and local adaptation in marine ecosystems is a worthy enterprise that is needed to answer important questions of relevance facing marine conservation and management (Selkoe *et al.* 2008).

Investigating putatively adaptive genetic variation along environmental gradients in several populations represents a promising way to screen for evidence of local adaptation over large geographic areas (Nielsen 2005; Savolainen *et al.* 2013). The potential explanatory power of such investigation has been substantially enhanced by the development of increasingly affordable genomic tools for next generation sequencing (Willette *et al.* 2014). To date, only a few seascape studies have taken advantage of these tools to explore both adaptive and neutral genetic patterns in marine species (Gagnaire *et al.* 2012; Bourret *et al.* 2013; Hess *et al.* 2013; Bourret *et al.* 2014; Guo *et al.* 2015; Tepolt & Palumbi 2015).

The American lobster (*Homarus americanus*) supports the most important fishery in Canada (http://www.dfo-mpo.gc.ca). Consequently, sustainability of this fishery is a major concern for fishers and managers. Implementing sustainable management procedures requires an accurate description of population structure (Reiss *et al.* 2009). This need led to previous studies that documented neutral genetic structure of this species by means of microsatellites (Kenchington *et al.* 2009) and more recently by RAD sequencing (Benestan *et al.* 2015). Both studies detected the existence of two genetic clusters separating northern and southern samples of this species. These genetic clusters coincide with the occurrence of a discontinuity in larval exchange between these two regions, which suggests that ocean currents may promote "neutral" genetic divergence in this species (see Supplementary materials). In addition, this species' range spans a strong thermal gradient (Aiken & Waddy 1986) but the possibility of adaptive differentiation among populations associated with this environmental gradient remains to be tested. Documenting adaptive genetic structure will augment our understanding of conservation units based on neutral genes and may help establish effective conservation strategies (Allendorf *et al.* 2010). In particular, identifying the genetic basis of local adaptation to temperature is a major goal of conservation biology since it could help predict how a species will respond to climate change (Savolainen *et al.* 2013).

Temperature represents a key selective agent that appears to drive adaptive divergence among populations of many marine invertebrate species (Sanford & Kelly 2011). This is likely the case for the American lobster, which has a broad distribution along the Atlantic coast of North America, from 35.25°N in Cape Hatteras, North Carolina, to 51.73°N in the Strait of Belle Isle, Labrador (Lawton & Lavalli 1995). American lobsters are exposed to temperatures as low as -1°C and as high as 26°C (Aiken & Waddy 1986; Quinn &

Rochette 2015). Temperature has been shown to be an important determinant of metabolism (Qadri *et al.* 2007), behaviour (Crossin *et al.* 1998), and several life history traits of this species (Lawton & Lavalli 1995). In particular, sea surface temperature (SST) during summer months is critically important to lobster larvae, affecting their survival, development, and distance dispersed after hatching (MacKenzie 1988; Quinn *et al.* 2013).

Studies that searched for evidence of adaptive genetic variation have mostly used traditional population differentiation (PD) approaches (Jensen *et al.* 2016), which aim to identify loci putatively under selection by comparing the genetic differentiation index (F_{ST}) of each locus to values expected under a null model of neutral evolution (Francois *et al*. 2016; Jensen *et al*. 2016). One advantage of this approach is that it does not require *a priori* information concerning the environmental forces that act as selective pressures. Environmental-association (EA) analyses represent an alternative and/or complementary avenue to PD approaches that may allow detecting adaptive patterns missed by PD methods (Pritchard *et al.* 2010; Rellstab *et al.* 2015; Francois *et al.* 2016) insofar as the environmental variables investigated are relevant to genetic structure. They tend to provide evidence for adaptive genetic variation by seeking correlations between environmental variables and allele frequencies (reviewed in Rellstab *et al.* 2015). Both PD and EA approaches are prone to false positive associations (Frichot *et al.* 2012; Villemereuil & Gaggiotti 2015; Rellstab *et al.* 2015; Francois *et al.* 2016), but they can each detect loci under selection not identified by the other approach. Combining PD and EA approaches may thus provide an efficient strategy to identify patterns and causes of local adaptation (Rellstab *et al.* 2015; Gagnaire *et al.* 2015) while guarding against false positives (Villemereuil *et al.* 2014; Francois *et al.* 2016).

The goal of this study was to perform one of the first seascape genomics studies in a marine invertebrate by assessing the potential role of spatial distribution, ocean currents, and temperature in shaping both putatively neutral and adaptive genetic structure in American lobster. We jointly performed PD analyses and EA approaches (see Methods) on samples of egg-bearing female lobsters from 19 locations spanning most of the species' range. We then applied multivariate redundancy analyses to estimate the relative contribution of spatial distribution, ocean currents, and temperature to neutral and adaptive genetic patterns. Finally, we implemented a BLAST search on the best candidate SNPs defined by both PD analyses and EA approaches to identify genes with molecular functions potentially involved in local

adaptation to temperature among American lobster inhabiting different locations.

4.4. Results

Dataset definition: neutral versus putatively adaptive markers

A total of 13,688 filtered and informative SNPs within 8,094 sequences were successfully genotyped from 562 egg-bearing female American lobsters (Table S1). The number of SNPs per sequence ranged from 1 to 7, with about 48.5% of the sequences containing 1 or 2 SNPs. Missing genotype data per SNP averaged 7.2%. BAYESCAN detected 10,544 SNPs (77.0%) putatively neutral, 3,119 SNPs (22.8%) putatively under balancing selection and 35 SNPs (0.2%) putatively under divergent selection, at the 5% significance level. Based on the q-value model, we found 22 SNPs showing decisive evidence for selection with a Bayes factor > 100 (Figure 4.1). ARLEQUIN identified 12,275 putatively neutral SNPs (89.7%), 164 SNPs putatively under divergent selection (1.2%) and excluded 1,249 SNPs (9.1%) due to too much missing genotype data. At the same significance level ($P < 0.05$), OUTFLANK identified 41 SNPs under divergent selection. BAYESCAN, OUTFLANK and ARLEQUIN analyses shared 28 SNPs identified as being putatively under divergent selection (Figure 4.2a) for which F_{ST} values varied between 0.0321 and 0.1780 among the 19 sampling sites compared to an average F_{ST} value of 0.0018 over all markers. These 28 candidate SNPs were used for downstream analyses of adaptive genetic structure. Similarly, we used the 9,770 putatively neutral SNPs detected by both BAYESCAN and ARLEQUIN for downstream analyses of neutral genetic structure.

Environmental factors shaping neutral and adaptive genetic structure

Based on the Kaiser-Guttman criterion, 10 PCs were meaningful and kept for the 9,770 putatively neutral SNPs, which accounted for more than 70.0% of the total putatively neutral genetic variation. For this putatively neutral genetic variation, one temperature descriptor (maximum annual winter temperature), two geographic vectors (dbMEM-1 and dbMEM-3; Table 2) and five vectors representing a network of ocean currents (AEM-1, AEM-2, AEM-4, AEM-7 and AEM-9; Table 4.2) were selected by the *ordistep* function and included in the RDA framework. The RDA was globally significant ($P = 0.001$) with an adjusted coefficient of determination (R^2_{adj}) of 0.317. The first two axes of the RDA

accounted for 16.7% and 10.8% of the genetic variation, respectively. By considering the most explanatory independent parameters selected by the *ordistep* function, the marginal ANOVA showed that one geographic vector (dbMEM-1) and four vectors representing ocean current networks (AEM-1, AEM-4, AEM-7 and AEM-9) were all significant predictors of the putatively neutral genetic variation ($P < 0.05$; Table 2). When partitioning the relative importance of spatial distribution and ocean currents on neutral genetic variation (partial RDA), spatial distribution (dbMEM-1) and ocean currents (AEM-1, AEM-4, AEM-7 and AEM-9) were both still significant but variation explained by ocean currents was three times (21.0%) that explained by spatial distribution (7.6%) (Table 4.2).

For the analysis based on the 28 SNPs putatively under divergent selection, we retained five PCs based on the Kaiser-Guttman criterion, which together accounted for 78.5% of the putatively adaptive genetic variation. Here, three temperature descriptors (mean summer, minimum annual, and maximum annual SST), and one geographic vector (dbMEM-1) were selected by the *ordistep* function and included in the RDA framework. The RDA was globally significant ($P = 0.004$) and revealed an adjusted coefficient of determination of 0.301. The first two axes of the RDA accounted for 35.9% and 6.5% of the genetic variation, respectively (Figure 4.4). The marginal ANOVAs for the RDA indicated that minimum annual, mean summer, and maximum annual SST were the most significant predictors of the putatively adaptive genetic variation ($P < 0.05$; Table 2). However, the ANOVA for the partial RDA showed that minimum annual SST was the only significant predictor of the putatively adaptive genetic variation ($R^2_{adj} = 0.281$, $P = 0.001$) when spatial distribution was taken into account (Table 4.2).

Population differentiation (PD) approaches versus Environmental Association (EA) analyses: overlapping SNPs

The LFMM analysis identified a total of 248 SNPs showing at least one significant association with the nine temperature parameters (Table 4.2). BAYESCENV was markedly more conservative and identified only 26 SNPs potentially linked to temperature. Correlation tests between minor allelic frequencies (MAF) and the nine temperature parameters revealed a set of 123 SNPs showing significant associations (81 positive: $r > 0.70$; 42 negative: $r < -0.70$) with at least one of the nine temperature parameters ($P < 0.001$). We identified seven overlapping SNPs (Figure 4.2b) among these three EA analyses based on different models

and assumptions (LFMM, BAYESCENV and Pearson correlation test), six of which were also among the 28 common SNPs detected by the three PD programs.

Clines in allele frequency

For the sPCA at the seven putatively adaptive SNPs identified by all EA analyses, we retained only the first positive eigenvalue since an abrupt decrease in eigenvalues was observed after it (Figure 4.4), which may indicate the boundary between true patterns and non-interpretable structures. The linear regression of the genetic locality scores extracted from the sPCA against spatial distribution and environmental factors revealed that the best predictors of locality scores were minimum annual SST (R^2 = 0.382, P = 0.002) and mean winter SST (R^2 = 0.306, P = 0.008). dbMEM vectors were not significantly related to genetic locality scores (P > 0.05), whereas latitude and longitude were (R^2 = 0.178 and 0.157, P = 0.040 and 0.052 respectively), albeit less strongly so than the temperature (SST) parameters. Thus, the synthetic multi-locus cline of allele frequency at these SNPs showed a stronger association with either minimum annual SST or mean winter SST compared to latitude and longitude.

Gene ontology

A total of 432 candidate sequences contained the 505 unique SNPs significantly associated with temperature or defined as potentially under divergent selection by the genome scan analyses. The alignment of these candidate sequences to the complete transcriptome of the American lobster merged in a total of 122 contigs. The BLAST analysis on these 122 contigs against the SWISS-PROT database provided a total of 15 hits with an *E*-value smaller than 10^{-6}. From these 15 successfully annotated genes, five carried a non-synonymous SNP (Table 4.3). Only two of these non-synonymous SNPs - SNP *20131* and SNP *49442* - may have an impact at the protein-level since these substitutions lead to amino acid with different properties, which is not the case for the other four. The SNP *20131* is situated in the gene *GRID1*, which encodes glutamate receptor delta 1, a subunit of glutamate receptor channels that mediate most of the synaptic transmissions in the central nervous system (Guo *et al.* 2007). This mutation (Leu/Ile) is located in the extremity of the C-terminal protein that could interact with the N-ethylmaleimide-sensitive fusion (NSF) and soluble NSF attachment (SNAP) proteins, which are involved in glutamate activity. Similarly, the SNP *49442*, located in the *Vps16* gene, may interact with the SNP *20131* through the proteins NSF

and SNAP (Osten *et al.* 1998), which are both involved in ATPase activity pathway and then influence the metabolism activity in different thermal regimes. In the remaining nine synonymous polymorphisms, we also discovered the SNP *11147,* detected by the COR method ($r > 0.75$, $P < 0.001$ for mean year SST), which has a higher frequency of its alternate allele (T) in warmer populations than in colder populations (Figure 4.5). Interestingly, this SNP (A/T) is located near the active site of the *β-galactosidase* gene, which produces a hydrolase enzyme well known to be involved in molecular cold adaptation processes in several organisms (Table 4.3; reviewed in D'Amico *et al.* 2002).

4.5. Discussion

Despite the socio-economic importance of the American lobster in the Northwest Atlantic, we have very limited understanding of how the marine environment affects this species' genetic structure. In response to this knowledge gap, we conducted what may be the broadest seascape genomics study to date on a non-model invertebrate species. Using 13,688 RAD-sequencing markers we applied traditional population genetics approaches (population differentiation (PD) and environmental association (EA) analyses) jointly with more general multivariate statistical frameworks (RDA and sPCA) in an attempt to gain new insights into the key determinants of genetic structure and local adaptation in this species. Our results revealed that both geographic distance but more importantly ocean currents were involved in explaining and shaping neutral genetic population structure, whereas minimum annual sea-surface temperature (SST) was identified as a main potential selective agent driving local adaptation. From the combination of statistical analyses, we detected three candidate genes (*GRID 1*, *Vps16*, *β-galactosidase*), including one gene (*β-galactosidase*) with allele frequencies exhibiting a pronounced temperature-associated cline. This *β-galactosidase* gene has been identified as an important functional gene involved in cold adaptation in many microorganisms (Hoyoux *et al.* 2001; Karasova *et al.* 2002) because it produce an enzyme that may have a higher catalytic acivity toward low temperatures, and may play a similar role in American lobster.

Drivers of neutral and adaptive genetic structure

Marine species are typically characterised by high gene flow and weak genetic

structure (Waples 1998). Nonetheless, there is a growing number of seascape studies highlighting the role of geographic distances and ocean currents in shaping patterns of marine species' population structure (White *et al*. 2010; Amaral *et al*. 2012; Iacchei *et al*. 2013; Jorde *et al*. 2015). White *et al.* (2010) highlighted the benefits of using oceanographic data to advance our ability to interpret population structure of species with pelagic larval stages and high gene flow. They demonstrated that ocean currents better explained genetic patterns of the whelk, *Kelletia kelletii*, than geographic distance. Similarly, another recent study by Jorde *et al*. (2015) revealed that both geographic distances and larval drift with currents help elucidate large-scale genetic differentiation patterns in northern shrimp, *Pandalus borealis*. In agreement with these studies, we found that ocean currents (21%) were more useful in explaining genetic structure in American lobster than geographic distances alone (7.6%).

In agreement with Benestan *et al.* (2015), the most significant Moran Eigenvectors maps (dbMEM-1; Figure 4.3a), which represent the influence of distances on neutral genetic structure, highlighted the North and South dichotomy resulting in two genetic groups of lobster. For both regions the most significant Asymmetrical Eigenvectors maps (AEM-4; Figure 3b), representing larval dispersal within a single generation, indicated that the Gaspé and the Scotian Shelf Currents impact neutral genetic structure (Figure 3c). Indeed, the Gaspé Current is likely to carry pelagic larvae along the Gaspé Peninsula towards the southern Gulf of St Lawrence and western coast of Cape Breton, connecting sampling sites in this area (GAS, MAL, MAG and DIN) that showed very low and non-significant F_{ST} values (Supplementary material, Figure S4.3). Similarly, the Scotian Shelf current could contribute to "homogenizing" lobsters in and near the eastern Gulf of Maine, potentially causing the lack of significant genetic divergence previously observed among offshore (OFF and BRO) and inshore (LOB; Figure S4.3) sampling sites near the south-western part of the Scotian Shelf. However, current-mediated drift of larvae from the Gulf of St. Lawrence to the Gulf of Maine almost never occurred within one generation (Figure S4.2). Over multiple generations some connectivity likely occurs between these regions, following a "stepping stone" model of gene flow, which would prevent complete isolation of lobsters in these two regions; however, this would not be enough to homogenize them, thus supporting the observed north-south genetic divide observed for this species (Benestan *et al.* 2015 and present study). On average, lobster larvae drift approximately 129 km between hatch and settlement, with the majority

(90 %) drifting ≤ 410 km (Quinn, Chassé, and Rochette, in prep). Therefore, genetic dissimilarities observed between sites in the north and south regions, as well as between far-apart sites within these regions (e.g., TRI and GAS), are likely due, at least in part, to the limited amount of current-mediated larval exchange between them.

Importantly, these findings provide empirical support for modeled estimates of larval drift and connectivity for this species (Quinn 2014) and they demonstrate that ocean currents play a meaningful role in shaping American lobster neutral population genetic structure. Nevertheless, larval connectivity via ocean currents "only" explained approximately 21.0% of the neutral genetic variation observed among lobsters from our 19 study locations. This could be partly due to limitations of the dispersal modeling system we used, which at present lacks some aspects of lobster biology (*e.g.,* larval behavior, mortality, egg production) that could impact dispersal patterns, but for which information from across the species' range is currently unavailable (Quinn, 2014). Processes occurring at other points in the lobster's life cycle (*e.g.,* movement by adults on the sea floor, post-larval swimming and settlement behaviours) might also play a role in structuring lobster populations (Campbell & Stasko 1986; Chiasson *et al.* 2015) and would thus lead to different connectivity patterns than inferred by larval dispersal alone. Additionally, processes occurring over multiple generations could lead to different patterns than those observed in single-generation simulations and should thus be comprehensively investigated in the future.

We used Redundancy Analysis (RDA) instead of performing a linear regression between Euclidian or oceanographic distances and F_{ST}, which has been the most common approach used in seascape studies thus far (White *et al.* 2010; Godhe *et al.* 2013; Jorde *et al.* 2015). However, the assumption of independence is violated when performing linear regressions on F_{ST} values, which may make this approach statistically inappropriate (Boldina & Beninger 2016). The approach we used overcame this issue by synthesizing multivariate genetic data (SNPs) into vectors that were compared to Moran Eigenvectors maps (dbMEM) of geographic distances and Asymmetrical Eigenvectors maps (AEM) of larval dispersal mediated by ocean currents. Moreover, these methods depicted a greater influence of ocean currents and geographic distances on genetic variation than if we had used Euclidian distances or latitude and longitude data in a linear regression analysis. Indeed, performing RDA based on latitude and longitude alone would have resulted in $R^2_{adj} = 0.030$ (details not

shown), which is four times lower than the R^2_{adj} = 0.115 obtained with dbMEM variables. Our study therefore provides evidence of the relevance of considering dbMEM for future landscape studies, especially when the spatial context is potentially non-linear (see Garroway *et al.* 2013; Breyne *et al.* 2014).

The effects of demographic history and isolation by distance on genetic variation can confound effects of environmental variables, potentially leading to incorrect interpretations regarding local adaptation (Excoffier & Ray 2008). It is therefore important to account for the spatial distribution of populations or sample locations when attempting to assess the effect of environmental factors on genetic variation. To that end we used a partial RDA to investigate genetic variation in lobster and found that when accounting for effects of spatial distribution of sample locations SST was not a significant explanatory variable of neutral genetic variation, whereas adaptive genetic differences were significantly related to minimum annual SST. SST likely provides the best available index of spatial variation in selection imposed by temperature on all life stages of lobsters (see Methods), and our results suggest that spatially varying selection in American lobster populations is mainly driven by minimal temperatures encountered by larval or benthic stages. Spatially varying selection is a signature of local genetic differentiation caused by disparate *in situ* mortalities within a single generation (Endler 1986). Spatially varying selection has been evidenced in several marine species, for example American eel (*Anguilla rostrata*: Gagnaire *et al.* 2012; Laporte *et al.* 2016) and acorn barnacle (*Semibalanus balanoides:* Schmidt & Rand 2001; Véliz *et al.* 2004). Following the method proposed by Gagnaire *et al.* (2012), we also revealed that the genetic cline based on the seven candidate SNPs identified commonly by EA approaches was better explained by minimum annual SST (R^2_{adj} = 0.382) than by geography (R^2_{adj} = 0.178 for latitude and 0.157 for longitude). This suggests again that the effect of temperature prevails over that of the spatial structure alone.

Here, we highlighted that minimum annual SST may be a potential selective agent driving local adaptation. Whereas SST estimates are correlated to bottom temperatures (Drinkwater & Gilbert 2004; Brickman & Drozdowski 2012a), which describe the environment occupied by sampled benthic stages (adults) of the lobster life cycle, SST is most likely to be experienced by pelagic larval phase where it could be a significant source of mortality. For instance, *in situ* observations showed that postlarvae tend to remain in waters

above 12°C (Annis 2005) and an increase in mortality below that temperature has been documented in experimental conditions (MacKenzie 1988). It is also noteworthy that larvae originating from a cold-water region have been found to exhibit a shorter development time in cold water than larvae originating from a warm-water region (Quinn *et al.* 2013), which may also suggest that lobsters are adapted to the thermal regime they occupy. However, minimum annual SST occurs during winter months, which is a period when the larval phase is already over. Therefore, this outcome might suggest that cold-tolerance is more important for the benthic life stages than larvae, where some juveniles/adults may be better able than others to tolerate certain low temperature and will remain in the population, through the process of natural selection.

Combining Population Differentiation and Environmental Association approaches

Detecting local adaptation occurring in complex landscapes is not optimally achieved using a single approach (Rellstab *et al.* 2015). Combining population differentiation (PD) and environmental association (EA) approaches to detect candidate loci of thermal adaptation not only reduces false positive discoveries, but also maximize our chances of detecting potential signals of selection (Francois *et al.* 2016). Recently, Vatsiou *et al*. (2016) showed that combining seven analyses for the detection of selective sweeps could greatly increase the ability to pinpoint the most likely genomic regions under selection. In this study we employed three different analyses for each approach (Figure 4.2), which led to the identification of 505 candidate SNPs, a small fraction of which (six SNPs) were identified by all six analyses. Overall, we found that EA analyses identified more candidate markers (370 SNPs) than PD analyses (170 SNPs). These outcomes are in agreement with a simulation study demonstrating that EA approach have more power to detect loci under divergent selection than PD approach (Villemereuil *et al.* 2014), which is not surprising given that the former (but not the latter) utilize environmental information (here SST) to depict signals of selection.

We found 28 candidate genes that were identified by all three PD analyses, which represent only 16.5% of all outliers detected by at least one of these analyses. The number of outliers discovered by BAYESCAN and OUTFLANK tests (36 and 41 outliers respectively) was about four times lower than the number found by ARLEQUIN (123 outliers). This outcome is in agreement with results of a simulation studies showing that ARLEQUIN

consistently found more outliers and had highest type I and type II errors in their simulation scenarios in comparison to other methods such as BAYESCAN (Narum & Hess 2011). In contrast, BAYESCAN and OUTFLANK performed much more similarly by finding 80% of the same candidate SNPs. OUTFLANK identified slightly more candidate SNPs than BAYESCAN (41 against 36) although it is supposed to have a lower false discovery rate than the latter (Whitlock & Lotterhos 2015). However, the slightly higher identification rate of OUTFLANK does not necessarily result from more false positives (type I errors) but could also be due to fewer false negatives (type II errors). In species exhibiting isolation by distance (IBD), such as American lobster, a large number of false positives may be detected when testing for SNPs under selection. In the presence of IBD, Whitlock & Lotterhos (2015) recommended using other methods (*e.g.* OUTFLANK, Fdist2, FLK) than BAYESCAN because of its higher rate of false positive in such circumstances. Here, we followed this recommendation and underlined that BAYESCAN and OUTFLANK gave very similar results in an IBD system where F_{ST} is very low, which was never shown before. Indeed, the assumptions is that BAYESCAN may handle the differences between heterozygosity among loci better in a cases of less structured populations (*e.g.* $F_{ST} < 0.005$), which was different from the system tested ($F_{ST} > 0.05$) by Whitlock & Lotterhos (2015).

We identified only a small subset of seven overlapping SNPs (1.8%) that displayed temperature-associated clines in all three of the genotype-temperature association tests we conducted. Villemereuil *et al.* (2014) similarly found on average from 1 to 5% of overlap between loci considered as positives by all three analyses they used, which were very similar to ours; LFMM, BAYESCAN (we used BAYESCENV, but results were 90% similar to those obtained with BAYESCAN) and a simple linear regression analysis (similar to our COR method). This low number of overlapping SNPs reiterates the high degree to which outcomes differ between analytical approaches. Since Villemereuil *et al.* (2014) revealed that these methods tend to agree more on true positives, consistency among methods can be used to account for the errors that each analysis makes and improve the identification of true positives. Nevertheless, none of the SNPs detected by all PD and EA approaches combined was among the most likely candidate to thermal adaptation detected by the BLAST. More broadly, we found that several candidate SNPs were only detected by one analysis, including the three strongest SNPs candidate (SNP *49442,* SNP *20131* and SNP *11147*). As each

approach has its advantages and disadvantages (Rellstab *et al.* 2015), our results reiterate the importance of utilizing several analyses and approaches in the field of landscape genomics.

