

Un rôle pour la science de l'information géographique en écologie moléculaire: la détection de régions du génome soumises à la sélection naturelle

Stéphane Joost^{*,***} – Aurélie Bonin^{**} – Pierre Taberlet^{**} – Régis Caloz^{***} ¹

**Institut de Zootechnique, Université Catholique du Sacré Cœur, Piacenza, Italie
Stephane.Joost@econogene.eu*

*** Laboratoire d'Ecologie Alpine, Université Joseph-Fourier, Grenoble, France
{Aurelie.Bonin, Pierre.Taberlet}@ujf-grenoble.fr*

**** Laboratoire de Systèmes d'Information Géographique, Ecole Polytechnique Fédérale de Lausanne, Suisse
Regis.Caloz@epfl.ch*

¹ et le Consortium ECONOGENE (<http://www.econogene.eu>)

RÉSUMÉ. Cet article présente une méthode d'analyse spatiale originale basée sur le concept de coïncidence spatiale et sur le calcul simultané de nombreuses régressions logistiques dont le rôle est de détecter des régions du génome soumises à la sélection naturelle. Son application permet d'identifier les mêmes régions que celles mises en évidence par une approche théorique en génomique des populations et, en outre, d'identifier les paramètres environnementaux responsables de la sélection, ce qui était difficilement faisable jusqu'ici. Cette recherche ouvre des perspectives d'applications pour la science de l'information géographique en écologie moléculaire.

ABSTRACT. We introduce a new method to detect signatures of natural selection in the genome based on the application of spatial analysis, with the contribution of Geographical Information Systems, environmental variables, molecular data, and multiple univariate logistic regressions. Its use allows the identification of the same genomic regions as revealed by a theoretical approach in population genomics, but also the uncovering of environmental parameters responsible for selection which was so far difficult. This research also opens up avenues for GIScience research in molecular ecology.

MOTS-CLÉS : analyse spatiale, science de l'information géographique, écologie moléculaire, génomique environnementale, sélection naturelle.

KEYWORDS: spatial analysis, GIScience, molecular ecology, landscape genomics, natural selection.

1. Introduction

Au cours de son voyage sur le Beagle, Darwin avait remarqué la différence de taille entre les becs des pinsons qui vivaient sur les différentes îles des Galapagos (Darwin, 1839). Cela le conduisit à la conclusion qu'au fil des générations, les oiseaux évoluent de manière à correspondre à l'environnement dans lequel ils vivent. Les becs des pinsons sont différents parce que la sélection naturelle couplée à l'isolement géographique les a forcés à évoluer de façon à ce qu'ils s'adaptent aux particularités respectives des différentes îles. Comme cela a été démontré dans de nombreux cas, l'isolement géographique est à même de déclencher des processus évolutifs divergents, puis la spéciation correspondante (de Duve, 2005). L'exemple fameux des pinsons de Darwin met en évidence le fait que les organismes sont directement influencés par les caractéristiques de leur environnement, via les processus génétiques comme cela sera découvert plus tard grâce notamment aux travaux de Thomas Morgan sur la Drosophile (en 1910), et de Ronald Fisher, John B.S. Haldane et Sewall Wright qui ont permis de proposer dans les années 1940 une première théorie synthétique de l'évolution.

Dans son cadre naturel, l'information génétique est scellée dans un contexte géographique. Par conséquent, l'information spatiale est un élément dont il faut tenir compte lorsque l'on essaie d'analyser les ressources génétiques existantes et de comprendre les mécanismes évolutifs qui les ont façonnées. La démarche que nous présentons ici met clairement l'accent sur l'information géographique et montre comment cette dernière peut fournir aux sciences de la vie des moyens alternatifs pour exploiter les données produites en biologie moléculaire. Beaucoup de projets de recherche que ce soit sur le séquençage du génome¹ (humain, du poulet, etc.), en génétique des populations ou en biologie de la conservation, participent à la collecte de grandes quantités d'échantillons biologiques, dont une part croissante est géographiquement référencée. Les techniques avancées en biologie moléculaire rendent possibles l'extraction et l'analyse d'énormes quantités de données génétiques à partir de ces échantillons biologiques. Dans ce contexte, il est souhaitable que différents types d'approches interdisciplinaires soient exploités de manière à aborder cette information si complexe sous différents angles, de manière à procurer des moyens complémentaires pour la comprendre (Woese, 2004). La science de l'information géographique est justement une approche qui aborde l'information génétique d'un point de vue original, susceptible de répondre à cette demande.

Dans cet article, nous présentons une méthode d'analyse spatiale développée dans le but d'identifier des régions dans les chromosomes qui sont potentiellement impliquées dans des processus d'adaptation à l'environnement local. Cette présentation est illustrée par les résultats de trois études, respectivement sur la

¹ Ensemble du matériel génétique (ADN) d'un organisme. Les gènes ne constituent qu'une petite partie du génome total, le reste étant de l'ADN dit « non-codant ».

grenouille rousse, le grand charançon du pin et le mouton, initialement réalisées dans le cadre d'autres travaux.