Finding a candidate gene for thermal adaptation

Numerous marine invertebrates have evolved biochemical adaptations to reduce the negative consequences of unfavourable changes in temperature (Hochachka & Somero 2014). By combining population PD and EA approaches we identified a total of 505 SNPs as potential selection targets among the 19 sampling sites. We found only 15 SNPs in coding regions of known genes in the SWISSPROT database, which is not surprising given that the genome of the American lobster has not been sequenced and a large fraction of genes remains without any annotation (Pavey *et al.* 2012). Among these markers we discovered two non-synonymous polymorphisms (SNP *49442,* SNP *20131*) and one synonymous polymorphism and (SNP *11147*) with putative functions that are compatible with the hypothesis of adaptive selection acting on encoded protein. The *20131* SNP is located in the *Grid1* gene, which may play key roles in synaptic plasticity (Guo *et al.* 2007) and was found to be potentially involved in high-altitude adaptation in Tibetan pigs (*Sus scrofa;* Ai *et al.* 2014). The SNP *49442* belongs to the *Vsp16* gene, which is involved in vacuole protein sorting and organelle assembly in *Saccharomyces cerevisae* (Sato *et al.* 2000) and showed upregulated expression in sweet corn (*Zea mays*) under heat stress (Li *et al.* 2015). These findings suggest that these two SNPs may also be involved in thermal adaptation in American lobster, although more research will be needed to determine what their functions may be in this species as well as their protein structures.

The synonymous SNP *11147* is located near the active site of the β-galactosidase gene. β-galactosidases have a wide phylogenetic distribution, encompassing plants, animals and microorganisms (Wallenfels & Weil 1972). The β-galactosidase gene produces a cold-adapted enzyme, which hydrolyzes lactose into galactose and glucose and has a stable enzymatic activity at temperatures below 8°C. While the SNP *11147* is synonymous, there is a growing body of evidence demonstrating that synonymous polymorphism may face strong selection and could alter the phenotype by influencing several important cellular processes (*e.g.* transcription, splicing, mRNA transport or translation, enzyme activity and production; reviewed in Plotkin & Kudla 2011). Here, for most of the sampling sites, we found that a greater proportion of individuals occupying warmer habitats had the alternate allele (T)

compared to lobsters living in colder habitats. Nevertheless, this is not true for four sites (CAR, GAS, OFF and BRO), where allele frequencies do not match well with the mean summer SST. We have no explanation for this gene-temperature mismatch at the CAR and GAS sites, but we can envision two reasons for the mismatch at the OFF and BRO sites. First, SST may not be the best predictor of allele frequency at these sites since they are located offshore, where lobsters occupy deeper (up to 200 meters deep) and likely colder waters during summer months. However, during winter months temperature tends to actually be higher in deeper than in shallower water in this part of the species range, and adult lobsters make seasonal migrations from shallow to deep in the fall-winter to experience warmer conditions over the winter months (Robichaud & Campbell 1991). Secondly, it is plausible that there is a lot of mixing between animals sampled in LOB and those in OFF-BRO (see Figure S4.2, S3). Overall, our results suggest that functional *β-galactosidase* SNP may play a key role in the thermal adaptation of American lobster populations inhabiting varying temperatures regimes in the Northwest Atlantic. Nevertheless, the pattern we see needs to be investigated more in a future study.

Future directions

Considering processes that govern genetic structure with a broad perspective is crucial for understanding the forces that impact species' demography and evolution. Here, our best RDA model, which included spatial distribution, ocean currents, and SST explained 31.7% of the neutral genetic structure. Consequently, much of this genetic structure remains unexplained. This result is comparable to that obtained by Selkoe *et al.* (2014) who found that variation in the genetic patterns of nine Hawaiian marine species cannot be explained by geography, dispersal ability and habitat factors alone (R^2_{adj} ranging from 0.11 to 0.66). Indeed, part of the neutral genetic variation is the result of random processes (resulting from genetic drift and mutation) and it is unlikely that it would ever be possible to explain 100% of this structure. Still, other biophysical and geographical proprieties of habitats such as bathymetry, bottom temperature, productivity, salinity, colonization history, pollution, and anthropogenic movements may also contribute to demographic isolation across American lobster populations. For example, previous marine genetic studies have shown that bottom temperature in northern shrimp (*Pandalus borealis*; (White *et al.* 2010; Godhe *et al.* 2013; Jorde *et al.* 2015) as well as bathymetry in cusk (*Brosme brosme*: Knutsen *et al.* 2009) and

deep-sea sharks (*Centroscymnus coelolepis*: Catarino *et al.* 2015) may affect genetic structure. While these factors can also influence genetic structure in American lobster, we did not have the necessary data to test this possibility. This may help explaining the relatively low level of variance explained by our models and shows that the influence of these factors should be investigated in future studies. On the other hand, SST likely co-varies with several of these other factors so the interpretation of temperature trends must be done cautiously, as is the case in any correlative study. Additionally, we only sampled females to investigate American lobster genetic structure, and thus we have not observed genetic structure in males, which could potentially be different (*e.g.,* if there is sex-biased dispersal).

Without additional resources on the American lobster genome, we were only able to produce a list of loci that are potentially under selection or linked to alleles under selection and link the variation at these loci with relevant environmental variables that provided the selective pressures (here, SST). This list of loci represents only a very small portion of the genome potentially under divergent selection; other targets of selection may have become lost when generating the libraries and sampling DNA fragments. Nonetheless, we detected three candidate genes that may have potential effects on thermal adaptation of American lobster populations. However, polymorphisms identified as potential targets of selection are usually only statistically linked with close targets of adaptive significance, and performing a site-directed mutagenesis experiment on β-galactosidase is required to draw firmer conclusions about this gene's function (Barrett & Hoekstra 2011).

4.6. Material and methods

Sampling and genotyping

Between May and August 2012 we sampled a total of 696 egg-bearing female American lobsters from 19 locations spanning most of the species' range along a pronounced gradient of sea-surface temperature (SST). We only sampled egg-bearing female since they are thought to display homing behaviour related to spawning and hatching grounds and therefore better informative about actual genetic population structure (Pezzack & Duggan 1986). Of the 19 sampling sites included in this study, 17 were previously analyzed in Benestan *et al*. (2015) for other objectives than seascape genomics, namely potential for

population assignment. Yet, adding two new sites led us to resume bioinformatics analyses from the beginning. Sampling, DNA extraction, RAD-sequencing library preparation, sequencing with Illumina technology, and bioinformatic analyses using STACKS v. 1.09 program (Catchen *et al.* 2013) followed the methods described in Benestan *et al*. (2015). From the dataset generated in that study, we developed a set of 13,688 filtered SNPs, which excluded SNPs that were not genotyped in at least 80% of the individuals and 70% of the locations, or did not show a minor allele frequency of at least 0.05 in all locations (see Table 2.1 and Benestan *et al.* (2015) for justification).

Population differentiation (PD) approach

We searched for loci with a level of population differentiation exceeding neutral expectations using three F_{ST}-based outlier analyses. First, we used the software OUTFLANK (Whitlock & Lotterhos 2015), which calculates a likelihood based on a trimmed distribution of F_{ST} values to infer the distribution of F_{ST} for neutral markers. OUTFLANK was run with default options (LeftTrimFraction=0.05, RightTrimFraction=0.05, Hmin=0.1, 19) and identified outlier SNPs across the 19 sites based on the Q-threshold of 0.05. Secondly, we detected outlier SNPs with BAYESCAN v. 2.1 (Foll & Gaggiotti 2008), a Bayesian method based on a logistic regression model that separates locus-specific effects of selection ("adaptive" genetic variation) from population-specific effects of demography ("neutral" genetic variation). BAYESCAN runs were implemented using prior model (pr_odds) of 10,000, as recommended by Lotterhos & Whitlock (2015), including a total of 10,000 iterations and burn-in of 200,000 steps. Finally, we also identified outlier SNPs using ARLEQUIN v. 3.5 (Excoffier & Lischer 2010), which was run using 100,000 simulations and 1,000 demes. ARLEQUIN is based on the infinite island model that integrates heterozygosity and simulates a distribution for neutrally distributed markers. A Q-value of 0.05 was used as threshold for statistical significance for OUTFLANK and BAYESCAN and a P of 0.05 for ARLEQUIN. All outlier analyses were conducted on the entire data set divided according to sampling location. Using the results of these three analyses we divided our dataset in two categories, putatively neutral SNPs and SNPs putatively under divergent selection (SNPs putatively under balancing selection were removed), in order to then infer demographic and potentially adaptive processes (Beaumont & Balding 2004). A SNP was considered as being putatively under divergent selection if all three PD identified it as an

outlier.

Spatial structure and environmental factors

Spatial structure was modeled with Cartesian coordinates and distance-based Moran's eigenvector map (dbMEMs) variables obtained through a Euclidian distance matrix. These dbMEMs (hereafter spatial distribution) are independent vectors that summarize the spatial structure associated with the neighbourhood network (the distance matrix) across scales (Borcard & Legendre 2002), thereby representing a spectral decomposition of the spatial relationship among the study sites. A numerical simulation study has shown that analysis using dbMEMs is capable of detecting spatial structure at several scales, which can then be used to control for spatial correlation in tests of y ~x relationships (*e.g.,* genetic-environment relationships in seascape genetics; (Peres Neto & Legendre 2010). To calculate dbMEMs, we first converted degrees North latitude and West longitude to Cartesian coordinates with the *geoXY* function available in the *SoDa* package of R software v. 3.1.3 (Team R core 2014). Then, we computed a Euclidian distance matrix on the Cartesian coordinates using the *dist* function and we performed the *PCNM* function (permutations = 1000) on this matrix. The *PCNM* function, available in the *PCNM* package, transformed the spatial distances to rectangular data that are suitable for constrained ordination (Legendre *et al.* 2012).

Environmental factors considered in our seascape-genomic analyses were larval connectivity estimates based on ocean currents (see next paragraph for details) and nine estimates of sea-surface temperature (SST). We considered only SST (not bottom temperature) because empirical data were not readily available for all our sampling locations at daily intervals over multiple years. In contrast, well-validated bottom temperature data were not available over the spatial and temporal domains needed in the present study. While SST directly impacts planktonic lobster larvae, bottom temperature would be more representative of potential selection acting on benthic juvenile and adult lobsters. However, SST and bottom temperature tend to be correlated over much of the geographical domain studied here (Drinkwater & Gilbert 2004; Brickman & Drozdowski 2012a). Since temperature may affect different life-history stages of the American lobster at different times of the year (Aiken & Waddy 1986; see also Introduction), we estimated the following nine metrics of SST: maximum, minimum, average SST from April to September (spring and summer), from October to April (fall and winter) and over the entire year. We estimated these

nine SST indices for each year between 2002 and 2012, and in analyses used the average value of each index over these 11 years. The SST data for our 19 study locations were generated by the Remote Sensing Laboratory of the Maurice Lamontagne Institute and obtained from Observatoire global du Saint-Laurent-OGSL database (http://ogsl.ca), which contains geo-referenced SST along North America's coastlines with a nominal spatial resolution of 1.1 km and a 24-hr update frequency. The nine temperature metrics estimated for each sampling location are included as Supplementary materials (Table S4.1).

Larval connectivity values among our sampling sites, which reflect the estimated spread of larvae from a spawning site to a settlement site as a result of ocean currents, were derived from simulations with an individual-based biophysical dispersal model of American lobster larvae (Chasse & Miller 2010) coupled to a three-dimensional physical oceanographic model (CANOPA) of the Atlantic Shelf of eastern North America (longitude: 71.5°-54.9°W; latitude: 38.6°-52.0°N; Brickman & Drozdowski 2012b). The physical oceanographic model has a spatial resolution of 1/12° N or W (~9 km x 6 km, or 54 km^2) and simulations were run over eight years, from 2005 to 2012. During each simulation, clusters of larvae were released every 12 hours in the months of June-September, when larval release and drift occur in nature (MacKenzie 1988; Quinn *et al.* 2013; Quinn & Rochette 2015). Larvae were released in same quantity and at same time in all cells of our model domain that fell within the lobster's historical range (Pezzack 1992), with a total of 2.16 x 10^9 larvae released per year per ~54 km^2 cell (Quinn 2014, preliminary values based on those from Chassé and Miller 2010). Larvae drifted passively at the surface, and no mortality was included. Time spent drifting was controlled based on (i) water temperature experienced by larvae, (ii) temperature-dependent development equations derived from laboratory studies on this species (MacKenzie 1988; Quinn *et al*. 2013), and (iii) settlement beginning halfway through stage IV (Cobb *et al.* 1989) and occurring where bottom temperature was ≥ 10°C (Chiasson *et al.* 2015). Positions of larvae within the flow field were tracked at 5-min time steps, which allowed the number and origin of settling larvae for each model cell to be determined. Additional details concerning this model can be found in Quinn (2014) and Quinn, Chassé, and Rochette (*in prep*).

For determination of connectivity, the model's domain was divided into 5400 km^2 geographic blocks ("source-sink areas", n = 338 total) containing 100 oceanic model cells

each (see Figure S4.1), among which dispersal probabilities were calculated (Figure S4.2). One of the 19 study sites (named BON) fell outside the model's geographic domain and could not be used to make pairwise estimates of connectivity (Table 4.1, Figure S4.1 and S4.2). In each year, the number of larvae released from and settling in each of the remaining 18 sites' blocks was calculated, as was the number of larvae exchanged by each pair of blocks. Larval connectivity between each pair of sites was determined based on whether dispersal probability was 1, they are said to be connected (yes, 1) or not connected (0, no) across all 8 years of model simulations (Figure S4.2) between the two sites of a pair, and then used to calculate Asymmetric Eigenvector Maps (AEM). AEM is a spatial eigenfunction method developed to model multivariate (*e.g.* genetic data) spatial distributions generated by an asymmetric, directional physical process, such as current-driven larval dispersal (Blanchet *et al.* 2011). The nodes-by-edges matrix, which translates the larval connectivity matrix into a vector of weights at each site (based on the absence/presence of connectivity), was constructed with 18 nodes (*i.e.* sites) and 25 edges (*i.e.* connectivity links obtained from our matrix data). From this matrix, the calculation of AEM resulted in thirteen AEM vectors (hereafter ocean currents) reflecting the ocean currents network obtained from our biophysical dispersal model.

Redundancy Analysis (RDA): linking genetic structure to environmental factors

We conducted redundancy analyses (RDA) to investigate the relative contribution of spatial distribution, ocean currents, and temperature to both neutral and putatively adaptive genetic structure at all study sites except BON (outside of the model range). RDA is a direct extension of multiple regression to model multivariate response data (Legendre & Gallagher 2001). We first attempted to reveal the relationship between spatial distribution (using dbMEMs vectors) and/or ocean currents (using AEM vectors) with the neutral and adaptive genetic structure using a RDA. Then, we implemented a partial RDA, which partitioned the total explained genetic variation among spatial distribution, ocean currents, and temperature (SST) to investigate the separate and joint influences of spatial distribution and environmental variables on genetic structure, thus overcoming collinearity issues. Using sampling sites as subjects, we assessed the variability in minor allele frequencies (MAF) of SNPs (response variables) that could be explained by our explanatory variables (spatial distribution, ocean currents, SST). MAF were calculated in Plink v.1.9 using all 13,688 SNPs

available, as recommended by Manel *et al.* (2013). Because the PCNM method performs better on detrended data, *i.e.* data from which the broad-scale trend has been removed (Borcard & Legendre 2002), we applied a detrending on the raw MAF data using the *decostand* function with the hellinger method available in *vegan* package in R (Oksanen *et al.* 2007). Next, we performed principal component analysis (PCA) of the MAF data and only retained the meaningful principal component factors (PCs) with eigenvalues greater than 1, according to the Kaiser–Guttman criterion (Yeomans & Golder 1982). The independent parameters that best explained variability in the PC factors were selected through a stepwise procedure elimination (forward and backward giving similar results) using the *ordistep* function available in the *vegan* package. The *ordistep* function selects variables in order to build the "optimal" model, which is the model with the highest adjusted coefficient of determination (R^2_{adj}). In the case of the partial RDAs, the effect of spatial distribution on genetic structure was "subtracted" (dbMEM vectors used as covariables) and constrained ordination was performed on the residual variability of the genetic data.

For all four tests (neutral and adaptive datasets: RDA and partial RDA), analyses of variance (ANOVAs; 1000 permutations) were performed to assess the global significance of the RDAs and marginal ANOVAs (1000 permutations) were run to determine which environmental factors were significantly correlated with genetic variation. RDAs were computed using the *rda* function available in the *vegan* package in R software. We performed an RDA and a partial RDA on all the 9770 putatively neutral SNPs and on the 28 SNPs that were identified as putatively under divergent selection by all three tests based on the PD approach (see results). The remaining SNPs identified as being under balancing selection by BAYESCAN and ARLEQUIN were not considered further.

Environmental association (EA) approach

We used three approaches to identify a set of best candidate SNPs for local adaptation. First we used the Pearson test *via* the *cor.test* function available in R software and identified all SNPs that showed statistically significant associations ($P < 0.001$ and $r > 0.70$ or $r < -0.70$) between their allele frequencies and at least one of the nine temperature parameters (called COR in Figure 24.b). Then we searched for SNP-environment associations using two environmental association programs that take into account population genetic structure: BAYESCENV (Villemereuil & Gaggiotti 2015) and LFMM v.1.4 (Frichot *et al.*

2013). BAYESCENV detects genetic signature of selection by identifying loci that show large positive F_{ST} (outside of the neutral model F_{ST} distribution) values that are significantly correlated to environmental variables. We set the neutral model of F_{ST} distribution with $P < 0.05$. LFMM uses a hierarchical Bayesian mixed model based on the residuals of PCA to take into account population genetic structure. We applied a $P < 0.05$ with a Bonferroni correction, which corresponded to SNPs that showed a z-score higher than 5 (p = 2*pnorm(-abs(z)) as recommended by Frichot *et al.* 2013). The number of genetic clusters ($k = 2$) needed for running LFMM analyses was determined by a Discriminant Analysis of Principal Component (DAPC) described in Benestan *et al*. (2015).

We defined the set of best candidate SNPs for local adaptation to temperature as those SNPs that were found to be associated with at least one of the nine temperature parameters in all three of the analyses (COR, LFMM and BAYESCENV). We then performed a spatial Principal Component Analysis (sPCA) on these best candidate SNPs with the R package *adegenet* (Jombart 2008), which accounted for spatial distribution and allowed us to assess whether variation in our best candidate SNPs was associated with environmental variables beyond what would be expected based on proximity of the different sites alone. For the sPCA, the spatial proximity network among localities was built using the neighbourhood-by-distance method based on latitude and longitude data. Then, we extracted the "locality scores" of this sPCA, which reflect genetic variability linked to spatial structure among sites, and used these to transform the genetic variation of candidate SNPs into a multi-locus geographic clines.

Gene ontology

We attempted to detect whether any of the 505 candidate SNPs (belonging to 432 candidate sequences) identified as potentially adaptive matched any of those listed in the SWISS-PROT database (Bairoch & Apweiler 2000). First, we BLASTed all 432 candidate sequences containing the 505 candidate SNPs against the complete transcriptome of the American lobster (Fraser Clark and Spencer Greenwood, University of Prince Edward Island, *personal communication*). Since the RAD candidate sequences were only 90 bp in length, this step helped increase the length of the RAD candidate sequences and hence reduced the number of false positives found when performing a BLAST search of these query sequences on a gene database. After this initial BLAST-screen, we found a total of 122 contigs

(extracted from the complete transcriptome data) that contained the 432 candidate sequences. These contigs were used as query sequences in a more stringent BLAST search conducted on the well-annotated SWISS-PROT database. Minimal *E*-value threshold of 1 x 10^{-6} and a homology of sequences of more than 70% were required for our BLAST analysis. This yielded a set of candidate SNPs successfully identified as belonging to known genes, giving in the SWISS-PROT database. The codon containing the SNP was identified in the contig sequence translated in the six reading frames. To ascertain whether a given mutation was synonymous or non-synonymous, the codon containing the SNP variants was translated into an amino acid according to the location of the start codon. Gene ontology (GO) annotation terms were then associated to the synonymous and non-synonymous candidate SNPs.

4.7. Data accessibility

DNA sequences demultiplexed with barcodes: NCBI SRA

- Bioproject Acc#: PRJNA281764
- BioSample Acc#s: SAMN03492800

The following files from this study are available from the Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.5vb8v

- *Homarus americanus*, RAD sequences for putative 13688 SNPs
- All inputs used for the PD and EA analyses
- Environmental data (spatial distribution, larval dispersal, sea surface temperature)
- All the home-made scripts used to perform the analyses

4.8. Acknowledgements

This research is part of the 'Lobster Node' of the NSERC Canadian Fisheries Research Network (CFRN). We are grateful to the fishermen of the Lobster Node without whom this project would have been impossible. Project design and work plan were developed in collaboration with scientists from the Department of Fisheries and Oceans (B. Sainte-Marie, J. Chassé, M. Comeau, J. Tremblay), representatives of fishermen associations and the facilitator of the Lobster Node, M. Allain. We would like to thank A. Boudreau, V. Brzeski, Y. Carignan, Clearwater, B. Comeau, M. Comeau, JP. Allard, M. Deraspe, N. Davis, S.

Delorey, C. Denton, R. Doucelle, J. Grignon, M. Haarr, R. MacMillan, G. Paulin and M. Thériault who helped to collect the samples. We are very grateful to T. Gosselin and P. Legendre for their help in bioinformatic and statistical analyses. The manuscript was improved by comments from A.L. Ferchaud. We are also grateful to Associate Editor Paul Bentzen and four anonymous reviewers for their constructive input on an earlier version. This research was funded by the NSERC CFRN. L. Benestan was supported by a doctoral fellowship from NSERC CFRN and Réseau Aquaculture Québec (RAQ), and funds from LB's Canadian Research Chair in Genomics and Conservation of Aquatic Resources. Larval dispersal modeling was carried out with contributions from J. Chassé (DFO) and computing resources provided by ACENET, the regional advanced research-computing consortium for universities in Atlantic Canada. ACENET is funded by the Canada Foundation for Innovation, the Atlantic Canada Opportunities Agency, and the provinces of Newfoundland & Labrador, Nova Scotia, and New Brunswick.

4.9. Tables

Table 4.1. Number of putative SNPs retained following each filtering step.

FROM READS TO SNPS	**SNP count**
STACKS CATALOG	199,664
POPULATION FILTERS	
Genotyped	
> 80% of the samples	74,512
> 70% of the populations	
MAF FILTERS	
Global MAF > 0.05	18,034
Local MAF > 0.1	
COVERAGE FILTER	
From 10 to 100x	17,831
HWE FILTERS	
F_{IS} between -0.3 and 0.3	
$H_{OBS} < 0.5$	13,688

Table 4.2. RDA and partial RDA result for each response variable ("neutral" or "adaptive" genetic variation) in relation to the explanatory variables included in the model. Significant explanatory variables are indicated with the following symbols: * P < 0.05; ** P< 0.01; *** P = 0.001.

SNPs	Analyses	Selected variables (ordistep function)			P model	R2 adj
		Environmental	Spatial	Connectivity		
9770 neutral	RDA	Maximum winter SST	dbMEM-1* dbMEM-3	AEM-1* AEM-2 AEM-4*** AEM-7*** AEM-9*	0.001	0.317
	Partial RDA		dbMEM-1**	 AEM-1** AEM-4 ** AEM-9 **	0.005 0.003	0.076 0.21
28 outliers	RDA	Minimum annual SST** Mean summer SST* Maximum annual SST*	dbMEM-1		0.004	0.301
	Partial RDA	Minimum annual SST***			0.001	0.081

* P < 0.05; ** P< 0.01; *** P = 0.001

Table 4.3. Characterization of high-quality BLAST matches.

Characterization of high-quality BLAST matches obtained in comparison of American lobster RAD-sequencing SNP against American lobster transcriptome and then against SWISSPROT database. The five SNPs that involve an amino acid change are listed as well as the one located in beta-galactosidase gene. SNPs in bold are located in genes with putative functions that are compatible with the hypothesis of thermal adaptive selection acting on encoded protein.

Uniprot ID	Program	Loci	Gene names	Species	E-value	Hit length	Amino acid change	Uniprot GO
Q3LXA3	ARLEQUIN	11427	*DAK*	*Homo sapiens*	*7.00E-85*	*573*	*ACC /GCC= Thr/Ala*	*TP-binding, FAD-AMP lyase activity, glycerone kinase activity, metal ion binding, triokinase activity, carbohydrate and fructose metabolic processes, fructose catabolic process to hydroxyacetone phosphate and phosphorylation, glycerol metabolic process, innate immune response, regulation of innate immune response*
Q62640	**LFMM**	**20131**	***Grid1***	***Rattus norvegicus***	***5.00E-10***	***119***	***CTA /ATA= Leu/Ile***	***Extracellular-glutamate-gated ion channel activity, ionotropic glutamate receptor activity and ionotropic glutamate receptor signaling pathway, ion transmembrane transport, social behavior and synaptic transmission, glutamatergic***
Q920Q4	**OUTFLANK**	**49442**	***Vps16***	***Mus musculus***	***6.00E-120***	***454***	***ACT /CCT= Thr/Pro***	***Actin biding, endosomal transport, intracellular protein transport and regulation of SNARE complex assembly, regulation of vacuole fusion, non atophagic, vacuole organization, viral entry into host cell***
Q95SX7	LFMM	21341	*RTase*	*Drosophila melanogaster*	*4.00E-14*	*248*	*AAT /AGT= Asn/Ser*	*RNA-directed DNA polymerase activity*
Q96MW7	COR	12449	*T1GD1*	*Homo sapiens*	*9.00E-08*	*108*	*GCT /GTT= Ala/Val*	*Tigger transposable element-derived protein 1*
Q81W92	**COR**	***11147***	***GLB1L2***	***Homo sapiens***	***2.00-119***	***616***	***CTT /CTA= Leu/Leu***	***Carbohydrate metabolic processes***

4.10. Figures

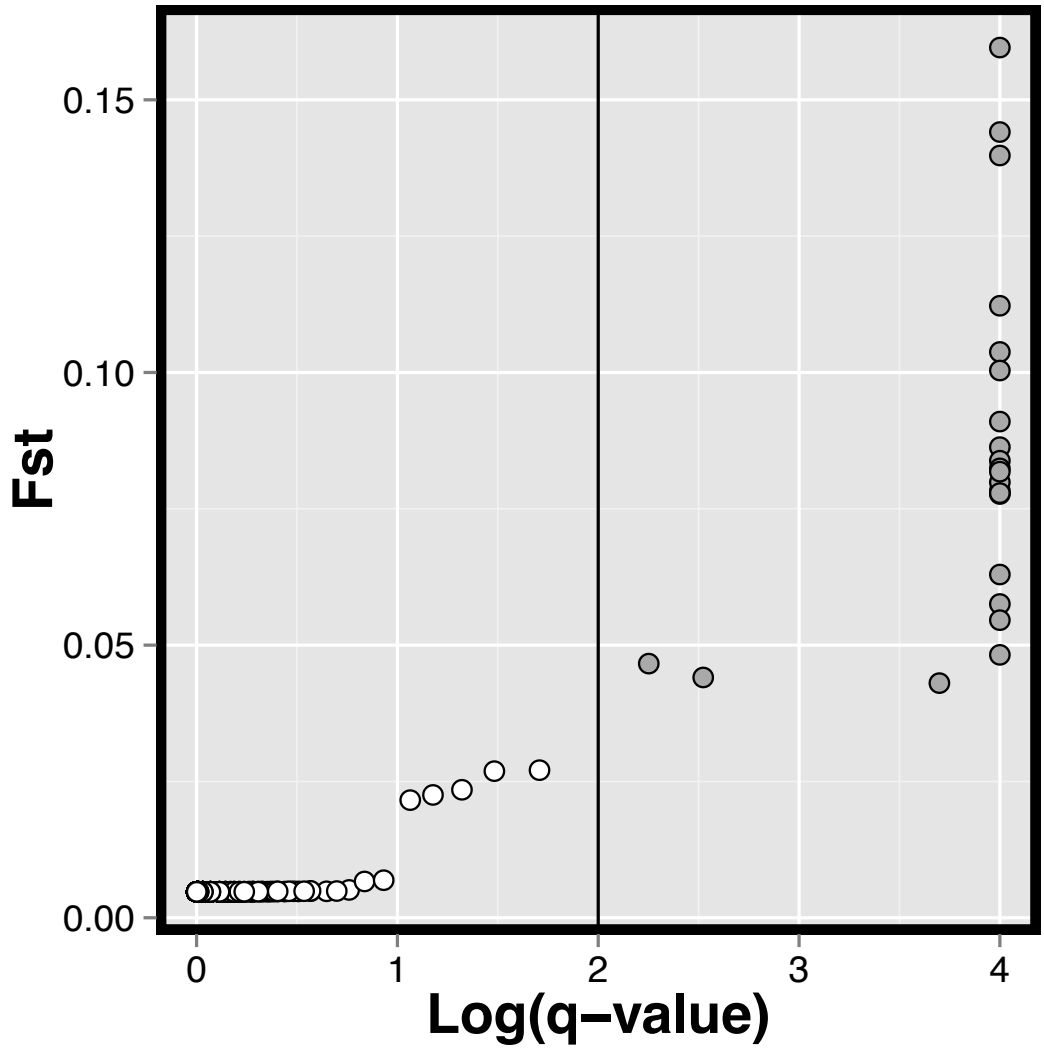


Figure 4.1. Bayesian test for selection on individual SNPs.

Bayesian test for selection on individual SNPs in BAYESCAN v. 1.21. SNPs to the right of the vertical black line represent outliers with a Bayes factor >100 (Log (Q-value) > 2).

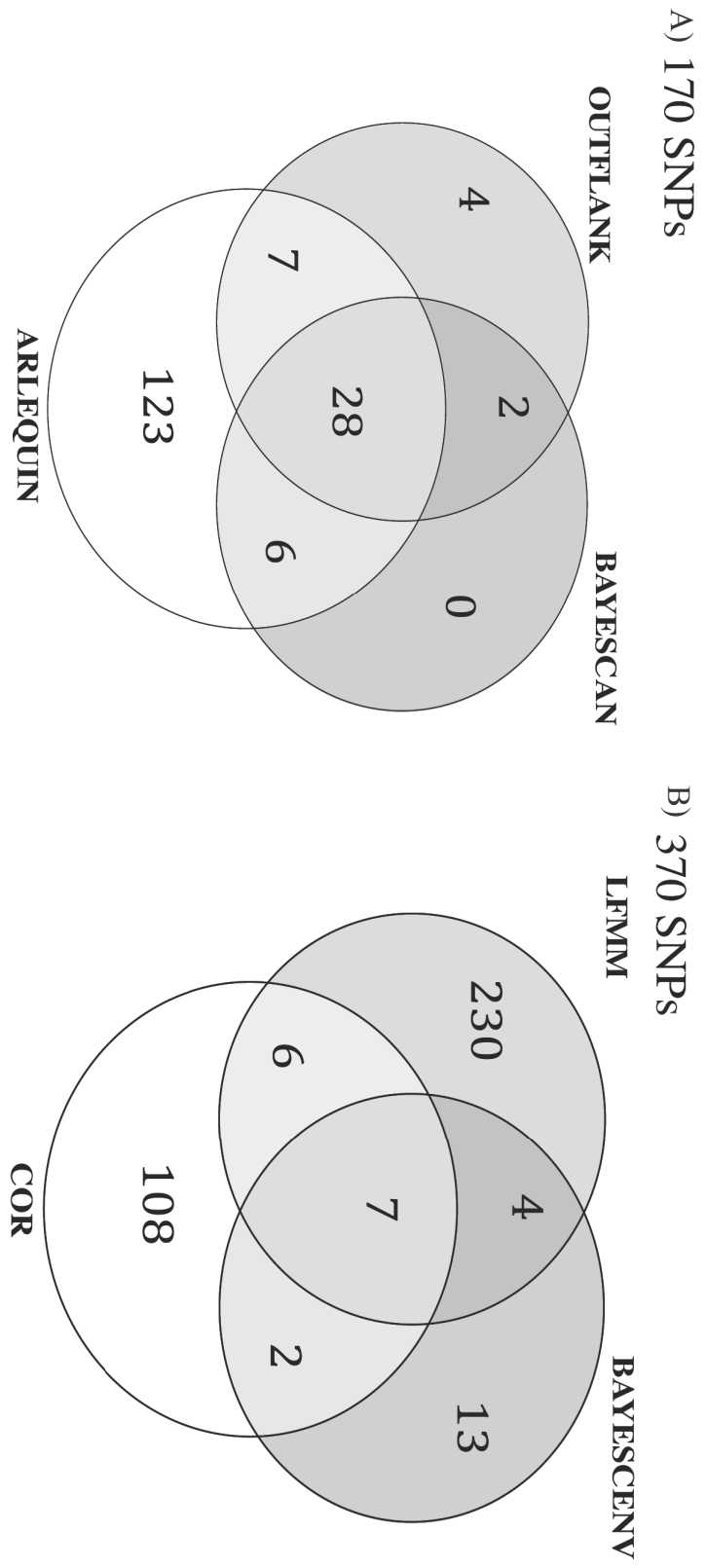


Figure 4.2. Number of SNPs identified as putatively under selection.

Number of SNPs identified as putatively under selection using A) Three genome scan methods and B) Three environment association analyses. The total number of SNPs is reported in the upper left corner of each panel.

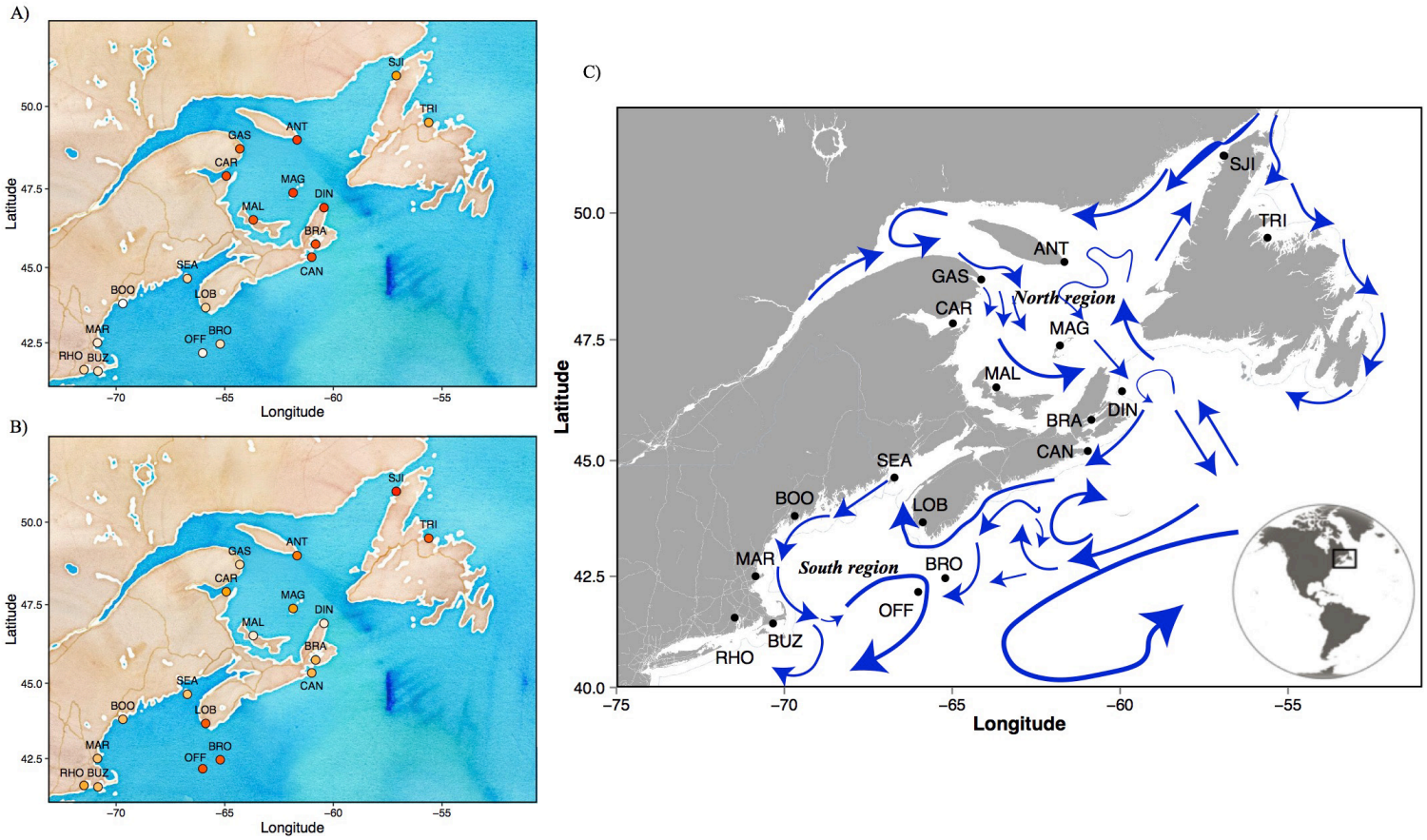


Figure 4.3. Mapping of environmental data.

Mapping of environmental data: A) Map view of the values of the first distance based Moran's eigenvecto (dbMEM-1) attributed to each site. Similarity in shading represents similarity in dbMEM-1 values. B) Map view of the values of the fourth asymmetric Eigenvector Maps vector (AEM-4), representing connectivity via larval dispersal attributed to each site. Similarity in coloring represents similarity in AEM-4values. C) Map showing the 18 sampling sites of the present study related to ocean circulation (blue arrow) along the eastern seaboard of Canada with permission of Brickman & Drozdowski 2012a.

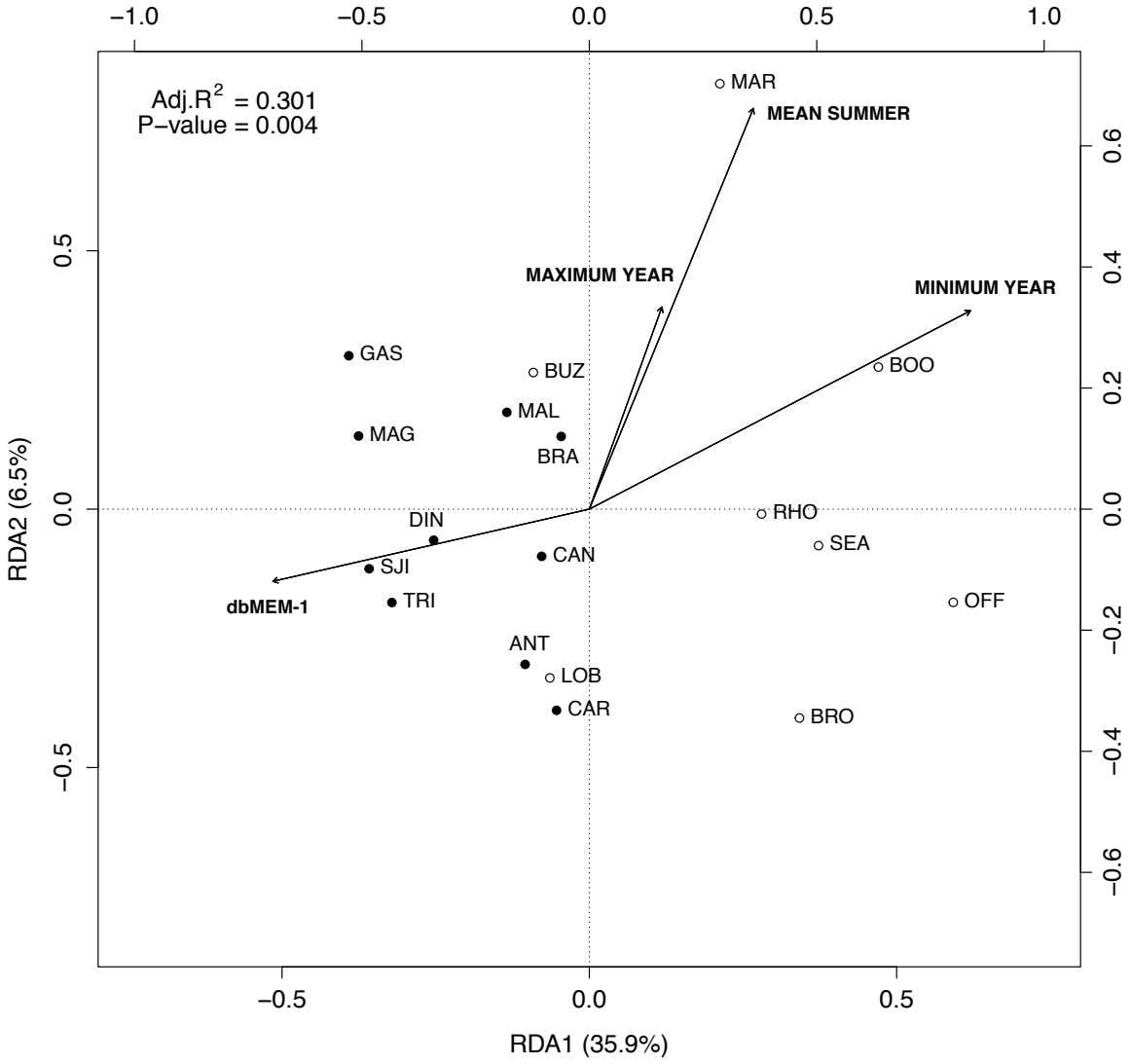


Figure 4.4. Redundancy Analysis (RDA).

Redundancy Analysis (RDA) performed on the 28 SNPs putatively under divergent selection. RDA axes 1 (35.9%) and 2 (6.5%) show American lobster from 18 localities in relation to geographic vectors (dbMEM-1) and temperature descriptors (minimum annual, maximum annual and mean sea surface temperature), which are illustrated by black arrows. Lobsters from the "south region" are in white and those from the "north region" are in black. Positions of PC factors are according to scales on top and right axes. The RDA was globally significant and revealed an adjusted coefficient of determination of 0.301.

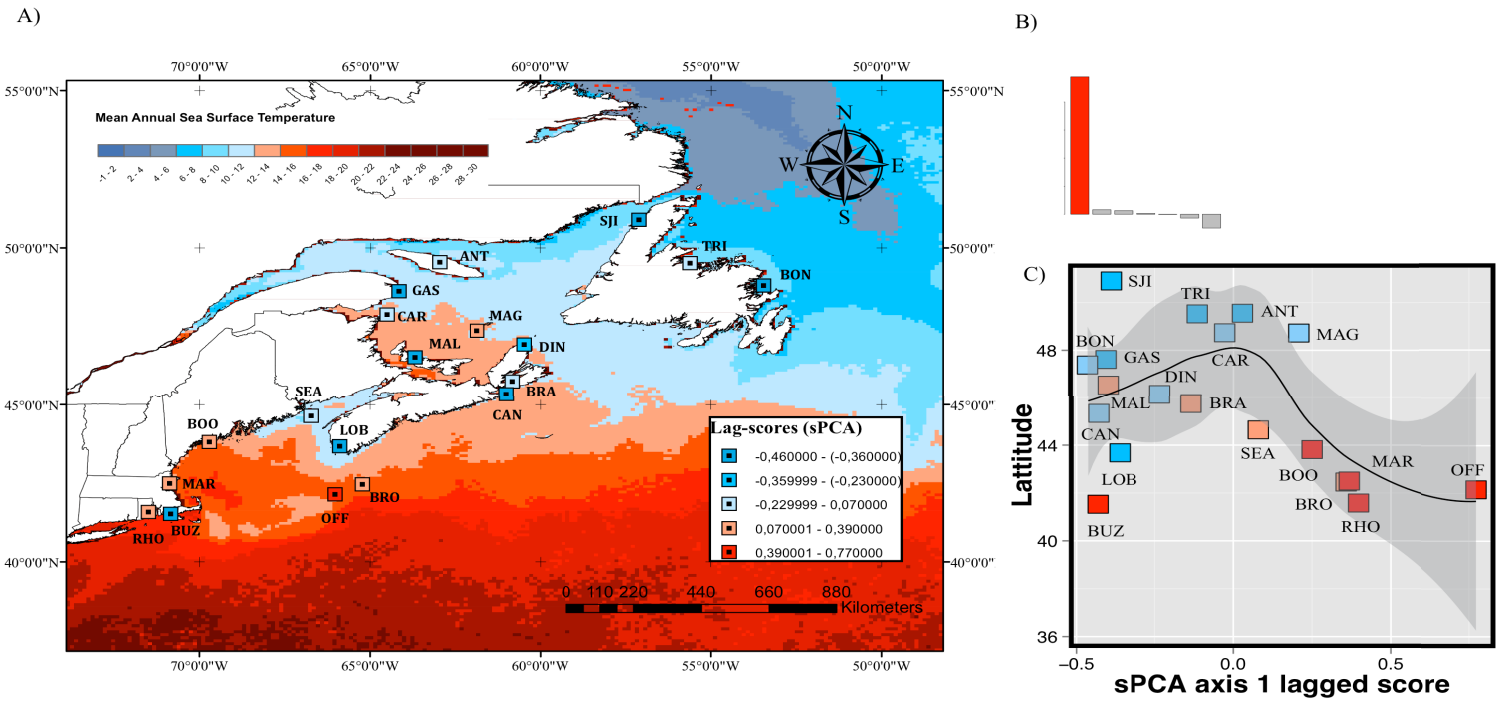


Figure 4.5. Spatial Principal Component Analysis (sPCA)

A) Synthetic multi-locus putatively adaptive variation in American lobsters from 19 sampling sites. This spatial analysis was based on genetic variation at the seven SNPs that were significantly associated with explanatory variables and detected commonly using three environmental association analyses (see Figure 4.2b, LFMM, BAYESCENV and COR). The 19 sampling sites are represented on the map by squares colored according to each locality's lagged score on the first principal component. Mean annual sea-surface temperatures, averaged from January to December 2012, are represented on the same color scale. B) Barplot of the positive and negative eigenvalues obtained when running the sPCA. Here, the first and positive eigenvalue retained is indicated in red. C) Graphical representation of the synthetic multilocus cline considering the relationship between latitude and sPCA lagged score. The gradient of colors represents the mean annual sea surface temperature (see A).

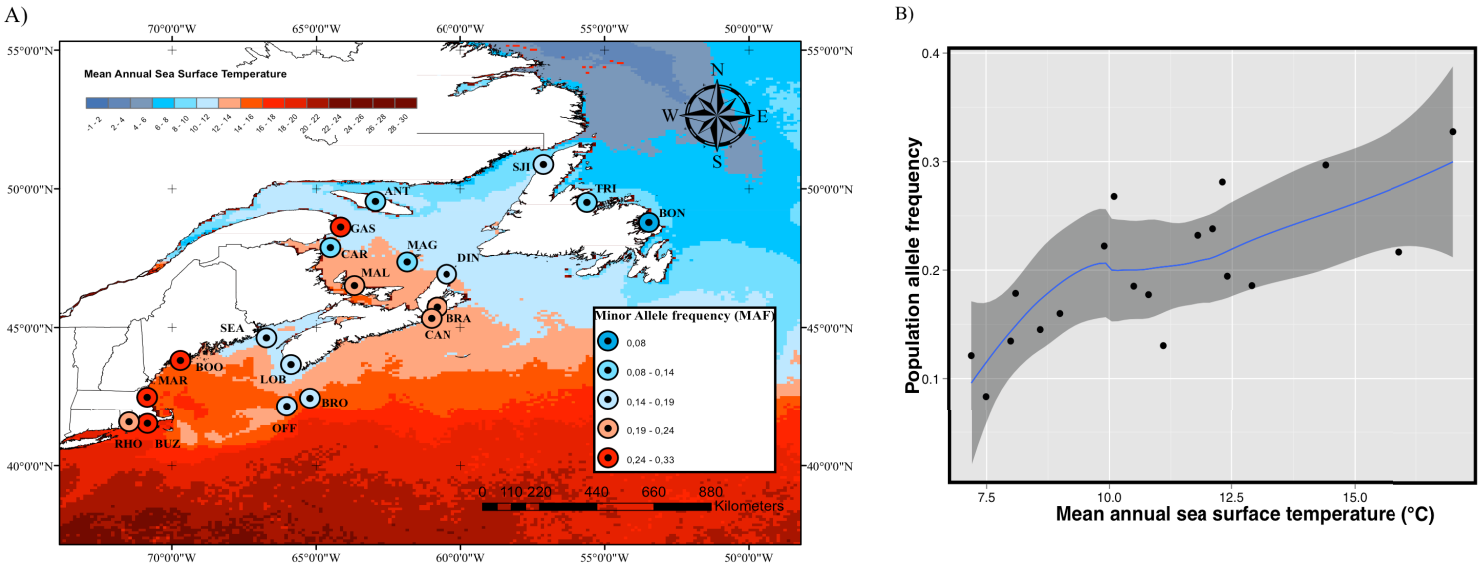


Figure 4.6. Galactosidase gene characterisation.

A) Map showing the minor allele frequency of the alternative allele (T) of the galactosidase gene (Haβ-GAL-1) at each of our 19 study sites in relation to the mean annual sea surface temperature (SST) (2012) over our study domain, and B) correlation between minor allele frequency of Haβ-GAL-1 and mean annual SST (2012), including loess smoothing function and confidence interval (grey area)

4.11. Supplementary

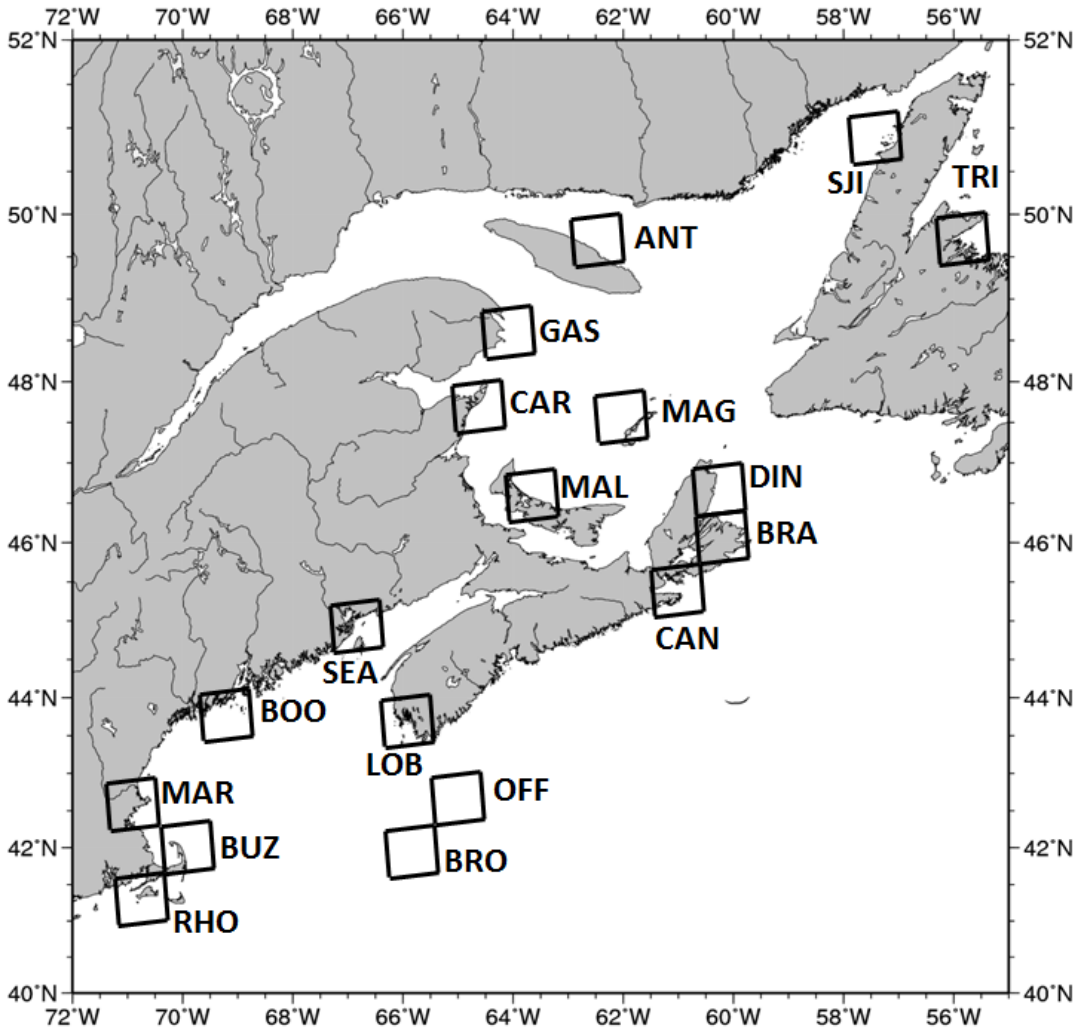


Figure S4.1. "Source-sink" areas map.

Plot of 5400 km^2 drift model "source-sink-areas" containing sampling sites used in this study. The boundaries of the map correspond to the boundaries of the model domain (Brickman & Drozdowski 2012). Note that one site (BON) is not included here because it was located outside of the model domain.

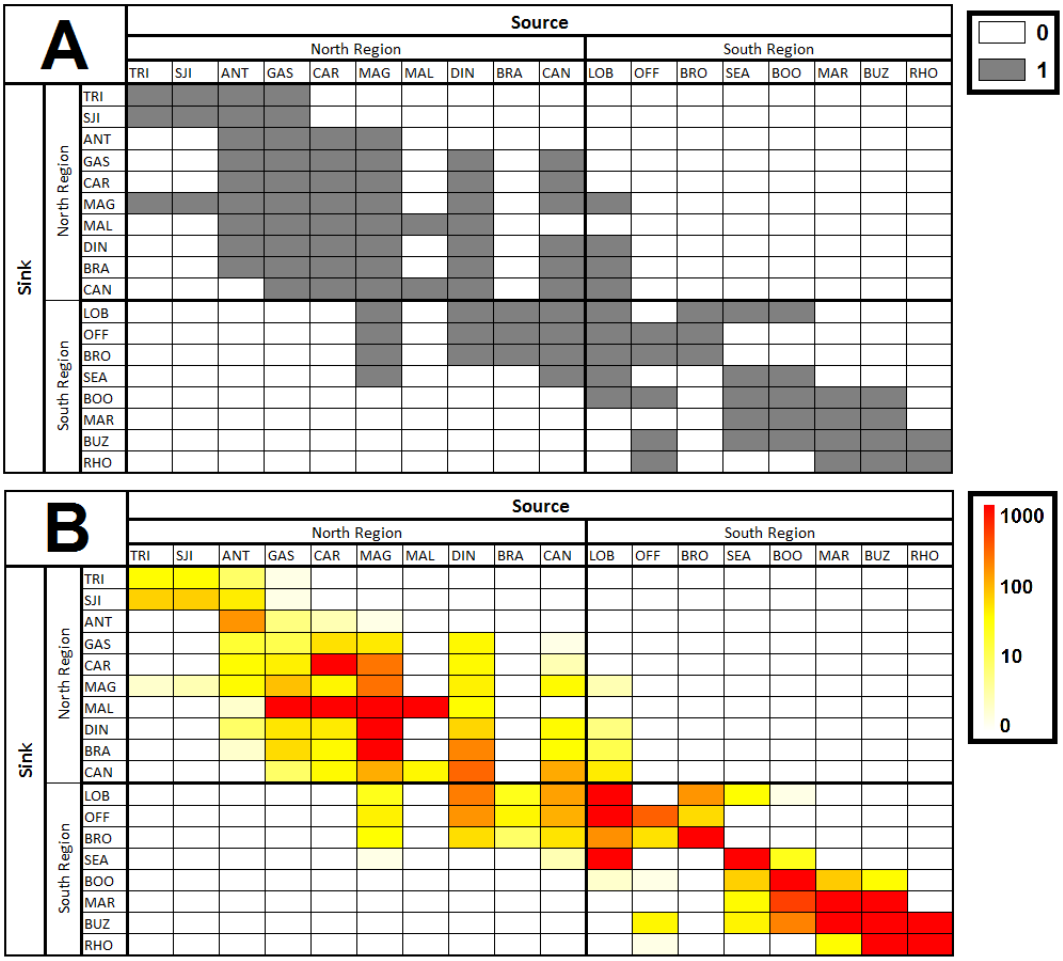


Figure S4.2. Connectivity matrices.

Connectivity matrices among the 18/19 sample sites that fell within the drift model's domain (site BON not included). (A) Model-predicted incidence of connectivity among sites (0 = not connected, 1 = connected), and (B) numbers of larvae exchanged by each pair of sites (source = larval release point, sink = place of settlement) averaged across eight years of simulations (2005-2012).

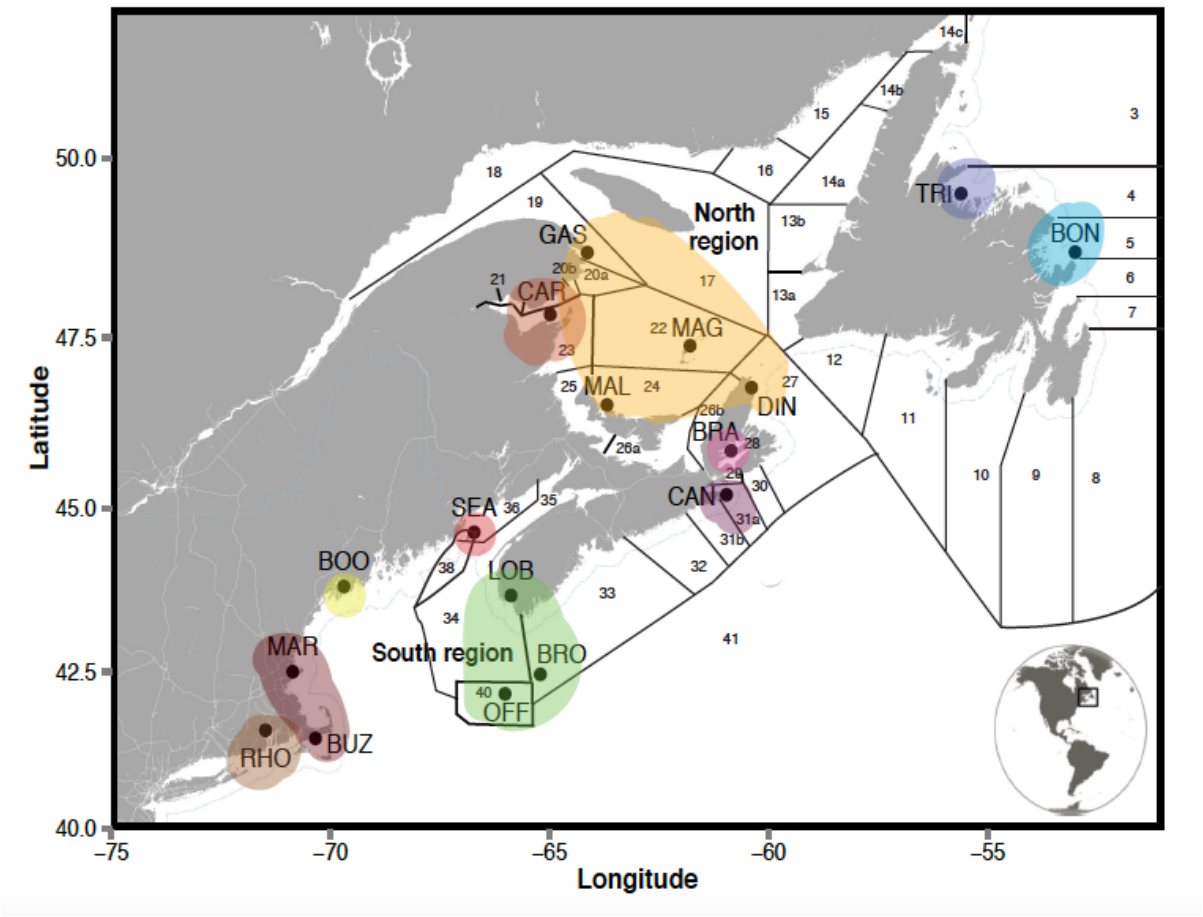


Figure S4.3. Map representing the 11 differentiated genetic populations.

Map representing the 11 differentiated genetic populations out of the 17 sampling sites analyzed by Benestan *et al.* (2015).

Chapitre 5. Sex matters: gender information is critical for unbiased population structure inferred from high-density SNP genotyping.

Soumis sous : Benestan L, Normandeau E, Rycroft N, Atema J, Rochette R and Bernatchez L. *Current Biology*.

5.1. Résumé

Les marqueurs sexuels sont confrontés à différents processus évolutifs comparativement aux marqueurs autosomiques, et peuvent donc introduire un biais dans les estimations des paramètres génétiques. Ici, nous présentons un exemple convaincant démontrant comment un échantillonnage avec un sexe-ratio déséquilibré et un génotypage incluant quelques marqueurs liés au sexe peut conduire à de fausses interprétations sur la structure de la population et ainsi mener à des recommandations erronées en terme de gestion. Ici, notre objectif initial était d'étudier l'étendue de la différenciation génétique entre deux écotypes de homards américains (*Homarus americanus*) occupant des habitats côtiers (INS) et au large (OFF). Les analyses multivariées ont révélé deux groupes génétiques qui correspondent à des individus mâles et femelles au lieu d'être liés à des groupes INS et OFF. À partir de notre ensemble de données initial, nous avons créé plusieurs sous-échantillons en faisant varier le sexe-ratio (à savoir la proportion de mâles ou de femelles sur 100 individus échantillonnés) présent dans les groupes INS et OFF, résultant dans un continuum de sex-ratio. Nous avons ensuite démontré qu'une différenciation génétique significative pouvait être observée dans un contexte de panmixie et était strictement due à un sexe-ratio non équilibré dans l'échantillonnage (sex-ratio < 0,3). Nous avons également découvert que 12 marqueurs liés au sexe étaient la cause sous-jacente de cette différenciation génétique entre les mâles et les femelles. Le retrait de ces 12 marqueurs dans les analyses a ensuite révélé une structure génétique non significative, quel que soit le sex-ratio. Pour les futures études génomiques, nous recommandons donc de collecter l'information relative au sexe des individus échantillonnés, ce qui est rarement fait, comme le démontre notre recherche exhaustive sur la littérature existant en génomique des populations marines. Pourtant, cet effort permettrait également d'acquérir plus de connaissances sur les systèmes de détermination sexuelle de nombreuses espèces non-modèles et les mécanismes moléculaires à l'origine de cette différenciation, ce qui est encore mal documenté à l'heure actuelle.

5.2. Abstract

Sexual markers face different evolutionary processes than autosomal markers and may introduce bias in genetic parameter estimations. Here, we present a compelling example of how an unbalanced sex ratio in the samples and a few sex-linked markers may lead to false interpretations of population structure and related management recommendations. Our original goal was to investigate the extent of genetic differentiation between two ecotypes of American lobsters (*Homarus americanus*) occupying inshore (INS) and offshore (OFF) habitats. Multivariate analyses revealed two genetic clusters that correspond to male and female individuals instead of being related to INS and OFF groups. From our initial dataset, we created several subsamples by varying the sex ratio (*i.e.* number of males for each female in a location) present in INS and OFF groups, resulting in a sex ratio continuum. We then demonstrated that significant genetic differentiation could be observed in this panmictic context strictly due to an unbalanced sex ratio (sex ratio < 0.3). We also discovered that 12 sex-linked markers were the underlying cause of this genetic differentiation between males and females. Removing these 12 markers led to non-significant genetic structure, regardless of an unbalanced sex ratio or not. For future genomic studies, we therefore recommend collecting sex information for each sampled individual, which is rarely done, as exemplified by an exhaustive literature search for marine species. This would also help increase our understanding of sex determination systems and their molecular mechanisms, which is still poorly documented in many non-model species.

5.4. Results

Artefactual population structure caused by skewed sex ratio

Commercial fishermen collected American lobsters from 13 sites including nine inshore sites and five offshore sites along the Atlantic coast of North America (Figure S5.1; Table S5.1). Using 1,717 filtered single nucleotide polymorphisms (SNPs), we performed a Discriminant Analysis of Principal Components (DAPC) on 203 individuals (100 males and 103 females) successfully genotyped to investigate the extent of population structuring between offshore (OFF) and inshore (INS) locations. Instead of finding significant genetic differences between INS and OFF samples, the first axis of the DAPC highlighted a significant genetic differentiation between sexes, which explained 16.04% of all genetic variation (Figure 5.1b). We then performed a DAPC on a dataset containing only males for offshore and only females for inshore locations. As expected, the DAPC showed a highly significant signal of genetic differentiation between INS and OFF samples (F_{st} = 0.0056, 95% CI_{inf} = 0.0027 and CI_{sup} = 0.0088, P-value = 0.001), which in reality resulted from the skewed sex ratio of this artificial dataset (Figure 5.1c). This outcome contrasts with the panmictic structure observed between INS and OFF (F_{st} = 0.0001, CI_{inf} = -0.0004 and CI_{sup} = 0.0006, P-value = 0.4185; Figure 5.1a) when sex ratio is balanced (sex ratio in the original dataset is equal to 0.5).

To delineate the extent to which variable sex ratio influences the genetic structure being detected, we subsampled different proportions of male and female lobsters from INS and OFF, for a total of 50 individuals. This generated a gradient of sex ratios representing different sampling bias scenarios, from the most balanced (sex ratio = 0.5) to the most unbalanced sex ratio (sex ratio = 0). First, F_{st} between INS and OFF was highest and significant when sex ratio was completely unbalanced, *i.e* sex ratio equal to 0 (F_{st} = 0.00552, CI_{inf} = 0.00300 and CI_{sup}= 0.00923, P-value < 0.05). Then, F_{st} gradually decreased with an increasingly balanced sex ratio until being no longer significant and very small (F_{st} < 0.001, CI_{inf} < 0, P-value > 0.05) when sex ratio reached 0.3 (Figure 5.2). Finally, we detected a quasi-nul and non-significant F_{st} for a sex ratio of 0.5 (F_{st} = 0.00007, CI_{inf} = -0.00100 and CI_{sup}= 0.00126, P-value > 0.05).

Sex-linked markers in American lobster and their incidence on Fst

Out of the 1,717 filtered SNPs initially considered, BAYESCAN identified 12 highly differentiated markers between the sexes (see Supplementary procedures for details). Linkage disequilibrium (LD) estimation among these markers depicted the existence of two clusters of markers in high LD (Figure 5.3). One of the clusters comprised the markers showing the strongest genetic differentiation between both sexes ($F_{st} > 0.40$; Table 5.1). Six out of these markers displayed a heterozygosity excess in males ($H_o > 0.50$) and a heterozygosity deficit in females ($H_o < 0.02$), thus providing evidence for a male heterogametic system with well-differentiated sex markers.

We investigated the incidence of the number of sex-linked markers on the index of genetic differentiation (F_{st}) calculated between INS and OFF, considering three scenarios, where sex ratio in sampling was unbalanced at different degrees (0, 0.1, 0.2; there was no effect of a reduced number of markers when sex ratio > 0.3). First, we observed high and significant F_{st} values when no sex-linked marker was removed for the three scenarios. Then, F_{st} progressively decreased with the removal of sex-linked markers (in descending order of their F_{st} values) until reaching a small and non-significant F_{st} value when we removed at least 11 out of 12 sex-linked markers.

Functional annotation of sex-specific markers and sex-linked SNP markers

Finally, we explored the identity of candidate polymorphisms involved in sexual differentiation in American lobster. From the 11 sequences containing the 12 sex-linked SNP markers, only two had a significant match (more than 90% of nucleotide identity) with the American lobster transcriptome (Fraser Clark and Spencer Greenwood, University of Prince Edward Island, *personal communication*). The polymorphisms associated to these two sequences both occurred in the 3'UTR region of unique gene IDs found in SWISSPROT database. These genes, *SULT1B1* and *cwf19*, are involved in steroid metabolism and mRNA splicing, which are known molecular pathways influencing sex determination in several fishes (Devlin & Nagahama 2002), namely the European eel (*Anguilla anguilla;* Churcher *et al.* 2015) and Greenland halibut (*Scophthalmus maximus;*Ribas *et al.* 2015).

5.3. Discussion

Sex-ratio bias in high-density SNP genotyping studies

In principle, the use of high throughput sequencing technology generates markers randomly distributed throughout the target genome (Davey *et al.* 2011). Therefore, markers linked to sexual chromosome are expected to be present in all genomic datasets developed on species with a genetic basis for sex determination (Gamble & Zarkower 2014). Despite the ubiquity of these sexual markers, almost no population genomic study on marine species has collected information on the sex of samples being analyzed (see below, Table S5.1). Yet, we clearly demonstrate that the occurrence of such markers jointly with an unbalanced sex ratio can create spurious population structure. This may in turn lead to misinterpreting the species' biology and possibly improper management recommendations. Such bias is to be particularly critical for high gene flow species typically characterized by a very weak population structuring (F_{st} < 0,01), such as marine organisms. In such cases, only a few highly differentiated markers (here 0.6% of all markers genotyped) can generate a signal of significant genetic differentiation in an otherwise totally panmictic population. These outcomes stress the need for collecting sex information of individual samples to draw accurate conclusions about population structure of non-model species using genome-wide data sets.

Sex ratio is an important characteristic of a population, which is tightly related to its dynamic (Ranta *et al.* 1999; Miller & Inouye 2013). Gaining information about the sex ratio of a population represents valuable information for an efficient and well-designed management plan, especially as sex ratio can vary widely in nature. For instance, one particular feature linked to sex-ratio is sex-biased dispersal, which is widespread in birds and mammals (Pusey 1987) but still poorly investigated in marine organisms (Burgess *et al.* 2015). Moreover, identifying sex-linked markers to define the sex of the individuals sampled may open the door for further studies documenting sex-biased dispersal as well as overcoming the influence of an unbalanced sex ratio on the analyses of genetic structure.

Addressing bias in sex ratio for population genomic studies of marine species

Marine population genomic studies have become increasingly frequent in recent years, from one study published in 2010 to a total 45 meeting our criteria and published up to

now. Among these 45 studies, only four (Galindo *et al.* 2010; Bruneaux *et al.* 2013; Johnson *et al.* 2014; Johnston *et al.* 2014; Benestan *et al.* 2015) collected information about the sex of the individuals sampled. Yet, most of these studies had a comparable sampling effort relative to the present study (N = 156 on median) as well as a relatively small number of individuals sampled per location (N per location ranging from 20 to 38 on median), and thus are also likely subject to the same bias discussed here. In the majority of these studies, the number of markers genotyped was higher than ours (8,489 SNPs on median) but since we demonstrated that only 11 sex-linked markers (0.6% of our total dataset) were sufficient to create a false signal of genetic structure, a greater number of markers is unlikely to overcome the influence of sexual markers in a low differentiated system such as that observed in the majority of marine species. Moreover, it would be expected that the number of sex-linked markers would increase proportionally with the total number of SNPs being genotyped.

Sex determination in the American lobster

In crustaceans, as in other invertebrates, sex is determined either by a male heterogamety (XX/XY) or by a female heterogamety system (ZZ/ZW). However, sex chromosomes are difficult to identify because chromosomes within Decapoda order (*e.g.* lobster, crabs) are numerous, generally very small and punctiform (Legrand *et al.* 1987; Lécher *et al.* 2011). Although markers associated to sex determination can now be easily identified within thousands of markers/sequences generated by NGS technology such as Restricted Association DNA (RAD) sequencing, as shown here (but see also Gamble & Zarkower 2014) most of the crustacean sex determining systems are poorly known and still understudied (Legrand *et al.* 1987). Here, we provide the first evidence of male heterogametic system in *Homarus americanus*, which is in agreement with one review reporting that male heterogamety was more common in Crustacea than female heterogamety, as it is for the majority of invertebrate species (Legrand *et al.* 1987). In addition, we demonstrated the possibility to efficiently uncover the sex chromosome system of a non-model species using a genome-wide dataset.

Candidate genes for sexual differentiation

We identified two candidate genes for sexual differentiation, *SULT1B1* involved in steroid metabolism, and *cwf19* gene that acts on pre-RNA splicing. Steroids play important roles in regulating physiological functions related to reproduction and sex differentiation in

fishes (James 2011). More broadly, several publications already pointed out that sulfotransferase genes, such as *SULT1*, were linked to sex determination in rats (*Mus musculus*; Dunn *et al.* 1999), mussels (*Mytilus galloprovincialis*; Atasaral Şahin *et al.* 2015), European eel (*Anguilla anguilla;* Churcher *et al.* 2015) and turbot (*Scophthalmus maximus;* Ribas *et al.* 2015). For instance, sulfotransferase 6B1-like gene (*SULT6B1*) was expressed at higher levels in the liver of sexually mature European eel males, suggesting that this gene may be associated with pheromonal communication during the reproduction of this species (Churcher *et al.* 2015). In addition, one sulfotransferase gene (*hs3st1l2*) was detected as a potential candidate gene for sex determination in turbot (*Scophthalmus maximus*), as it is associated with differential expression between sexes at sexual maturity (Ribas *et al.* 2015). Interestingly, this study also identified *cwf19* gene as a putative sex determining gene in the turbot (Ribas *et al.* 2015).

Remarkably, both candidate polymorphisms occurred in the 3'UTR region of *SULT1B1* and *cwf19* gene. The 3'UTR regions plays an important role in post-transcriptional control of gene expression, and thus may affect the level of protein being expressed (Hesketh 2004). Several studies have already pointed out that polymorphisms in 3'UTR region modulate the level of transcription of downstream genes (Barrett *et al.* 2012). Here, polymorphisms found in *SULT1B1* and *cwf19* gene could potentially affect their transcription, which makes sense in the light of previous work on eel and turbot that reported differences in the transcription profile for these genes between males and females (Churcher *et al.* 2015; Ribas *et al.* 2015). These two studies showed that *cwf19* gene was down regulated in turbot female with a fold change of -1.7, whereas *SULT6B1* was expressed at higher levels, with a fold change of 7.8, in the livers of sexually mature males. Admittedly, we stress that the functional annotation for these two genes in American lobster is for the moment hypothetical but clearly deserves further scrutiny.

5.4. Methods

Molecular techniques

Genomic DNA was extracted using Qiagen Blood and Tissue kits following the kit protocol. DNA quality was confirmed using visual inspection on 1% agarose gel followed by quantification with Quantit Picogreen dsDNA assay kits. RAD-sequencing libraries were prepared following the protocol from Benestan *et al.* (2015). Each individual was barcoded with a unique six-nucleotide sequence and 48 individuals were pooled per lane. Real-time PCR was used to quantify the libraries. Single read, 100 bp target length, sequencing was performed on an Illumina HiSeq2000 platform at the Genome Quebec Innovation Centre (McGill University, Montréal, Canada).

Bioinformatics and genotyping

The libraries were de-multiplexed using the *process_radtags* program in STACKS v.1.29 (Catchen *et al.* 2013). Raw sequencing data was checked in FASTQC (Andrews 2015). Reads were truncated to 80 bp and adapter sequences were removed with CUTADAPT in order to obtain reads with the same length. The formation of RAD loci was allowed with a maximum of three nucleotide mismatches (M = 3), according to Ilut *et al.* (2014) and a minimum stacks depth of three (m = 3), among reads with potentially variable sequences (*ustacks* module in stacks, with default parameters). Then, reads were clustered *de novo* with each other to create a catalogue of putative RAD tags (*cstacks* module in stacks, with default parameters). In the *populations* module of STACKS v.1.29 and following consecutive filtering steps, we first retained SNPs genotyped in at least 80% of the individuals found in at least 9 of the 12 "populations" (Table S5.2). Potential homeologs were excluded by removing markers showing heterozygosity > 0.50, $F_{IS} < -0.30$ $F_{IS} > 0.30$ within samples. Only SNPs with a minor allele frequency > 0.02 were retained for the analysis. The resulting filtered VCF files were converted into the file formats necessary for the following analyses using PGDspider v.2.0.5.0 (Lischer & Excoffier 2012).

Discriminant Analysis of Principal Component (DAPC)

We performed a Discriminant Analysis of Principal Components (DAPC) in the R package *adegenet* (Jombart *et al.* 2010) using prior information on group of origin (offshore

and inshore). Then, we evaluated the optimal number of discriminant functions (n=60) to retain according to the optimal α-score obtained from our data (Jombart *et al.* 2010).

Genome scan and LD calculation

We searched for outlier loci with unusually high level of divergence between sexes using an F_{ST}-based outlier analysis. We detected outlier SNPs with BAYESCAN v. 2.1 (Foll & Gaggiotti 2008), a Bayesian method based on a logistic regression model that separates locus-specific effects from population-specific effects of demography. BAYESCAN runs were implemented using prior model (pr_odds) of 10,000, as recommended by Lotterhos & Whitlock (2015), including a total of 10,000 iterations and burn-in of 200,000 steps. These outlier analyses were conducted on the entire data set divided according to sex information.

We calculated Linkage Desequilibrium (LD) between pairs of SNPs using the *geno-r2* command available in VCFTOOLS. We then transformed the LD dataframe obtained with VCFTOOLS into a suitable LD matrix ready to analyze with the *heatmap* command accessible in R environment.

Markers annotation

First, we blasted the 12 candidate SNPs against the complete transcriptome of the American lobster (Fraser Clark and Spencer Greenwood, University of Prince Edward Island, *personal communication*). Nine of the 12 candidate SNPs were found to belong to a contig extracted from the complete transcriptome data. We used this set of contigs as query sequences in a blast search conducted on SWISS-PROT database (Bairoch & Apweiler 2000). Minimal *E*-value threshold of 1×10^{-6} and a homology of sequences of more than 70% were required for our blast analysis. This yielded a set of candidate SNPs successfully identified as belonging to known genes. Gene ontology (GO) annotation terms were then associated to the candidate SNPs.

Literature search

We conducted a literature search of marine genomic studies published in peer-reviewed journals through July 2016 using the Web of Knowledge bibliographic database. A search of Web of Science for the key words (i) “genomics” AND “marine” AND “SNP” yielded 17 hits, (ii) “population structure” AND “marine” AND “SNP” yielded 37 hits, (iii) “RAD sequencing” AND “marine” yielded 31 hits and (iv) “population genomics” AND

"marine" yielded 165 hits. From these hits, several criteria were used to determine which studies to include in our analyses. First, the paper had to focus on an animal marine species (vertebrate or invertebrate) and should use a set of SNPs markers > 1000. Secondly, the paper had to refer to population genomics or related areas such as outlier identification because these are the target areas of research that are likely prone to be influenced by the sex ratio bias in sampling. After removing non-animal marine and non-genomic hits, we ended up with a total of 45 publications listed in Table S5.1.

5.4. Acknowledgements

We are grateful to the fishermen without whom this project would have been impossible. We would like to thank B. Sutherland and J.S Moore for the constructive discussions on the topic. The « Lobster Node » of the NSERC CFRN funded this research. L. Benestan was supported by a doctoral fellowship from NSERC CFRN and Réseau Aquaculture Québec (RAQ), and funds from LB's Canadian Research Chair in Genomics and Conservation of Aquatic Resources.

5.5 Tables

Table 5.1. Genetic information of 12 sex-linked markers.

Observed heterozygosity (H_o), expected heterozygosity (H_E), Wright's coefficient of inbreeding (F_{is}) tested for Hardy-Weinberg equilibrium (P-value) and genetic differentiation index (F_{st}) between sexes (females, n=100; males, n=103) for 12 highly sex-linked markers identified with BAYESCAN. All markers showed a significant F_{st} between sexes (P-value < 0.002). Markers showing the strongest genetic differentiation between sexes and belonging to the same LD cluster (see Figure 5.5) are in bold characters.

	Females				**Males**				**F_{st}**
Marker	H_o	H_e	F_{is}	P-value	H_o	H_s	F_{is}	P-value	
1951841	**0.010**	**0.010**	**0.000**	**1.000**	**0.605**	**0.496**	**-0.220**	**0.027**	**0.560**
3534313	**0.000**	**0.000**	**--**	**--**	**0.634**	**0.498**	**-0.273**	**0.008**	**0.543**
2341697	**0.291**	**0.504**	**0.423**	**0.002**	**0.311**	**0.383**	**0.188**	**0.056**	**0.514**
703660	**0.011**	**0.011**	**0.000**	**1.000**	**0.628**	**0.501**	**-0.253**	**0.013**	**0.470**
1713801	**0.000**	**0.021**	**1.000**	**0.006**	**0.615**	**0.499**	**-0.231**	**0.029**	**0.440**
2033018	**0.011**	**0.032**	**0.664**	**0.020**	**0.563**	**0.498**	**-0.130**	**0.162**	**0.425**
2879520	**0.011**	**0.032**	**0.664**	**0.022**	**0.524**	**0.493**	**-0.064**	**0.371**	**0.401**
434792	0.021	0.041	0.493	0.039	0.484	0.389	-0.244	0.017	0.214
1757708	0.280	0.415	0.326	0.003	0.500	0.485	-0.031	0.462	0.166
1525333	0.261	0.496	0.473	0.001	0.323	0.411	0.215	0.033	0.141
2341745	0.000	0.044	1.000	0.001	0.591	0.496	-0.192	0.052	0.108
794307	0.156	0.373	0.581	0.001	0.156	0.162	0.037	0.525	0.077

5.6. Figures

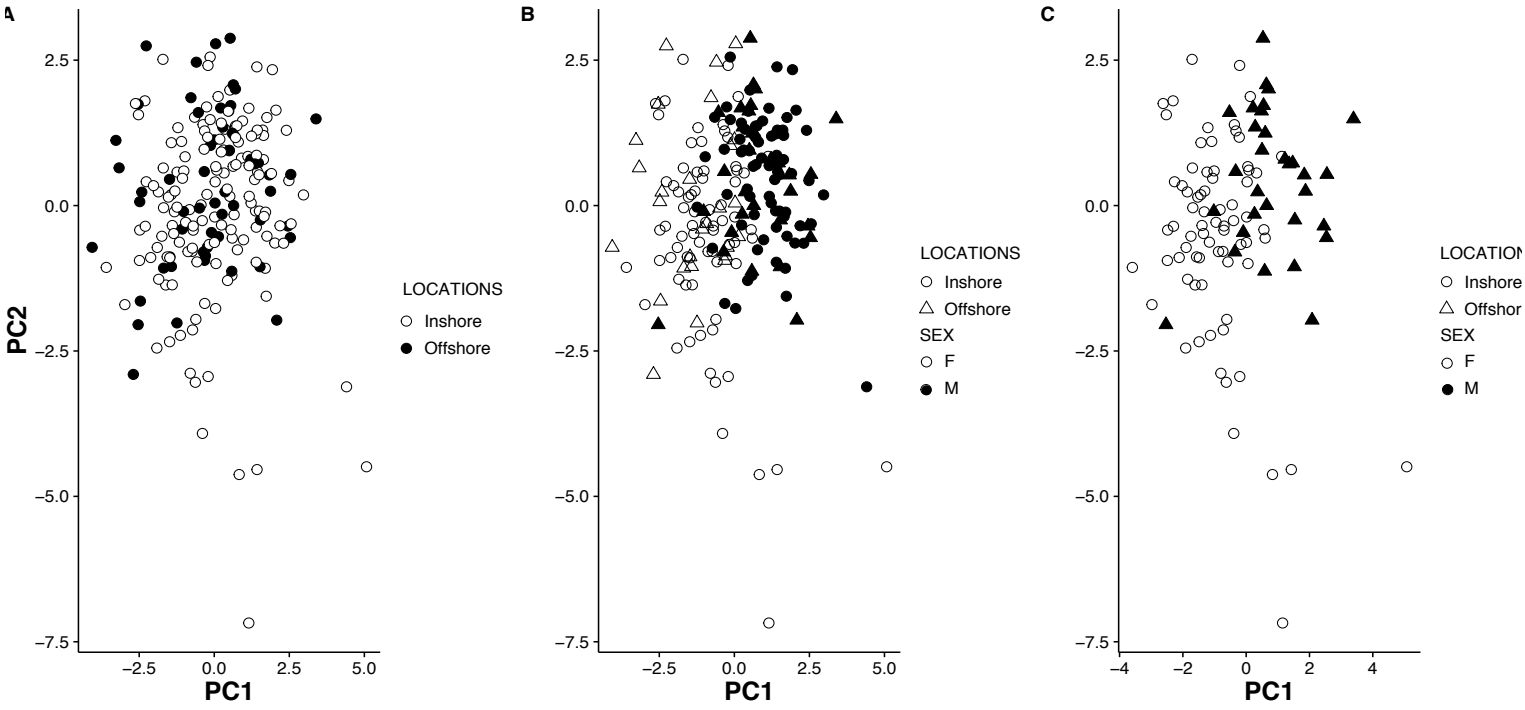


Figure 5.1. Discriminant analysis of principal components (DAPC)

Discriminant analysis of principal components (DAPC) of genetic differentiation depending on the sampling scenario. For each sampling scenario, we performed a DAPC based on 1,717 single nucleotide polymorphic markers (each point represents one individual). (A) For the first scenario, we tested the evidence of a genetic structure between offshore (black symbols) and inshore (white symbols) locations. Individuals from the inshore and offshore regions are represented by white and black symbols, respectively. (B) Then, we analyzed our results regarding the offshore/inshore (triangles/circles) clustering and female/male (white/black) information. (C) Finally, we subsampled only males in offshore locations and only females in inshore locations and we highlighted the existence of a genetic structure due strictly to a skewed sex ratio in sampling.

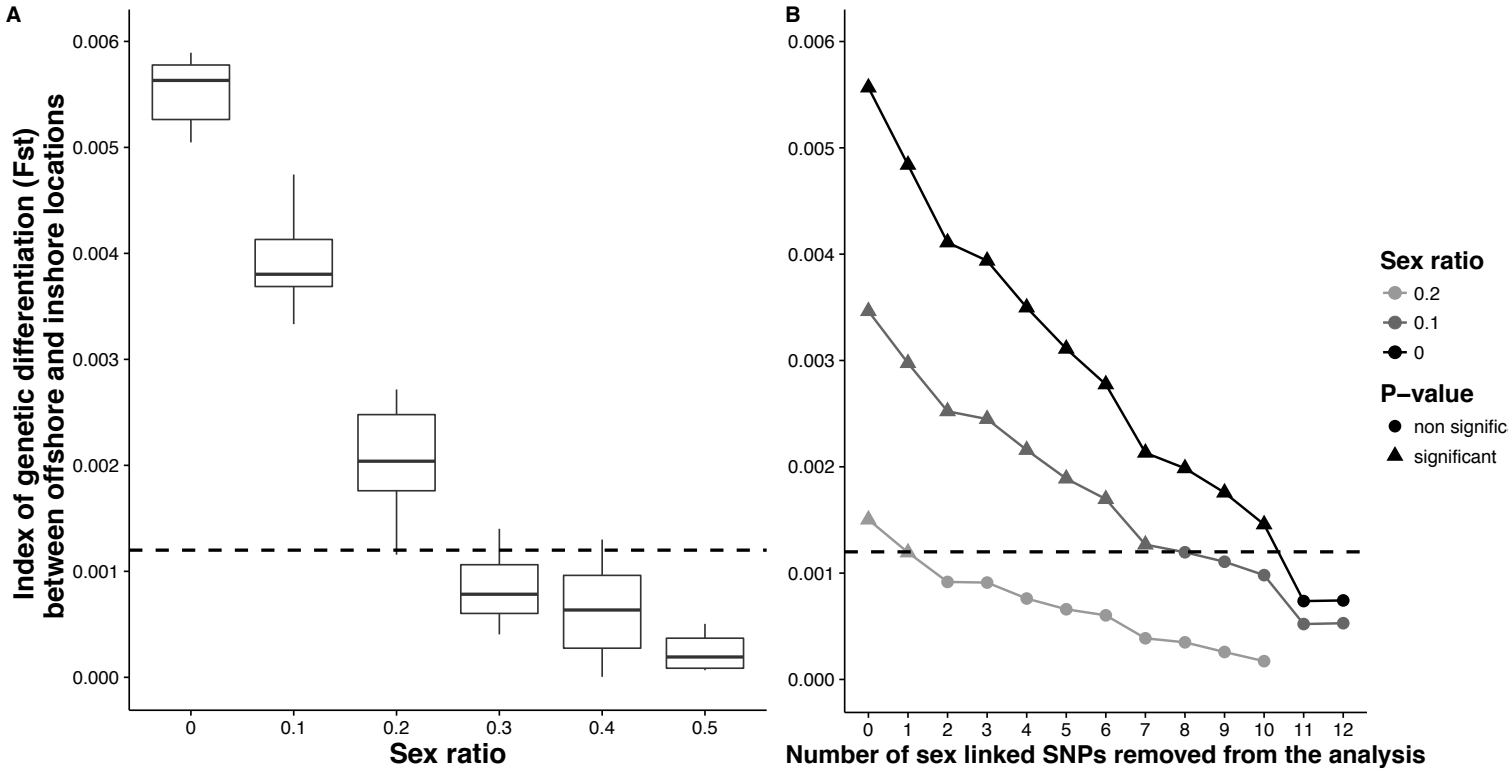


Figure 5.2. Boxplot and line graph showing the influence of sampling sex ratio and sex-linked SNPs on the index of genetic differentiation (Fst) between inshore and offshore locations.

(A) The boxplot represents the index of genetic differentiation (F_{st}) between offshore and inshore (y-axis) lobsters in subsamples of 100 individuals with varying sex ratio. The first scenario of sampling corresponds to a sex ratio equal to 0, meaning that no males were sampled in offshore whereas 50 female were sampled in inshore and vice-versa. We tested six scenarios ranging from a complete unbalanced sex ratio (*i.e.* sex ratio equal to 0) to a perfectly balanced sex ratio (*i.e.* sex ratio equal to 0.5). The vertical limits of the box represent one standard deviation around the mean (n = 10 subsamples), the horizontal line within the box is the median, and the whiskers extend to the 25th and 75th percentiles. The dashed line in black indicates the threshold below which F_{st} values are no longer significant at $P < 0.05$.

(B) The line graph displays the index of genetic differentiation (F_{st}) as a function of the number of sex-linked markers removed from the analysis considering three sampling scenario with different degrees of sex ratio bias (0, 0.1, 0.2). Sex-linked markers are removed in descending order according to their F_{st} values (see Table 5.1). The dashed line in black indicates the threshold below which F_{st} values are no longer

significant at $P < 0.05$. Sex ratio of 0.3, 0.4 and 0.5 were not included in this analysis because F_{st} values were not significant in these cases (see A)

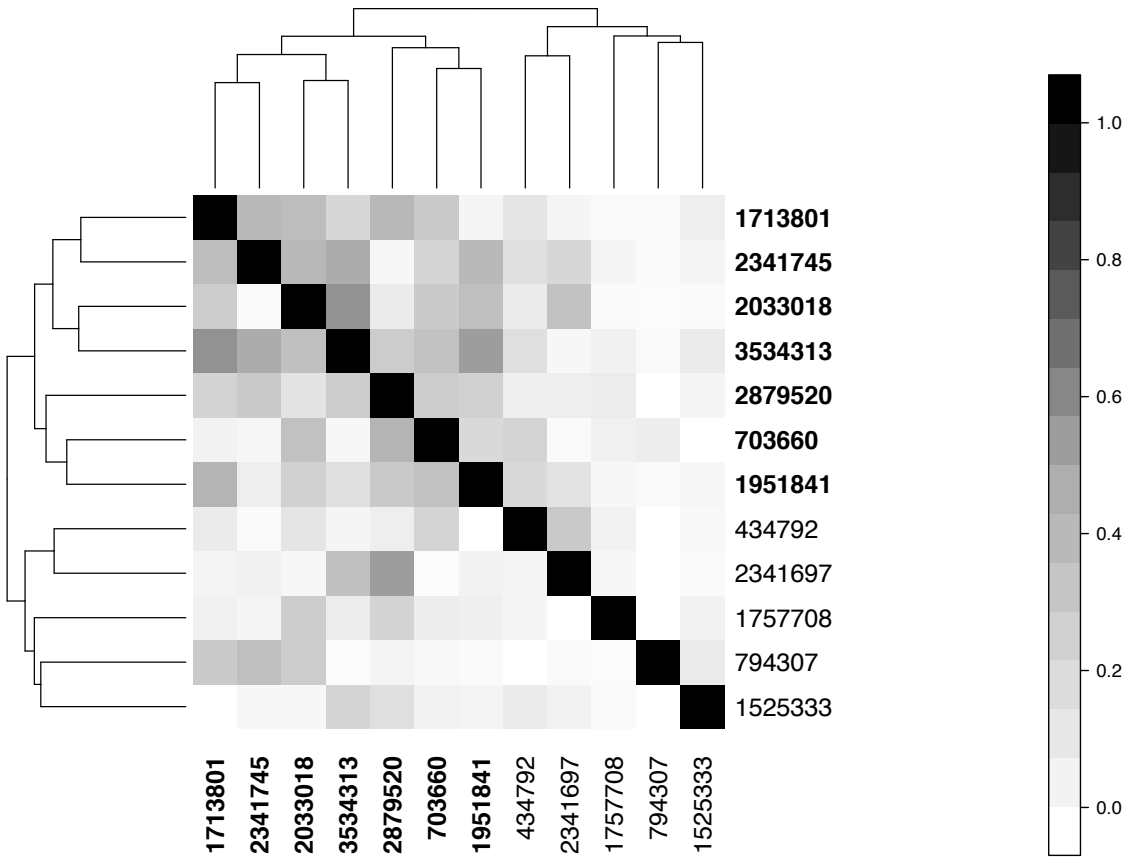


Figure 5.3. Heatmap of the Linkage Disequilibrium (LD)

Heatmap illustrating the linkage disequilibrium (LD) for the 12 highly sex-differentiated markers, considering all the males and females sampled in inshore (INS) and offshore (OFF) locations. Each row and column represents a specific SNP. The shades represent different ranges of LD values, from low (pale grey, 0.0) to high (in black, 1.0). The gene trees shown on the heatmap, based on LD values, suggest two LD clusters. The SNPs clustering in the LD cluster that is the most strongly linked to sex determination are shown in bold.

5.7. Supplementary materials

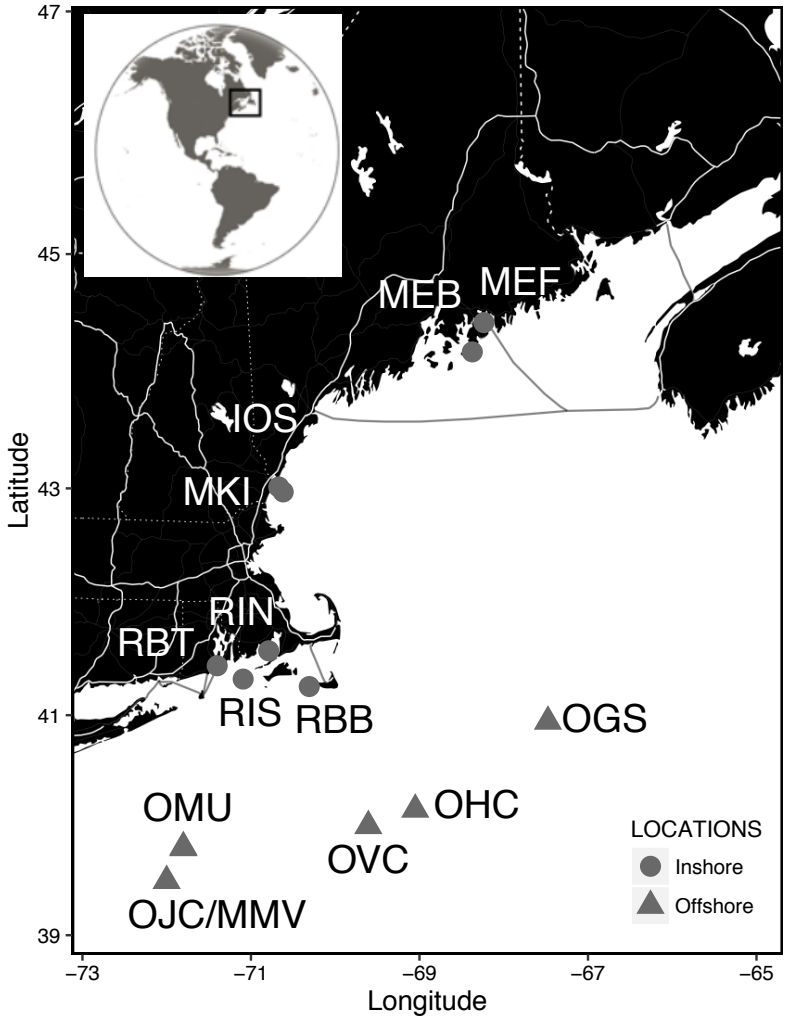


Figure S5.1. Sampling locations

Offshore sampling locations (OFF) are shown in black labels with a grey triangle and inshore locations (INS) in white labels with a grey circle. Offshore locations are Georges Basin (OGS; n=10), Hydrographers Canyon (OHC; n=16), Jones Canyon (OJC; n=12), MacMaster Canyon (OMU; n=10) and Veatch Canyon (OVC; n=16). Inshore locations are Isle of Shoals (IOS; n=14), Blue Hill Bay (MEB; n=20), Frenchmans Bay (MEF; n=19), Kittery (MKI; n=20), Brown's bank (RBB; n=17), Beavertail (RBT; n=16), Narragansett Bay (RIN; n=13), Rhode Island Sound Bay (RIS; n=7).

Table S5.1. Marine population genomics studies.

Marine population genomics studies that focus on population differentiation and/or outlier identification. We indicated the name of the authors (Study), the species studied (Organism), the method used (Method), the number of individuals sampled, the number of SNPs genotyped (SNPs), if the authors collected and used gender information for their analyses (Sex), the number of samples per population (N/per population) and the overall population differentiation index the authors estimated from their dataset (Fst).

Study	Organism	Method	N	SNPs	Sex	N/per population	Fst
Benestan *et al.* 2015	*Homarus americanus*	RAD sequencing	586	10,156	Yes	30 to 36	0.0018
Berg *et al.* 2015	*Gadus morhua*	SNP-arrays	194	8,809	No	8 to 48	0.0002 to 0.0709
Berg *et al.* 2016	*Gadus morhua*	SNP-arrays	141	8,168	No	42 to 51	0.00123 to 0.000861
Boehm *et al.* 2015	*Hippocampus erectus*	RAD sequencing	23	11,708	No	5 to 9	0.0454 to 0.1012
Bourret *et al.* 2013	*Salmo salar*	SNP-arrays	1,431	6,176	No	20 to 72	0.025 to 0.758
Bruneaux *et al.* 2013	*Gasterosteus aculeatus*	RAD sequencing	288	6,834	Yes	48	Unknown
Cammen *et al.* 2015	*Tursiops truncatus*	RAD sequencing	156	7,431	No	12 to 26	Unknown
Candy *et al.* 2015	*Thaleichthys pacificus*	RAD sequencing	494	4,104	No	22 to 71	0.000 to 0.0128
Chu *et al.* 2014	*Nucella lapillus*	RAD sequencing	30	4,000	No	Unknown	0.0004 to 0.0474
Corander *et al.* 2013	*Clupea harengus*	RAD sequencing	2	4,756	No	6	0.005
Ferchaud *et al.* 2014	*Gasterosteus aculeatus*	RAD sequencing	60	33,993	No	20	0.056 to 0.111

Ferchaud *et al.* 2016	*Gasterosteus aculeatus*	RAD sequencing	177	28,888	No	20	0.002 to 0.458
Galindo *et al.* 2010	*Littorina saxatilis*	454 sequencing	30	2,454	Yes	15	0.03
Guo *et al.* 2015	*Gasterosteus aculeatus*	RAD sequencing	10 pools (360)	30,871	No	36	0.02825
Hess *et al*. 2013	*Entosphenus tridentatus*	RAD sequencing	518	4,439	No	4 to 35	0.021
Holenhole *et al*. 2010	*Gasterosteus aculeatus*	RAD sequencing	100	45,000	No	20	0.00203 to 0.1391
Jackson *et al*. 2014	*Epinephelus striatus*	RAD sequencing	620	4,234	No	14 to 32	0.002
Jacobsen *et al.* 2014	*Anguilla anguilla; Anguilla rostrata*	RAD sequencing	60	328,300	No	8 to 15	0.041
Johnston et al 2014	*Salmo salar*	SNP-arrays	503	4353	Yes	49 to 260	0.0103
Lal *et al.* 2015	*Pinctada margaritifera*	RAD sequencing	156	5,243	No	32-50	0.046
Lamichnaey et al. 2012	*Clupea harengus*	Transcriptome assembly	400	440,817	No	50	Unknown
Laporte *et al.* 2016	*Anguilla anguilla; Anguilla rostrata*	RAD sequencing	179	14,755	No	21 to 24	0.000 to 0.001
Larson *et al.* 2014	*Oncorhynchus tshawytscha*	RAD sequencing	270	10,944	No	47 to 56	0.003 to 0.098
Moan *et al.* 2016	*Engraulis encrasicolus*	RAD sequencing	128	5,638	No	24 to 64	Unknown
Moore *et al.* 2015	*Salmo salar*	SNP-arrays	9,142	3,192	No	9 to 100	0.043
Moura *et al.* 2014	*Orcinus orca*	RAD sequencing	115	3281	No	6 to 21	0.0346 to 0.334

O Brieuc *et al.* 2015	*Oncorhynchus tshawytscha*	RAD sequencing	414	9,107	No	21 to 41	0.000 to 0.33
Pavey *et al.* 2015	*Anguilla rostrata*	RAD sequencing	379	42,424	No	21 to 24	< 0.001
Pecoraro C. *et al.* 2016	*Thunnus albacares*	RAD sequencing	100	6,772	No	10	0.0273
Picq *et al*. 2016	*Hypoplectrus spp*	RAD sequencing	126	97,962	No	13 to 43	0.0042
Pocwierz-Kotus *et al.* 2015	*Gadus morhua*	SNP-arrays	95	7,944	No	26 to 40	0.034
Pujolar *et al.* 2014	*Anguilla anguilla*	RAD sequencing	259	50,354	No	30 to 37	< 0.001
Reitzel *et al.* 2013	*Nematostella vectensis*	RAD sequencing	30	2,759	No	4 to 7	0.286 to 0.622
Rodríguez-Ezpelet *et al.* 2016	*Scomber scombrus*	RAD sequencing	122	29,394	No	15 to 29	0.0157 to 0.039
Rougemont *et al.* 2016	*Lampetra planeri; Lampetra fluviatilis*	RAD sequencing	338	8,962	No	29 to 53	0.042 to 0.207
Sodeland et al 2016	*Gadus morhua*	SNP-arrays	378	9,187	No	43 to 48	0.000 to 0.0189
Stockwell *et al.* 2015	*Scarus niger*	RAD sequencing	81	4,253	No	24 to 30	0.007
Xu *et al.* 2016	*Bathymodiolus platifrons*	RAD sequencing	28	9,307	No	10 to 18	0.0126
Zhang *et al.* 2016	*Larimichthys polyactis*	RAD sequencing	24	27,556	No	12	< 0.001
Bradbury *et al.* 2010	*Gadus morhua*	EST library	300	1,641	No	15 to 26	Unknown

Jones *et al.* 2012	*Gasterosteus aculeatus*	SNP-arrays	121	1,159	No	4 o 6	0.031 to 0.383
Therkildsen *et al.* 20	*Clupea harengus*	EST library	508	1,047	No	14 to 37	0.000 to 0.086
Teplolt *et al. 2015*	*Carcinus maenas*	EST library	84	10 809	No	12	0.003 to 0.134
Bay & Palumbi 2014	*Acropora hyacinthus*	EST library	23	15,399	No	10 to 13	Unknown
De Wit & Palumbi 2013	*Haliotis rufescens*	EST library	26	21,579	No	1 to 13	0.0003

Table S5.2. Information on the sampling.

Information on the sampling: code, group, location, sample date, latitude, longitude and number of individuals successfully genotyped (N_{GEN}). Samples were taken by Atema and Gerlach (*unpublished*).

Code	Group	Location	Sample date	Lattitude	Longitude	N_{GEN}
MEF	Inshore	Frenchmans Bay, ME	2007-2008	44.421786°	-68.235314°	19
MEB	Inshore	Blue Hill Bay, ME	2008	44.172878°	-68.372556°	20
MKI	Inshore	Kittery, ME	2010	43.012019°	-70.669742°	20
RIN	Inshore	Narragansett Bay, RI	2007	41.574047°	-71.330722°	13
RIS	Inshore	Rhode Island Sound	2008	41.285772°	-71.092631°	7
MMV	Offshore	Marthas Vineyard, MA	2010	39.800000°	-71.802500°	15
IOS	Inshore	Isle of Shoals, NH	2010	42.969078°	-70.617383°	14
OHC	Offshore	Hydrographers canyon	2010	40.150000°	-69.050000°	16
OGS	Offshore	Georges Basin	2010	40.943108°	-67.475186°	10
OVC	Offshore	Veatch canyon	2010	40.000908°	-69.606944°	15
OJC	Offshore	Jones canyon	2010	39.500028°	-72.000778°	12
RBB	Inshore	Brown's Bank, RI	2010	41.322503°	-71.092622°	17
RBT	Inshore	Beavertail, RI	2010	41.441219°	-71.400367°	16
OMU	Offshore	McMaster canyon	2010	39.808889°	-71.802500°	9

Table S5.3. Number of putative SNPs retained following each filtering step.

FROM READS TO SNPS	**SNP count**	**Loci count**
STACKS CATALOG	119,811	26,371
POPULATION FILTERS		
Genotyped		
> 80% of the samples	26,544	5,935
> 80% of the populations		
MAF FILTERS		
Global MAF > 0.02	4,148	2,737
Local MAF > 0.2		
COVERAGE FILTER		
From 10 to 100x	4,075	2,685
HWE FILTERS		
Hardy-Weinberg equilibrium (P-value < 0.05) for 60% of the locations	2,553	2,110
F_{IS} between -0.3 and 0.3	1,869	1,484
H_{OBS} < 0.5	1,767	1,424
Linkage Disequilibrium < 0.8	1,717	

Chapitre 6 – Conclusion générale

Les objectifs de cette thèse s'inscrivaient dans une volonté d'améliorer l'état de nos connaissances sur la structuration génétique des populations du homard d'Amérique, afin d'établir des recommandations pour un plan de gestion durable de l'espèce. Plus spécifiquement, notre travail visait à évaluer l'ampleur de la divergence neutre et adaptative des populations de homard d'Amérique ainsi qu'à déterminer l'influence des facteurs géographiques et environnementaux sur cette divergence. Pour cela, nos travaux de recherche ont bénéficié des récentes avancées technologiques en écologie moléculaire qui ont rendu propice l'investigation simultanée des patrons démographiques et adaptatifs chez une espèce non modèle, ici le homard d'Amérique. Une approche de génomique du paysage marin nous a permis (i) de comprendre comment le mouvement des individus au travers du paysage marin peut influencer les patrons démographiques mis en lumière par les outils génomiques et (ii) d'identifier les variables environnementales qui ont une influence significative sur les patrons adaptatifs et qui peuvent donc avoir un impact critique sur la persistance des populations dans un contexte de changement climatique. Par ailleurs, notre projet de recherche a fait partie des premières études de génomique des populations marines utilisant les techniques de RAD-*sequencing*. Par conséquent, nous avons dans un premier temps défini et établi les bases d'un cadre méthodologique pertinent à l'analyse de ce type de données. D'autre part, nous avons identifié les sources de biais potentielles inhérentes à de telles analyses, ce qui nous a conduit à mettre en lumière, pour la première fois, le système de déterminination sexuelle et ses bases moléculaires chez le homard d'Amérique, *via* l'analyse d'un jeu de données de type RAD-*sequencing*. Dans une perspective de recherche appliquée, cette thèse s'inscrit comme une première étape vers la délimitation des unités génétiques du homard d'Amérique dans l'est du Canada et la mise en regard de ces unités avec le plan de gestion. Globalement, l'atteinte de nos objectifs a stimulé la cohésion entre la génétique des populations et la gestion des pêches en permettant d'envisager l'approche génomique comme un outil prometteur permettant de répondre à plusieurs problématiques actuelles de gestion.

6.1. Retour vers les principaux résultats

Le premier chapitre a apporté la plus complète étude de génomique des populations sur le homard d'Amérique dans l'Est du Canada, puisque nous avons récolté et génotypé plus d'échantillons et de marqueurs génétiques qu'aucune autre étude ne l'avait fait auparavant. Par une approche de type RAD-*sequencing*, nous avons été en mesure de démontrer l'existence de deux unités régionales (nord/sud) composées de onze sous-unités génétiques différenciées à plus fine échelle. Nos analyses sont venues ainsi confirmer la structure génétique régionale déjà observée par Kenchington *et al.* (2009), c'est à dire la divergence entre les populations de la région nord (Golf de Saint Laurent et Terre Neuve) avec celle de la région sud (Golfe du Maine et Baie de Fundy) de l'Est du Canada. Néanmoins, cette cohérence a uniquement été observée à l'échelle régionale. En effet, l'augmentation du nombre de marqueurs analysés nous a amené à raffiner les patrons de structuration génétique précédemment décrits et à en révéler des nouveaux. Par exemple, c'est la première fois qu'un phénomène d'isolement par la distance a été mis en évidence chez cette espèce (non détecté par Kenchington *et al.* (2009)). Ce phénomène suggére une forte influence de la distribution spatiale sur les processus démographiques, ce qui est attendu pour de nombreuses espèces marines (Palumbi 1994; 2003). De plus, le génotypage de milliers de marqueurs sur des centaines d'individus nous a donné l'opportunité d'effectuer des tests d'assignation populationnelle. Basé sur la structure régionale mise en évidence par notre étude et celle de Kenchington *et al.* (2009), nous avons obtenu un fort succès d'assignation régional (*i.e.* nord/sud), ce qui nous a indiqué que le signal génétique à l'échelle régionale était suffisamment fort pour mettre en place un outil de traçabilité utile aux pêcheurs, aux consommateurs et aux gestionnaires. En testant la possibilité de développer un tel outil à l'échelle locale, nous avons délimité l'impact du nombre de marqueurs et du nombre d'échantillons sur le succès d'assignation populationnelle en s'inspirant des travaux déjà effectués à partir de marqueurs microsatellites (voir Cornuet *et al.* 1999). Nous avons ainsi souligné l'influence du nombre de marqueurs et du nombre d'échantillons sur le succès d'assignation et nous avons délimité le seuil à partir duquel ce succès commençait à être optimal dans un contexte de faible structuration génétique ($F_{ST} < 0.01$). Ce résultat a permis de donner des recommandations aux futures études de génomique des populations souhaitant réaliser des tests d'assignation populationnelle chez des espèces faiblement différenciées (Benestan *et al.* 2015).

Bien que les tests d'assignation populationnel soient un outil très prometteur pour la gestion des pêches (*e.g.* fraude, écolabels), il est important de noter que nos succès d'assignation populationnelle étaient élevés à l'échelle régionale uniquement, contrairement à ce qui avait été démontré dans notre premier papier (Benestan *et al.* 2015). En parallèle à ce résultat, nous avons constaté que ce succès d'assignation est grandement influencé par le nombre d'individus échantillonnés par population. Ce succès est alors optimal lorsqu'un minimum de 50 à 100 individus est échantillonné par population. Plus ce nombre est grand plus l'estimation de ces fréquences alléliques est précise et juste. Or, en se basant sur les 11 unités génétiques précédemment identifiées par nos calculs de F_{ST}, nous nous retrouvions avec des populations qui ont en majorité une trentaine individus échantillonnés, ce qui est largement inférieur à 50 et donc sous-optimal. Par ailleurs, nous avons remarqué que l'algorithme de classement était extrêmement sensible à un débalancement dans le nombre d'individus par population, ce qui nous a empêché d'effectuer des tests d'assignation populationnelle à l'échelle locale en considérant plus de 50 individus. Nous avons donc été dans l'incapacité de déterminer si le faible succès d'assignation populationnel observé à l'échelle locale était dû à un nombre insuffisant d'individus échantillonnés et/ou à un signal populationnel trop faible pour reconnaître correctement la population d'origine de l'individu. Une prochaine étude qui vise à génotyper un plus grand nombre de homards par site d'échantillonnage ($n \geq 100$) permettra de répondre à cette question.

Suite à l'expertise acquise du premier chapitre, le deuxième chapitre avait pour objectif de décrire chacune des étapes clés nécessaires à l'élaboration et à la mise en place d'un projet de génomique de la conservation. Notre travail s'est concentré à décrire et référencer les différentes méthodes d'analyse disponibles afin d'avoir une vision globale des possibilités d'analyses offertes (Benestan *et al.* 2016a). Nous avons déterminé qu'il était primordial de construire un tel cadre d'analyse à partir d'une question scientifique qui va ensuite diriger l'ensemble des décisions concernant le plan d'échantillonnage, le séquençage, le génotypage et le type d'analyse à utiliser. De plus, les méthodes de filtration des marqueurs génétiques constituent une étape clé dans les analyses de génomique des populations, car elles ont un impact considérable sur les résultats produits. Or, il n'existe pas à l'heure actuelle de *consensus* sur le sujet. Ici, nous avons référencé ces méthodes de filtration ainsi que les nombreuses approches qui y sont associées (*e.g.* balayage génomique, algorithme de regroupement). Globalement, notre travail a fourni à la communauté scientifique une liste

non exhaustive des approches disponibles en génomique de la conservation en indiquant leurs limites et leur pertinence par rapport à une question scientifique donnée. Ultimement, ce travail visait à promouvoir la transparence et la standardisation des méthodes d'analyse dans le domaine de la génomique de la conservation (Benestan *et al.* 2016a).

Par une approche de génomique du paysage marin, le troisième chapitre a permis de mettre en lumière et de quantifier l'influence de la distribution spatiale, des courants océaniques et des SST sur la structuration génétique potentiellement neutre et adaptative des populations (Benestan *et al.* 2016b). Nous avons découvert que la structuration génétique potentiellement neutre était davantage influencée par les courants océaniques que par la distribution spatiale. Ce résultat va de concert avec les nombreuses études de paysage marin qui ont déjà démontré l'avantage de considérer les courants marins comme facteur d'influence des processus démographiques chez les espèces marines (White *et al.* 2010b; Jorde *et al.* 2015). Pour ce chapitre, nous avons développé une approche novatrice qui intègre des méthodes d'écologie du paysage, telles que les *distance-based Moran's Eigenvector Maps* (db-MEM) et les *Asymetric Eigenvector Map* (AEM; Borcard & Legendre 2002; Blanchet *et al.* 2008), à nos données d'écologie moléculaire. Cette approche a permis de remédier aux problèmes de non-indépendance statistique des échantillons qui prévalaient dans les régressions linéaires simples entre F_{ST} et matrice de connectivité larvaire, pourtant communes aux analyses de génomique du paysage marin (Riginos & Liggins 2013). Nous avons ensuite déterminé la part de structure génétique potentiellement adaptative en utilisant à la fois des approches de différentiation génétique (*Population Differentiation* ou PD) et d'association environnementale (*Environmental Association* ou EA) afin de maximiser nos efforts pour détecter les signatures génomiques de la sélection naturelle (Rellstab *et al.* 2015; Francois *et al.* 2016). Cet effort méthodologique apparaît être de plus en plus pertinent puisque de nombreuses méthodes alternatives sont envisagées pour pallier à l'inefficacité des PD à détecter certaines traces de sélection (*e.g.* faible changement de fréquence allélique). De façon intéressante, les deux approches de PD et EA ont conduit aux mêmes résultats en identifiant la température minimale annuelle comme principal facteur environnemental d'influence sur la variation génétique adaptative. Ici, il est à noter que ce facteur a toujours une influence significative sur la variation génétique poentiellement adaptative même après avoir soustrait l'effet de la distribution spatiale sur cette variation. Cette étape était en effet primordiale dans notre contexte d'étude où un isolement par la distance avait été mis en

évidence (Benestan *et al.* 2015). Parmi l'ensemble des marqueurs génétiques identifiés comme potentiellement sous sélection divergente, nous avons détecté des gènes dont les fonctions moléculaires s'accordent avec notre hypothèse d'adaptation locale à la température. La fonctionnalité de ces gènes candidats à l'adaptation à la température devra être confirmée par des analyses de transcriptomique ou mutagénèse dirigée (Barrett & Hoekstra 2011; Benestan *et al.* 2016a).

Finalement, le quatrième chapitre concernait l'utilité des marqueurs sexuels dans les analyses de génomique des populations. À l'aide d'un jeu de données comprenant à la fois des mâles et des femelles, nous avons remarqué qu'il était possible de détecter une structuration génétique au sein d'une population panmictique dans une situation où l'échantillonnage mâle-femelle était biaisé. De façon surprenante, nous avons découvert que la cause de cette structuration génétique venait d'un ensemble de 12 marqueurs liés au sexe qui avaient été gardés et biaisaient ainsi les résultats des analyses multivariées. En enlevant plus de 90% de ces marqueurs, les analyses multivariées ne révélaient alors plus aucune structuration génétique même dans le cas où l'échantillonnage montrait un sexe-ratio (SR) très biaisé (100% de mâles échantillonnés dans un site *versus* 100% de femelles échantillonnés dans un deuxième site). Malgré cet important biais généré par un SR déséquilibré, peu d'études en génomique des populations marines collectent l'information sur le sexe des individus échantillonnés. Plus spécifiquement, chez cette espèce dont les SR peuvent être biaisés en nature (*e.g.* dans des conditions de faible salinité; Jury & Watson 2013), l'identification de cette source d'erreur était primordiale afin de révéler de façon exacte la structuration génétique des populations de homard d'Amérique. Ce chapitre illustre clairement l'utilité de recueillir l'information sur le sexe des individus échantillonnés pour s'affranchir de ce biais potentiel. D'autre part, nous démontrons également la possibilité de déterminer le système de détermination sexuelle d'une espèce non modèle à partir des données de génomiques actuelles.

6.2. Contributions

Dans un premier temps, il convient de souligner le caractère précurseur et novateur du travail de cette thèse. En effet, nous présentons ici un des premiers travaux de recherche examinant la divergence neutre et adaptative des populations d'une espèce marine non modèle *via* l'utilisation des plus récentes ressources génomiques disponibles (ici, RAD-*sequencing*). Ces travaux de recherche ont été accompagnés par de nombreux défis méthodologiques inhérents au travail de laboratoire (*e.g.* préparation des librairies), aux analyses bio-informatiques (*e.g.* développement d'un *pipeline*) et aux tests statistiques effectués (*e.g.* utilisation des variables de db-MEM et d'AEM). En relevant l'ensemble de ces défis, cette thèse a donc fortement contribué à la pleine expansion du domaine de la génomique marine et a clairement illustré l'utilité des outils génomiques pour la gestion et la conservation. Plus particulièrement, le niveau de résolution atteint avec l'analyse de milliers de marqueurs SNPs a permis l'identification de l'architecture génomique sous-jacente aux processus démographiques et adaptatif d'une espèce marine exploitée, ce qui constitue encore à l'heure actuelle un des défis majeur de la génomique marine (Hemmer-Hansen *et al.* 2014).

Malgré les applications prometteuses des nouveaux outils de génomiques (Allendorf *et al.* 2010), leur utilisation a aussi été accompagnée par une prise de conscience des biais liés au développement de tels outils (Davey *et al.* 2013; Benestan *et al.* 2016a). Ces types de biais que nous nous sommes forcés de référencer et de décrire méticuleusement (*e.g.* filtration, sexe-ratio) peuvent engendrer des sources d'erreurs non négligeables dans l'estimation des paramètres génétiques (Benestan *et al.* 2016a). Dans le cas où la structuration génétique est très faible ($F_{ST} < 0.01$), identifier et caractériser ces sources d'erreurs potentielles a permis au domaine de la génomique de la conservation de se construire sur des bases solides. Cet effort a aussi aidé à définir les limites de l'application des résultats de génomique aux problématiques de gestion et de conservation des populations. Cette thèse illustre donc le cheminement méthodologique effectué par la communauté scientifique depuis l'arrivée de ces nouveaux outils (2010) à leur utilisation actuelle (2016).

L'application combinée des méthodes de détection de la sélection naturelle (*e.g*. PD et EA) sur un jeu de données empirique tel que le nôtre est également une première dans le domaine. En effet, de nombreuses études théoriques basées sur des jeux de données simulées ont déjà combiné ces approches alors que du côté des études empiriques, beaucoup se limitent à l'utilisation d'une seule méthode. Cet exercice a ainsi dévoilé la nécessité de tester plusieurs méthodes pour parvenir à repérer tout signal potentiel de la sélection naturelle sur le génome. En effet, les trois gènes identifiés comme les plus probables candidats potentiels à l'adaptation thermique ont tous été détectés par une seule et unique méthode. Par conséquent, l'utilisation d'une seule méthode aurait réduit considérablement notre capacité à identifier des cibles potentielles de la sélection. Cette démonstration va potentiellement orienter le domaine de la génomique adaptive vers une vision plus intégratrice des méthodes de détection. Cette vue d'ensemble pourra ainsi pousser les futures études à passer d'une perspective de « mono-outil » à une dynamique de « poly-outil ».

Notre étude offre un cadre méthodologique innovateur au domaine de la génomique du paysage marin, en ayant transformé les variables représentant les courants marins (ici la matrice de connectivité larvaire) en vecteurs AEM. Cette technique ayant déjà fait ses preuves en écologie du paysage a été pour la première fois appliquée à un contexte de génomique du paysage marin. Cette approche statistique évite les pièges de la colinéarité et de la non-indépendance des données qui se retrouvait dans les regressions linéaires entre les valeurs de Fst et les matrices de connectivités larves. Les avantages d'une telle approche laissent suggérer sa prévalence dans les futures analyses de génomique du paysage marin. Globalement, l'application de cette approche par notre étude contribue à développer cette nouvelle perspective.

6.3. Perspectives

Les aspects développés au cours de cette thèse posent les premières fondations aux travaux de recherche en génomique sur le homard d'Amérique. De nombreux éléments restent à explorer et à aborder. Par exemple, l'interprétation d'une faible structuration génétique en terme de processus démographiques et adaptatifs se heurte à de nombreuses limites méthodologiques qui rendent difficile l'intégration directe de ces résultats à la gestion des pêches. Par exemple, une des premières difficultés demeure l'identification de

populations démographiquement indépendantes (nombre de migrants $m < 0.1$ selon Hastings 1993), qui doivent être gérées et considérées séparément dans l'évaluation des stocks. Alors que les gestionnaires sont intéressés à quantifier le nombre de migrants *m* au sein d'un stock donné, la différentiation génétique estimée à partir des marqueurs moléculaires nous informe sur le N_em (où N_e est la taille efficace de la population et N_em représente le nombre effectif de migrants qui se reproduisent et contribuent à la prochaine génération). Une équation simple déterminée par Wright (1943) existe entre le F_{ST} et N_em : $F_{ST} \sim 1/(1 + 4 N_em)$. Au delà du fait que le modèle (*i.e.* en île) et les conditions (*e.g.* équilibre d'HWE, populations non chevauchantes) supportant cette équation soient rarement respectés chez les populations naturelles, il a été démontré que la relation asymétrique existant entre N_em et les valeurs de F_{ST} implique que le calcul de *m* devient imprécis au delà d'une certaine valeur de N_e ($N_e > 10^3$; Waples & Gagiotti 2006). Par conséquent, chez la majorité des espèces marines, qui ont communément des $N_e > 10^3$, définir avec suffisamment de précision les ensembles d'individus qui sont ou non démographiquement indépendants reste une tâche complexe à accomplir. C'est pourquoi, un nouvel échantillonnage qui comprendrait les mêmes sites étudiés ici sur plusieurs années (*i.e.* réplicats temporels) est nécessaire. En effet, cet échantillonnage fournirait un moyen efficace d'évaluer si les patrons de structuration génétique observés persistent au fil du temps, et par conséquent, si l'analyse génétique effectuée à partir d'échantillons prélevés sur une année (*i.e,* potentiellement relatif une génération) pourrait être assez fiable pour mettre en évidence les processus démographiques de l'espèce (*i.e*, relatifs à plusieurs générations). Ces réplicats temporels permettraient également de confirmer ou d'infirmer nos résultats, et ainsi de rendre compte de l'exactitude de la structuration génétique décrite dans la présente thèse.

Il est également important de noter que définir des unités génétiques à partir d'un test statistique effectué sur les $F_{ST,}$ comme ce fut le cas de notre étude, ne tient pas compte du fait que la différenciation populationnelle peut suivre un continuum. Ainsi, le résultat d'un tel test permet uniquement de rejeter l'hypothèse de panmictique et non de définir l'ampleur de connectivité entre les populations (Waples & Gagiotti 2006). De plus, cette méthode est très dépendante du pouvoir statistique (*e.g.* nombre d'individus, nombre de marqueurs) et peut amener à la détection de faibles différences génétiques statistiquement significatives alors qu'elles sont trop faibles pour réellement donner des informations biologiques pertinentes à utiliser dans un contexte de gestion (Waples *et al.* 2008). Bien que ce risque existe, des

études ont également démontré que des valeurs de F_{ST} faibles (F_{ST} ~ 0.001) mais significatives peuvent être obtenus dans un contexte de divergence phénotypique (Aykanat *et al.* 2015) ou même lorsqu'il y a peu d'échange entre les individus provenant de deux sites d'échantillonnage éloignés (Knutsen *et al.* 2011). Ces deux études empiriques témoignent du fait qu'une indépendance démographique peut être présente même lorsque la structuration génétique est très faible. De façon similaire, ce résultat a été démontré par Waples et al. (2008) à partir de l'analyse de divers scénarios démographiques simulés. Dans le cas des études empiriques, les auteurs surmontent les potentiels biais qui affectent l'interprétation d'une faible différentiation génétique en intégrant brillamment leurs résultats à une approche pluridisciplinaire où les données génétiques sont couplées à des connaissances sur les mouvements de l'espèce étudiée, ses traits d'histoire de vie, son comportement ou encore sa physiologie (voir Cooke *et al.* 2011). Dans le cadre d'une prochaine étude sur la structure génétique du homard d'Amérique, il serait par exemple judicieux d'analyser en synergie des outils de télémétrie et de génomique afin de tester si le flux génique estimé à l'aide des outils génomiques est réellement proportionnel au nombre de migrants identifiés à partir des outils télémétriques. Ce type d'étude serait d'autant plus pertinente à mettre en place depuis que nous avons récemment démontré qu'il était possible d'utiliser des tests d'assignation populationnelles, et donc inférer un nombre de migrants potentiels, chez une espèce faiblement structurée.

En vertu des connaissances acquises sur le nombre et la distribution spatiale des unités génétiques identifiées ici, il est devenu opportun de mettre en lumière cette structuration génétique à plus haute résolution spatiale (*i.e.* augmentation du nombre d'échantillons). En effet, les unités génétiques regroupant plusieurs sites d'échantillonnage (*e.g.* SGL), sont toutes proches géographiquement et le phénomène d'isolement par la distance semblait très présent car hautement significatif (P-valeur < 0.001). Aussi, à l'aide d'un échantillonnage à plus fine échelle, nous serions en mesure de tester à quelle distance géographique la présence d'un signal génétique persiste (*i.e.* autocorrélation spatiale) et ainsi inférer la capacité de dispersion de l'espèce. Par ailleurs, Legendre et *al*. (2015) ont récemment prouvé que le test de Mantel était une méthode inappropriée pour tester l'existence de patrons d'isolement par la distance, en partie parce que ce dernier est très sujet aux erreurs de type I. De plus, le coefficient de R^2 donné par ce test n'est pas interprétable et donc n'apporte aucune information pertinente sur l'ampleur de la corrélation existant entre

les données génétiques et celles spatiales. C'est pourquoi nous avons adopté la méthode de RDA combinée aux variables de db-MEM pour le troisième chapitre, et que nous recommandons fortement cette approche pour les prochaines études. Cette approche originale, appliquée à la génomique du paysage marin pour la première fois, pourrait être bonifiée en y incluant d'autres variables environnementales (*e.g.* température de fond de la mer, salinité) en plus des variables testés ici (*e.g.* température de surface de la mer). Par ailleurs, l'ajout de ces données pourrait expliquer une portion de la variance génétique neutre et adaptative encore inexpliquée par les variables que nous avions utilisées (voir Selkoe *et al*. 2016).

6.4. Vers une approche pratique de la génomique de la conservation

Malgré les promesses de la génomique, ses applications à des enjeux de conservation sont rares (Shafer *et al.* 2015 mais voir Garner *et al.* 2015). En effet, de nouveaux outils, défis et doutes, qui n'existaient alors pas en génétique de la conservation, ont fait surface. C'est dans ce contexte que la grande majorité de cette thèse s'est concentrée à améliorer notre compréhension des outils génomiques et à définir de façon précise leurs limites par rapport à l'information biologique révélée. Cette étape clé a ainsi permis au milieu de la génomique de la conservation de se développer afin de mieux répondre aux attentes des gestionnaires. En gestion des pêches plus particulièrement, les outils génomiques peuvent (i) aider à la délimitation d'unités de gestion durable (*i.e.*, en déquation avec la biologie de l'espèce) en identifiant des unités génétiques et en quantifiant leur connectivité au sein d'un espèce donnée, (ii) limiter la fraude et donner l'accès à des « ecolabels » en développant des outil de traçabilité et (iii) assister l'ensemencement de populations « pré-adaptées » aux conditions environnementales des site ensemencés en testant l'existence de patrons d'adaptation locale.

Les priorités actuelles en terme de gestion et de conservation sont majoritairement portées vers l'identification de la variation génétique adaptative et ses conséquences sur la valeur sélective ainsi que la quantification de la connectivité génétique et son influence sur les aléas démographiques. Bien que les outils génomiques mettent en évidence ces patrons adaptatifs et démographiques, nous insistons sur la nécessité d'inclure les informations biologiques propre à l'organisme étudié afin d'obtenir de plus précis estimés et d'émettre des recommandations de gestion appropriées. En effet, nos analyses des patrons démographiques et adaptatifs doivent respectivement être validées par des réplicats temporels et des

corrélations génotype-phénotype fortes avant qu'une application en terme de gestion ne soit mise en place. Cependant, nos efforts visant à faire le pont entre la génomique et la gestion des pêches a pour but ultime de stimuler la cohésion et la synergie de ces deux domaines, en proposant un exemple clair de génomique de la conservation en action.

Chapitre 7. Références bibliographiques

Ai H, Yang B, Li J *et al.* (2014) Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. *BMC genomics*, **15**, 834.
Aiken DE, Waddy SL (1986) Environmental Influence on Recruitment of the American Lobster Homarus americanus: A Perspective. *Canadian Journal of Fisheries and Aquatic Sciences*, **43**, 2258–2270.
Alex Buerkle C, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.
Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, **19**, 1655–1664.
Ali OA, O'Rourke SM, Amish SJ *et al.* (2015) RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. *Genetics*, genetics, **115**, 183665.
Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques. *Nucleic acids research*, **25**, 4692–4693.
Allendorf FW, Bassham S, Cresko WA *et al.* (2015) Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity*, **106**, 217–227.
Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.
Amaral AR, Beheregaray LB, Bilgmann K *et al.* (2012) Seascape Genetics of a Globally Distributed, Highly Mobile Marine Mammal: The Short-Beaked Common Dolphin (Genus *Delphinus*). *PloS one*, **7**, e31482.
Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.
Anderson ECAC, Waples RSWS, Kalinowski STKT (2008) An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 1475–1486.
Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.
Annis ER (2005) Temperature effects on the vertical distribution of lobster postlarvae (*Homarus americanus*). *Limnology and Oceanography*, **50**, 1972–1982.
Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: A workbench to detect molecular adaptation based on a F st -outlier method. *BMC bioinformatics*, **9**, 323.
Antao T (2015). *Bioinformatics with Python cookbook*. Packt Publishing Ltd.
Atasaral Şahin Ş, Romero MR, Cueto R *et al.* (2015) Subtle tissue and sex‐dependent proteome variation in mussel (*Mytilus galloprovincialis*) populations of the Galician coast (NW Spain) raised in a common environment (T Knigge, Ed,). *Proteomics*, **15**, 3993–4006.
Aykanat T, Johnston SE, Orell P *et al.* (2015) Low but significant genetic differentiation underlies biologically meaningful phenotypic divergence in a large Atlantic salmon population. *Molecular ecology*, **24**, 5158-5174.
Baird SJE (2015) Exploring linkage disequilibrium. *Molecular Ecology Resources*, **15**, 1017–1019.
Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*, **28**, 45–48.
Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population

assignment power? *Bioinformatics*, **19**, 1436–1438.
Banks SC, Piggott LMP, Williamson JE *et al.* (2007) Oceanic variability and coastal topography shape genetic structure in a long-dispersing sea urchin. *Ecology*, **88**, 3055–3064.
Barrett LW, Fletcher S, Wilton SD (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences*, **69**, 3613–3634.
Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.
Baums IB, Paris CB, Chérubin LM (2006) A bio-oceanographic filter to larval dispersal in a reef-building coral. *Limnology and Oceanography*, **51**, 1969–1981.
Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism. *PloS one*, **6**, e19315.
Beaumont MA & Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B: Biological Sciences*, **263**, 1619-1626.
Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.
Benestan L, Ferchaud A-L, Hohenlohe P *et al.* (2016a) Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular Ecology*.
Benestan L, Gosselin T, Perrier C *et al.* (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, **24**, 3299–3315.
Benestan L, Quinn BK, Maaroufi H *et al.* (2016b) Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (*Homarus americanus*). *Molecular Ecology*.
Benjamini Y & Hochberg Y (1994). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, **1,** 289-300.
Bernatchez L, Duchesne P (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 1–12.
Berry O, Tocher MD, Sarre SD (2004) Can assignment tests measure dispersal? *Molecular Ecology*, **13**, 551–561.
Blanchet FG, Legendre P, Borcard D (2008) Modelling directional spatial processes in ecological data. *Ecological Modelling*, **215**, 325–336.
Blanchet FG, Legendre P, Maranger R, Monti D, Pepin P (2011) Modelling the effect of directional spatial ecological processes at different scales. *Oecologia*, **166**, 357–368.
Boldina I, Beninger PG (2016) Strengthening statistical usage in marine ecology: Linear regression. *Journal of Experimental Marine Biology and Ecology*, **474**, 81–91.
Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.
Bourret V, Dionne M, Bernatchez L (2014) Detecting genotypic changes associated with selective mortality at sea in Atlantic salmon: polygenic multilocus analysis surpasses

genome scan. *Molecular Ecology*, **23**, 4444–4457.
Bourret V, Kent MP, Primmer CR *et al.* (2013) SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **22**, 532–551.
Bowler DE, Benton TG (2005) Causes and consequences of animal dispersal strategies: relating individual behaviour to spatial dynamics. *Biological Reviews*, **80**, 205–225.
Bowlby HD, Hanson JM & Hutchings JA (2007). Resident and dispersal behavior among individuals within a population of American lobster *Homarus americanus*. *Marine Ecology Progress Series*, **331**, 207-218.
Bradbury IR, Hubert S, Higgins B *et al.* (2010) Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 3725–3734.
Breyne P, Mergeay J, Casaer J (2014) Roe deer population structure in a highly fragmented landscape. *European Journal of Wildlife Research*, **60**, 909–917.
Brickman D, Drozdowski A (2012a) *Atlas of model currents and variability in Maritime Canadian waters*, **277**, vii-64.
Brickman D, Drozdowski A (2012b) *Development and validation of a regional shelf model for Maritime Canada based on the NEMO-OPA circulation model*, **278**, vii-57.
Brieuc M, Waters CD (2014) *A dense linkage map for Chinook salmon (Oncorhynchus tshawytscha) reveals variable chromosomal divergence after an ancestral whole genome* G3: Genes| Genomes|Genetics, **4**, 447–460.
Bruneaux M, Johnston SE, Herczeg G *et al.* (2013) Molecular evolutionary and population genomic analysis of the nine‐spined stickleback using a modified restriction‐site‐associated DNA tag approach. *Molecular Ecology*, **22**, 565–582.
Burgess SC, Baskett ML, Grosberg RK, Morgan SG, Strathmann RR (2015) When is dispersal for dispersal? Unifying marine and terrestrial perspectives. *Biological Reviews*.
Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, **3**, 1–27.
Campbell A (1986) Migratory Movements of Ovigerous Lobsters, *Homarus americanus*, Tagged off Grand Manan, Eastern Canada. *Canadian Journal of Fisheries and Aquatic Sciences*, **43**, 2197–2205.
Campbell A, Stasko AB (1986) Movements of lobsters (Homarus americanus) tagged in the Bay of Fundy, Canada. *Marine Biology*, **92**, 393–404.
Candy JR, Campbell NR, Grinnell MH, Beacham TD, Larson WA & Narum SR (2015). Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus*, Osmeridae), an anadromous Pacific smelt. *Molecular ecology resources*, **15**, 1421-1434.
Cano JM, Shikano T, Kuparinen A, Merila J (2008) Genetic differentiation, effective population size and gene flow in marine fishes: implications for stock management. *Journal of Integrative Field Biology*, **5**, 1–10.
Cantarel BL, Weaver D, McNeill N *et al.* (2014) BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC bioinformatics*, **15**, 1.
Castric V, Bernatchez L (2004) Individual assignment test reveals differential restriction to dispersal between two salmonids despite no increase of genetic differences with distance. *Molecular Ecology*, **13**, 1299–1312.
Catarino D, Knutsen H, Veríssimo A *et al.* (2015) The Pillars of Hercules as a bathymetric

barrier to gene flow promoting isolation in a global deep-sea shark (*Centroscymnus coelolepis*). *Molecular Ecology*, **24**, 6061–6079.
Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.
Chasse J, Miller RJ (2010) Lobster larval transport in the southern Gulf of St. Lawrence. *Fisheries Oceanography*, **19**, 319–338.
Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet*, **2**, e64.
Chiasson M, Miron G, Daoud D, Mallet MD (2015) Effect of Temperature on the Behavior of Stage IV American Lobster (*Homarus americanus*) Larvae. *dx.doi.org*, **34**, 545–554.
Chittleborough RG (1974) Home range, homing and dominance in juvenile western rock lobsters. *Marine and Freshwater Research*, **25**, 227.
Churcher AM, Pujolar JM, Milan M *et al.* (2015) Transcriptomic profiling of male European eel (*Anguilla anguilla*) livers at sexual maturity. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, **16**, 28–35.
Cobb JS, Wang D, Campbell DB (1989) Timing of settlement by postlarval lobsters (*Homarus americanus*): field and laboratory evidence. *Journal of Crustacean Biology*, **9**, 60–66.
Comeau M, Savoie F (2002) Movement of American lobster (*Homarus americanus*) in the southwestern Gulf of St. Lawrence. *Fishery Bulletin*, **100**, 181-192.
Cooke SJ, Hinch SG, Farrell AP *et al.* (2011) Developing a Mechanistic Understanding of Fish Migrations by Linking Telemetry with Physiology, Behavior, Genomics and Experimental Biology: An Interdisciplinary Case Study on Adult Fraser River Sockeye Salmon. *Fisheries*, **33**, 321–339.
Cooper RA & Uzmann JR (1971) Migrations and growth of deep-sea lobsters, *Homarus americanus*. *Science*, **171**, 288-290.
Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.
Costello MJ, Coll M, Danovaro R *et al.* (2010) A Census of Marine Biodiversity Knowledge, Resources, and Future Challenges (S Humphries, Ed,). *PloS one*, **5**, e12110.
Cowan DF, Watson WH, Solow AR & Mountcastle AM. (2007). Thermal histories of brooding lobsters, *Homarus americanus*, in the Gulf of Maine. *Marine Biology*, **150**, 463-470.
Crossin GT, Hinch SG, Cooke SJ, Welch DW (2007) Behaviour and physiology of sockeye salmon homing through coastal waters to a natal river. *Marine Biology*, **152**, 905-918.
Crossin G, Al-Ayoub S, Jury S, Howell W (1998) Behavioral thermoregulation in the American lobster Homarus americanus. *Journal of Experimental Biology*, **201**, 365–374.
Cui P, Lin Q, Ding F *et al.* (2010) A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, **96**, 259–265.
Cutler DJ, Jensen JD (2010) To pool, or not to pool? *Genetics*, **186**, 41–43.
D'Aloia CC, Bogdanowicz SM, Majoris JE, Harrison RG, Buston PM (2013) Self-recruitment in a Caribbean reef fish: a method for approximating dispersal kernels accounting for seascape. *Molecular Ecology*, **22**, 2563–2572.
D'Amico, S, Claverie P, Collins T, Georlette D *et al.* (2002). Molecular basis of cold adaptation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **357**, 917-925.

Darwin C (1872) *The Origin of Species.* London: John Murry.
Davey JW, Cezard T, Fuentes Utrilla P *et al.* (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.
Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.
Dawson MN, Hamner WM (2008) A biophysical perspective on dispersal and the geography of evolution in marine and terrestrial systems. *Journal of The Royal Society Interface*, **5**, 135–150.
DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.
Devlin RH, Nagahama Y (2002) Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture*, **208**, 191–364.
Dittman A, Quinn T (1996) Homing in Pacific salmon: mechanisms and ecological basis. *Journal of Experimental Biology*, **199**, 83–91.
Drinkwater K, Gilbert D (2004) Hydrographic variability in the waters of the Gulf of St. Lawrence, the Scotian Shelf and the eastern Gulf of Maine (NAFO Subarea 4) during 1991-2000. *Journal of Northwest Atlantic Fishery,* **34**, 85–101.
Dunn RT, Gleason BA, Hartley DP (1999) Postnatal ontogeny and hormonal regulation of sulfotransferase SULT1B1 in male and female rats. *Journal of Pharmacology and Experimental Therapeutics*, **290**, 319-324.
Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 121–1849.
Endler JA (1986) *Natural selection in the wild*. Princeton University Press. Princeton.
Ennis GP (1986) Stock Definition, Recruitment Variability, and Larval Recruitment Processes in the American Lobster, *Homarus americanus*: A Review. *Canadian Journal of Fisheries and Aquatic Sciences*, **43**, 2072–2084.
Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.
Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.
Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.
Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.
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.
Ferretti L, Ramos Onsins SE, Pérez Enciso M (2013) Population genomics from pool sequencing. *Molecular Ecology*, **22**, 5561–5576.
Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–993.
Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L (2014) Widespread Signals of

Convergent Adaptation to High Altitude in Asia and America. *The American Journal of Human Genetics*, **95**, 394–407.
Follesa MC, Cuccu D, Cannas R *et al.* (2009) Movement patterns of the spiny lobster *Palinurus elephas* (Fabricius, 1787) from a central western Mediterranean protected area. *Scientia Marina*, **73**, 499–506.
Francois O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.
Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology*, **46**, 493–505.
Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Molecular Ecology*, **10**, 2741–2752.
Frichot E, Schoville SD, Bouchard G, Francois O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.
Frichot É, Schoville S, Bouchard G, François O (2012) Landscape genomic tests for associations between loci and environmental gradients. *arXiv*.
Frisch AJ (2007) Short-and long-term movements of painted lobster (*Panulirus versicolor*) on a coral reef at Northwest Island, Australia. *Coral Reefs*, **26**, 311–317.
Fu C, Fanning LP (2011) Spatial Considerations in the Management of Atlantic Cod off Nova Scotia, Canada. *North American Journal of Fisheries Management*, **24**, 775–784.
Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, **27**, 489–496.
Futschik A, Schlötterer C (2010) The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, **186**, 207–218.
Gagnaire P-A, Broquet T, Aurelle D *et al.* (2015) Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications*, **8**, 769–786.
Gagnaire P-A, Normandeau E, Côté C, Møller Hansen M, Bernatchez L (2012) The genetic consequences of spatially varying selection in the panmictic American eel (*Anguilla rostrata*). *Genetics*, **190**, 725–736.
Galindo HM, Olson DB, Palumbi SR (2006) Seascape genetics: a coupled oceanographic-genetic model predicts population structure of Caribbean corals. *Current Biology*, **16**, 1622–1626.
Galindo J, Grahame JW, Butlin RK (2010) An EST-based genome scan using 454 sequencing in the marine snail Littorina saxatilis. *Journal of Evolutionary Biology*, **23**, 2004–2016.
Gamble T, Zarkower D (2014) Identification of sex-specific molecular markers using restriction site-associated DNA sequencing. *Molecular Ecology Resources*, **14**, 902–913.
Garner BA, Hand BK, Amish SJ *et al.* (2016) Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application. *Trends in Ecology & Evolution*, **31**, 81–83.
Garroway CJ, Radersma R, Sepil I *et al.* (2013) Fine-scale genetic structure in a wild bird population: the role of limited dispersal and environmentally based selection as causal factors. *evolution*, **67**, 3488–3500.
Gilbert KJ, Andrew RL, Bock DG *et al.* (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program

structure. *Molecular Ecology*, **21**, 4925–4930.
Gilg MR, Hilbish TJ (2003) The geography of marine larval dispersal: coupling genetics with fine-scale physical oceanography. *Ecology*, **84**, 2989–2998.
Godhe A, Egardt J, Kleinhans D *et al.* (2013) Seascape analysis reveals regional gene flow patterns among populations of a marine planktonic diatom. *Proceedings of the Royal Society of London B: Biological Sciences*, **280**, 20131599–20131599.
Goudet J (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184-186.
Guo B, DeFaveri J, Sotelo G, Nair A, Merilä J (2015) Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC biology*, **13**, 19.
Guo S-Z, Huang K, Shi Y-Y *et al.* (2007) A Case-control association study between the GRID1 gene and schizophrenia in the Chinese Northern Han population. *Schizophrenia Research*, **93**, 385–390.
Hand BK, Hether TD, Kovach RP, Muhlfeld CC (2015) Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, **61,** 146-154.
Hannah CG, Shore JA, Loder JW (2001) Seasonal Circulation on the Western and Central Scotian Shelf*. *Journal of Physical Oceanography,* **31**, 591–615.
Hansen MM, Hemmer-Hansen J (2007) Landscape genetics goes to sea. *Journal of Biology*.
Hanski, I. (1998). Metapopulation dynamics. *Nature*, **396,** 41-49.
Harding GC, Kenchington EL, Bird CJ, Pezzack DS, Landry DC (1997) Genetic relationships among subpopulations of the American lobster (*Homarus americanus*) as revealed by random amplified polymorphic DNA. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 1762–1771.
Hastings A (1993) Complex interactions between dispersal and dynamics: lessons from coupled logistic equations. *Ecology*, **74**, 1362.
Hedgecock D (1986) Is gene flow from pelagic larval dispersal important in the adaptation and evolution of marine invertebrates? *Bulletin of Marine Science*, **39**, 550-564.
Hedrick PW (2011) *Genetics of populations*.
Hemmer-Hansen J, Therkildsen NO, Pujolar JM (2014) Population genomics of marine fishes: next-generation prospects and challenges. *The Biological bulletin*, **227**, 117–132.
Herrnkind WF, Vanderwalker JA, Barr L (1975) Population dynamics, ecology and behavior of spiny lobsters, *Panulirus argus*, of St. John, USVI IV. Habitation, patterns of movement. *Science Bulletin of Natural History Museum Los Angeles County*, **20,** 31-45.
Hesketh J (2004) 3′-Untranslated regions are important in mRNA localization and translation: lessons from selenium and metallothionein. *Biochemical Society Transactions*, **32**, 990–993.
Hess JE, Campbell NR, Close DA, Docker MF, Narum SR (2013) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology*, **22**, 2898–2916.
Hochachka PW, Somero GN (2014) *Biochemical Adaptation*. Princeton University Press.
Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags (DJ Begun, Ed,). *PLoS Genet*, **6**, e1000862.
Hohenlohe PA, Day MD, Amish SJ *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, **22**, 3002–3013.

Howell WH, Watson WH III, Jury S (1999) Skewed sex ratio in an estuarine lobster (Homarus americanus) population. **18**, 193–201.
Hoyoux A, Jennes I, Dubois P *et al.* (2001) Cold-adapted beta-galactosidase from the Antarctic psychrophile Pseudoalteromonas haloplanktis. *Applied and Environmental Microbiology*, **67**, 1529–1535.
Hughes JB (2014) Variability of Chromosome Number in the Lobsters, *Homarus americanus* and *Homarus gammarus*. *Caryologia*. **35**, 279-289.
Iacchei M, Ben Horin T, Selkoe KA *et al.* (2013) Combined analyses of kinship and FST suggest potential drivers of chaotic genetic patchiness in high gene‐flow populations. *Molecular Ecology*, **22**, 3476–3494.
Incze LS, Naimie CE (2000) Modelling the transport of lobster (*Homarus americanus*) larvae and postlarvae in the Gulf of Maine. *Fisheries Oceanography*, **9**, 99–113.
James MO (2011) Steroid catabolism in marine and freshwater fish. *The Journal of Steroid Biochemistry and Molecular Biology*, **127**, 167–175.
Jensen JD, Foll M, Bernatchez L (2016) The past, present and future of genomic scans for selection. *Molecular Ecology*, **25**, 1–4.
Johnston SE, Orell P, Pritchard VL *et al.* (2014) Genome‐wide SNP analysis reveals a genetic basis for sea‐age variation in a wild population of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **23**, 3452–3468.
Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.
Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, **11**, 94.
Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.
Jorde PE, Søvik G, Westgaard JI *et al.* (2015) Genetically distinct populations of northern shrimp, *Pandalus borealis*, in the North Atlantic: adaptation to different temperatures as an isolation factor. *Molecular Ecology*, **24**, 1742–1757.
Jury S, Watson WH III (2013) Seasonal and sexual differences in the thermal preferences and movements of American lobsters. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 1650–1657.
Kalinowski ST (2010) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity*, **106**, 625–632.
Kanno Y, Vokoun JC, Letcher BH (2011) Fine-scale population structure and riverscape genetics of brook trout (*Salvelinus fontinalis*) distributed continuously along headwater channel networks. *Molecular Ecology*, **20**, 3711–3729.
Karasova P, Spiwok V, Mala S, Kralova B (2002) Beta-galactosidase activity in psychrotrophic microorganisms and their potential use in food industry. *Czech J. Food. Sci.*, **20**, 43–47.
Karnofsky EB, Atema J, Elgin RH (1989) Field observations of social behavior, shelter use, and foraging in the lobster, *Homarus americanus*. *The Biological bulletin*, **176**, 239.
Kelley JL, Yee M-C, Lee C *et al.* (2012) The possibility of de novo assembly of the genome and population genomics of the mangrove rivulus, *Kryptolebias marmoratus*. *Integrative and comparative biology*, **52**, 737–742.
Kelly S, MacDiarmid AB (2003) Movement patterns of mature spiny lobsters, *Jasus edwardsii*, from a marine reserve. *New Zealand Journal of Marine Freshwater Research*,

37, 149–158.
Kemppainen P, Knight CG, Sarma DK *et al.* (2015) Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Molecular Ecology Resources*, **15**, 1031–1045.
Kenchington EL, Harding GC, Jones MW, Prodöhl PA (2009) Pleistocene glaciation events shape genetic structure across the range of the American lobster, Homarus americanus. *Molecular Ecology*, **18**, 1654–1667.
Kershaw F, Rosenbaum HC (2014) Ten years lost at sea: response to Manel and Holderegger. *Trends in Ecology & Evolution*, **29**, 69–70.
Knutsen H, Jorde PE, André C, Stenseth NC (2003) Fine-scaled geographical population structuring in a highly mobile marine species: the Atlantic cod. *Molecular Ecology*, **12**, 385–394.
Knutsen H, Jorde PE, Sannæs H *et al.* (2009) Bathymetric barriers promoting genetic structure in the deepwater demersal fish tusk (*Brosme brosme*). *Molecular Ecology*, **18**, 3151–3162.
Knutsen H, Olsen EM, Jorde PE *et al.* (2011) Are low but statistically significant levels of genetic differentiation in marine fishes "biologically meaningful?" A case study of coastal Atlantic cod. *Molecular Ecology*, **20**, 768–783.
Kodama M, Brieuc MSO, Devlin RH, Hard JJ, Naish KA (2014) Comparative mapping between Coho Salmon (*Oncorhynchus kisutch*) and three other salmonids suggests a role for chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. *G3 (Bethesda, Md.)*, **4**, 1717–1730.
Kohn MH, Murphy WJ, Ostrander EA (2006) Genomics and conservation genetics. *Trends in Ecology & Evolution,* **21**, 629-637.
Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC bioinformatics*, **15**, 356.
Krück NC, Innes DI, Ovenden JR (2013) New SNPs for population genetic analysis reveal possible cryptic speciation of eastern Australian sea mullet (*Mugil cephalus*). *Molecular Ecology Resources*, **13**, 715–725.
Lamichhaney S, Barrio AM, Rafati N, Sundström G, Rubin CJ, Gilbert ER *et al.* (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, **109,** 19345-19350.
Laporte M, Pavey SA, Rougeux C *et al.* (2016) RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Molecular Ecology*, **25**, 219–237.
Larson WA, Seeb LW, Everett MV *et al.* (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, **7**, 355–369.
Lawton P, Lavalli KL (1995) Postlarval, Juvenile, Adolescent, and Adult Ecology. In: *Biology of the lobster Homarus americanus* (ed. Factor JR), pp. 47-88. Academic Press, San Diego, California.
Le SQ, Durbin R (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome research*, **21**, 952–960.
Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.
Legendre P, Borcard D, Blanchet FG, Dray S (2012) *PCNM: MEM spatial eigenfunction and principal coordinate analyses*. R package version.

Legendre P, Fortin M-J, Borcard D (2015) Should the Mantel test be used in spatial analysis? (P Peres-Neto, Ed,). *Methods in Ecology and Evolution*, **6**, 1239–1247.
Legrand JJ, Legrand-Hamelin E, Juchault P (1987) Sex determination in crustacea. *Biological Reviews*, **62**, 439–470.
Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, **17**, 183–189.
Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.
Lécher P, Defaye D, Noel P (2011) Chromosomes and nuclear DNA of Crustacea. *Invertebrate Reproduction & Development*, **27**, 85–114.
Li Y, Hu J, Liu J *et al.* (2015) Genome‐wide analysis of gene expression profiles during early ear development of sweet corn under heat stress (R Tuberosa, Ed,). *Plant Breeding*, **134**, 17–27.
Limborg MT, Helyar SJ, De Bruyn M *et al.* (2012) Environmental selection on transcriptome‐derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, **21**, 3686–3703.
Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.
Lotterhos KE & Whitlock MC (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular ecology*, **23**, 2178-2192.
Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.
Lowe WH & Allendorf F (2010). What can genetics tell us about population connectivity?. *Molecular ecology*, **19**, 3038-3051.
Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.
Lynch M, Bost D, Wilson S, Maruki T, Harrison S (2014) Population-genetic inference from pooled-sequencing data. *Genome Biology and Evolution*, **6**, 1210–1218.
MacKenzie BR (1988) Assessment of temperature effects on interrelationships between stage durations, mortality, and growth in laboratory-reared *Homarus americanus* Milne Edwards larvae. *Journal of Experimental Marine Biology and Ecology*, **116**, 87–98.
Malhis N & Jones SJ (2010). High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics*, **26**, 1029-1035.
Manel S, Holderegger R (2013) Ten years of landscape genetics. *Trends in Ecology & Evolution*, **28**, 614–621.
Manel S, Segelbacher G (2009) Perspectives and challenges in landscape genetics. *Molecular Ecology*, **18**, 1821–1822.
Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, **20**, 136–142.
Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.
Marko, P. B., & Hart, M. W. (2011). The complex analytical landscape of gene flow inference. *Trends in ecology & evolution*, **26**, 448-456.
McCauley DJ, Pinsky ML, Palumbi SR *et al.* (2015) Marine defaunation: Animal loss in the

global ocean. *science*, **347**, 1255641–1255641.
McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–1303.
Meirmans PG (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**, 2839–2846.
Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*.
Meirmans PG, Van Tienderen PH (2013) The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity*, **110**, 131–137.
Miller MR, Dunham JP, Amores A, Cresko WA & Johnson EA (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome research*, **17**, 240-248.
Miller TEX, Inouye BD (2013) Sex and stochasticity affect range expansion of experimental invasions (F Courchamp, Ed,). *Ecology Letters*, **16**, 354–361.
Mita S, Thuillet AC, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.
Mora C, Sale PF (2002) Are populations of coral reef fish open or closed? *Trends in Ecology & Evolution*, **17**, 422–428.
Morin PA, Luikart G, Wayne RK, group TSW (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.
Morse B, Rochette R (2016) Movements and activity levels of juvenile American lobsters Homarus americanus in nature quantified using ultrasonic telemetry. *Mar Ecol Prog Ser*, **551**, 155–170.
Narum SR (2006) Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conservation Genetics*, **7**, 783–787.
Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.
Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.
Narum, S. R., & Campbell, N. R. (2015). Transcriptomic response to heat stress among ecologically divergent populations of redband trout. *BMC genomics*, **16**, 1.
Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, **13**, 667–672.
Nielsen EE, Cariani A, Mac Aoidh E *et al.* (2012) Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature communications*, **3**, 851.
Nielsen R (2005) Molecular signatures of natural selection. *Annual review of genetics*, **39**, 197–218.
Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.
Nomura T (2008) Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evolutionary Applications*, **1**, 462–474.
Ogden R, Gharbi K, Mugue N *et al.* (2013) Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology*, **22**, 3112–3123.
Oksanen J, Kindt R, Legendre P, O'Hara B (2007) The vegan package. *Community ecology*

package 10.
Osten P, Srivastava S, Inman GJ *et al.* (1998) The AMPA Receptor GluR2 C Terminus Can Mediate a Reversible, ATP-Dependent Interaction with NSF and α- and β-SNAPs. *Neuron*, **21**, 99–110.
Ouborg NJ, Angeloni F, Vergeer P (2010) An essay on the necessity and feasibility of conservation genomics. *Conservation Genetics*, **11**, 643–653.
Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.
Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Molecular Ecology*, **13**, 55–65.
Palsboll P, Berube M, Allendorf FW (2007) Identification of management units using population genetic data. *Trends in Ecology & Evolution*, **22**, 11–16.
Palumbi SR (1994) Genetic Divergence, Reproductive Isolation, and Marine Speciation. *Annual Review of Ecology and Systematics*, **25**, 547–572.
Palumbi SR (2003) Population genetics, demographic connectivity, and the design of marine reserves. *Ecological applications*, **13**, 146–158.
Pauly D, Christensen V, Dalsgaard J, Froese R, Torres F (1998) Fishing Down Marine Food Webs. *science*, **279**, 860–863.
Pauly D, Christensen V, Guénette S, Pitcher TJ., Sumaila UR. *et al.* (2002). Towards sustainability in world fisheries. *Nature*, **418**, 689-695.
Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR (2012) What is needed for next-generation ecological and evolutionary genomics? *Trends in Ecology & Evolution*, **27**, 673–678.
Peres Neto PR, Legendre P (2010) Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography*, **19**, 174–184.
Perras M, Nebel S (2012) Satellite telemetry and its impact on the study of animal migration. *Nature Education Knowledge*, **3**.
Pezzack DS, Duggan DR (1986) Evidence of Migration and Homing of Lobsters (*Homarus americanus*) on the Scotian Shelf. *Canadian Journal of Fisheries and Aquatic Sciences*, **43**, 2206–2211.
Pinsky ML, Palumbi SR (2014) Meta-analysis reveals lower genetic diversity in overfished populations. *Molecular Ecology*, **23**, 29–39.
Piry S, Alapetite A, Cornuet JM *et al.* (2004) GENECLASS2: a software for genetic assignment and first-generation migrant detection. *The Journal of heredity*, **95**, 536–539.
Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, **12**, 32–42.
Pritchard JK, Pickrell JK, Coop G (2010) The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Current Biology*, **20**, R208–R215.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
Pudovkin AI, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, **144**, 383–387.
Puechmaille SJ (2016) The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: sub-sampling and new estimators alleviate the problem. *Molecular Ecology Resources*, **16**, 608-627.
Pujolar JM, Jacobsen MW, Als TD *et al.* (2014) Genome-wide single-generation signatures

of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514–2528.
Pujolar JM, Jacobsen MW, Frydenberg J *et al.* (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources*, **13**, 706–714.
Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, **2**, e431.
Pusey AE (1987) Sex-biased dispersal and inbreeding avoidance in birds and mammals. *Trends in Ecology & Evolution*, **2**, 295–299.
Qadri SA, Camacho J, Wang H *et al.* (2007) Temperature and acid-base balance in the American lobster *Homarus americanus*. *Journal of Experimental Biology*, **210**, 1245–1254.
Quinn BK, Rochette R (2015) Potential effect of variation in water temperature on development time of American lobster larvae. *ICES Journal of Marine Science: Journal du Conseil*, **72**, i79–i90.
Quinn BK, Sainte-Marie B, Rochette R, Ouellet P (2013) Effect of temperature on development rate of larvae from cold-water American lobster (*Homarus americanus*). *Journal of Crustacean Biology*, **33**, 527–536.
Ranta E, Kaitala V, Lindström J (1999) Sex in space: population dynamic consequences. *Proceedings of the Royal Society of London B: Biological Sciences*, **266**, 1155–1160.
Reiss H, Hoarau G, Dickey Collas M, Wolff WJ (2009) Genetic population structure of marine fish: mismatch between biological and fisheries management units. *Fish and Fisheries*, **10**, 361–395.
Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.
Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–4370.
Ribas L, Robledo D, Gómez-Tato A *et al.* (2015) Comprehensive transcriptomic analysis of the process of gonadal sex differentiation in the turbot (*Scophthalmus maximus*). *Molecular and Cellular Endocrinology*, **422**, 1–18.
Riginos C, Liggins L (2013) Seascape Genetics: Populations, Individuals, and Genes Marooned and Adrift. *Geography Compass*, **7**, 197–216.
Robichaud D, Rose GA (2011) Multiyear homing of Atlantic cod to a spawning ground. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 2325–2329.
Robichaud DA, Campbell A (1991) Annual and seasonal size-frequency changes of trap-caught lobsters (*Homarus americanus*) in the Bay of Fundy. *Journal of Northwest Atlantic Fishery Science*, **11**, 29–37.
Rochette R, Sainte Marie B, Allain M, Baker J et al. (2016) Co-constructed and collaborative research on productivity, stock structure and connectivity in the American lobster *Homarus americanus. Canadian Journal of Animal Science.*
Roesti M, Salzburger W & Berner D (2012). Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, **12**, 1.
Rowe S (2011) Movement and harvesting mortality of American lobsters (*Homarus americanus*) tagged inside and outside no-take reserves in Bonavista Bay, Newfoundland. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 1336–1346.

Ruzzante DE, Taggart CT, Cook D (1999) A review of the evidence for genetic structure of cod (Gadus morhua) populations in the NW Atlantic and population affinities of larval cod off Newfoundland and the Gulf of St. Lawrence. *Fisheries Research*, **43**, 79–97.

Sanford E, Kelly MW (2011) Local Adaptation in Marine Invertebrates. *Annual Review of Marine Science*, **3**, 509–535.

Sato TK, Rehling P, Peterson MR, Emr SD (2000) Class C Vps Protein Complex Regulates Vacuolar SNARE Pairing and Is Required for Vesicle Docking/Fusion. *Molecular cell*, **6**, 661–671.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Schwartz MK, Luikart G & Waples RS. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends in ecology & evolution*, **22**, 25-33.

Schiavina M, Marino IAM, Zane L, Melià P (2014) Matching oceanography and genetics at the basin scale. Seascape connectivity of the Mediterranean shore crab in the Adriatic Sea. *Molecular Ecology*, **23**, 5496–5507.

Schmidt PS, Rand DM (2001) Adaptive maintenance of genetic polymorphism in an intertidal barnacle: habitat- and life-stage-specific survivorship of Mpi genotypes. *evolution*, **55**, 1336–1344.

Selkoe KA, Aloia CCD, Crandall ED *et al.* (2016) A decade of seascape genetics: contributions to basic and applied marine connectivity. *Mar Ecol Prog Ser*, **554**, 1–19.

Selkoe KA, Gaggiotti OE, Bowen BW, Toonen RJ (2014) Emergent patterns of population genetic structure for a coral reef community. *Molecular Ecology*, **23**, 3064–3079.

Selkoe KA, Henzler CM, Gaines SD (2008) Seascape genetics and the spatial ecology of marine populations. *Fish and Fisheries*, **9**, 363–377.

Selkoe KA, Watson JR, White C *et al.* (2010) Taking the chaos out of genetic patchiness: seascape genetics reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species. *Molecular Ecology*, **19**, 3708–3726.

Shafer ABA, Wolf JBW, Alves PC *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, **30**, 78–87.

ShimadaY, Shikano T & Merilä J (2011). A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. *Molecular Biology and Evolution*, **28**, 181-193.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121–132.

Storfer A, Murphy MA, Evans JS *et al.* (2006) Putting the "landscape" in landscape genetics. *Heredity*, **98**, 128–142.

Swearer SE, Caselle JE, Lea DW, Warner RR (1999) Larval retention and recruitment in an island population of a coral-reef fish. *Nature*, **402**, 799.

Templeman W (1940) Embryonic Developmental Rates and Egg-Laying of Canadian Lobsters. *Journal of the Fisheries Research Board of Canada*, **5**, 71–83.

Tepolt CK, Palumbi SR (2015) Transcriptome sequencing reveals both neutral and adaptive genome dynamics in a marine invader. *Molecular Ecology*, **24**, 4145–4158.

Thomas L, Bell JJ (2013) Testing the consistency of connectivity patterns for a widely dispersing marine species. *Heredity*, **111**, 345–354.

Tin MMY, Rheindt FE, Cros E, Mikheyev AS (2015) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, **15**, 329–336.

Tracey ML, Nelson K, Hedgecock D, Shleser RA, Pressick ML (1975) Biochemical Genetics of Lobsters: Genetic Variation and the Structure of American Lobster (*Homarus americanus*) Populations. *Journal of the Fisheries Research Board of Canada*, **32**, 2091–2101.

Turgeon K, Hawkshaw SCF, Dinning KM *et al.* (2016) Enhancing fisheries education in Canada: The need for interdisciplinarity, collaboration, and inclusivity. (No. e2291v1). PeerJ Preprints.

Urrego-Blanco J, Sheng J (2014) Study on subtidal circulation and variability in the Gulf of St. Lawrence, Scotian Shelf, and Gulf of Maine using a nested-grid shelf circulation model. *Ocean Dynamics*, **64**, 385–412.

Valenzuela-Quiñonez F (2016) How fisheries management can benefit from genomics? *Briefings in functional genomics*, elw006–6.

van Heerwaarden J, van Zanten M & Kruijer W (2015). Genome-wide association analysis of adaptation using environmentally predicted traits. *PLoS Genet*, **11**, e1005594.

Vatsiou AI, Bazin E, Gaggiotti OE (2016) Detection of selective sweeps in structured populations: a comparison of recent methods. *Molecular Ecology*, **25**, 89–103.

Véliz D, Bourget E, Bernatchez L (2004) Regional variation in the spatial scale of selection at MPI* and GPI* in the acorn barnacle *Semibalanus balanoides* (Crustacea). *Journal of Evolutionary Biology*, **17**, 953–966.

Villemereuil P, Gaggiotti OE (2015) A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, **6**, 1248–1258.

Villemereuil P, Frichot E, Bazin E, Francois O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

Waddy SL, Aiken DE, De Klejin D (1995) Contro1 of growth and reproduction. *Biology of the American lobster (Homarus americanus)* (ed. Factor JR), pp. 217-266. Academic Press, London.

Wallenfels K, Weil R (1972) 20 β-Galactosidase. In: *The Enzymes*. pp. 617–663. Elsevier.

Wang J (2009) A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Molecular Ecology*, **18**, 2148–2164.

Waples RK, Seeb LW, Seeb JE (2015) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources*, **16**, 17–28.

Waples RS (1998) Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, **89**, 438–450.

Waples RS, Gagiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Waples RS, Punt AE, Cope JM (2008) Integrating genetic data into management of marine resources: how can we do it better? *Fish and Fisheries*, **9**, 423–449.

Watson FL, Miller RJ & Stewart SA (2013). Spatial and temporal variation in size at maturity for female American lobster in Nova Scotia. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 1240-1251.

Weersing K, Toonen RJ (2009) Population genetics, larval dispersal, and connectivity in marine systems. *Mar Ecol Prog Ser*, **393**, 1–12.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *evolution*, **38**, 1358.

White C, Selkoe KA, Watson J *et al.* (2010) Ocean currents help explain population genetic structure. *Proceedings. Biological sciences / The Royal Society*, **277**, 1685–1694.
Whitlock MC, Lotterhos KE (2015) Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST. *The American naturalist*, **186**, S24–36.
Wilkins J (2006) *Macroevolution: Its Definition, Philosophy and History*.
Willette DA, Allendorf FW, Barber PH *et al.* (2014) So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. *Bulletin of Marine Science*, **90**, 79–122.
Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.
Xue H, Incze L, Xu D, Wolff N, Pettigrew N (2008) Connectivity of lobster populations in the coastal Gulf of Maine. *Ecological Modelling*, **210**, 193–211.
Yeomans KA, Golder PA (1982) The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors. *The Statistician*, **31**, 221.
Yu X, Sun S (2013) Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, **14**, 1.
Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells (S-D Zhang, Ed,). *PloS one*, **9**, e78644.