La mise en évidence de signatures de sélection naturelle au sein du génome est fondamentale dans la mesure où elle peut permettre de mieux comprendre quels sont les gènes qui sont impliqués dans l'adaptation à tel ou tel facteur environnemental. La compréhension de ces processus est au centre des préoccupations des chercheurs en biologie de l'évolution (Storz, 2005). En outre, la connaissance de ces régions particulières du génome peut être exploitée en médecine, en élevage d'animaux domestiques (sélection naturelle et artificielle), ou en biologie de la conservation.

Cette contribution de la science de l'information géographique à la recherche menée dans ce domaine des sciences de la vie constitue un progrès remarquable dans la mesure où elle permet d'identifier des facteurs environnementaux potentiellement responsables de la sélection, ce qui était difficilement faisable jusqu'ici en utilisant les méthodes à disposition en génétique des populations.

2. Génétique et information géographique

La génétique est une discipline dont on parle beaucoup dans la société notamment en raison des grands débats publics à propos des organismes génétiquement modifiés (OGM) ou des thérapies géniques par exemple. Malgré cela, l'information génétique n'a été jusqu'ici que rarement analysée par les géographes ou les chercheurs en science de l'information géographique, alors que les travaux des généticiens des populations et des écologistes moléculaires ont permis de mettre au point des approches spatiales innovantes.

Dès 1920 environ, les fondateurs de la génétique des populations (Fisher, Haldane et Wright) ont pris en compte la distribution spatiale des populations étudiées et l'ont intégrée dans la plupart des modèles qu'ils ont alors établis (Epperson, 2003). La distance géographique entre les populations ou les habitats est un paramètre majeur des deux modèles les plus connus en génétique des populations, soit l'isolement génétique par la distance (« Genetic isolation by distance ») et le modèle infini en îles (« Infinite-island ») (Epperson, 2003; MacArthur et Wilson, 2001).

Selon Luigi Cavalli-Sforza, généticien des populations italo-américain, Arthur Mourant, un généticien anglais, fut le premier à représenter des fréquences de gènes sur des cartes géographiques et à exploiter pleinement ces dernières (Cavalli-Sforza et al., 1994). Mourant réalisa en effet l'une des premières études de la géographie des gènes en proposant que les Basques étaient les plus anciens habitants d'Europe et qu'ils avaient conservé une partie de leur constitution génétique ancestrale malgré leurs contacts avec les immigrants ultérieurs (Mourant, 1954).

Probablement influencé par cette lecture, Luigi Cavalli-Sforza eut l'idée dans les années 1950 de représenter sur des cartes géographiques la distribution spatiale de la

variation de la fréquence de gènes humains à travers le monde. L'application de cette idée a conduit à la rédaction de «The History and Geography of Human Genes» écrit avec ses collègues Paolo Menozzi et Alberto Piazza, et publié en 1994 (Cavalli-Sforza et al., 1994). L'ouvrage propose une interprétation des trajets qu'ont pu suivre les populations humaines pour coloniser la planète à partir de l'Afrique. Des anciennes routes de migration sont également mises en évidence, comme celle des éleveurs du Néolithique depuis le Moyen-Orient et la région du Croissant Fertile jusqu'en Europe occidentale. Ces travaux de Cavalli-Sforza et de son équipe sont les plus achevés concernant l'exploitation systématique de coordonnées géographiques dans le but de représenter des données génétiques et de les interpréter. Seule la notion de Système d'Information Géographique n'est pas présente, alors que des méthodes avancées d'analyse spatiale comme la variographie et la régionalisation sont utilisées pour interpoler ou « lisser » les surfaces entre les points d'échantillonnage.

Ces travaux, ainsi que les progrès effectués dans le domaine de la biologie moléculaire, du perfectionnement des logiciels de statistiques et de l'augmentation continue de la puissance des ordinateurs, ont largement contribué à l'émergence progressive de la génétique environnementale (*Landscape Genetics*) dès la fin des années 1980. Dans un premier temps, le président de la British Ecological Society intervient en 1989 pour reconnaître l'importance de l'étude des interactions entre l'environnement et le génome afin de mieux comprendre l'évolution (Berry, 1989). Puis des études sont menées au milieu des années 1990 pour tenter d'exploiter l'information génétique dans un contexte spatial avec l'aide des Systèmes d'Information Géographiques (Dave Galbraith, Royal Botanical Gardens²), ou pour relier la génétique et l'écologie (Jelinski, 1997), mais sans pour autant parvenir à faire une synthèse complète et satisfaisante de ce nouveau domaine. C'est finalement en 2003, avec un article intitulé «Landscape genetics: combining landscape ecology and population genetics» que Stéphanie Manel, Michael Schwartz, Gordon Luikart et Pierre Taberlet (Manel et al., 2003) parviennent finalement à clairement définir la discipline. En identifiant ses racines dans les travaux précurseurs effectués par de Candolle (1778-1841) et par Wallace (1823-1913), ils décrivent la génétique environnementale comme étant une combinaison de génétique des populations et d'écologie spatiale. Le but de ce domaine de recherche est de comprendre comment les caractéristiques géographiques et environnementales structurent la variabilité génétique tant au niveau des populations que des individus. Cette discipline étudie également les effets produits par les pressions environnementales, leurs implications en écologie et dans les processus évolutifs, afin de perfectionner les méthodes utilisées en biologie de la conservation.

Les travaux les plus récents issus de ces développements en génétique environnementale tirent profit de la récente évolution de certaines techniques en génomique. Grâce aux progrès des biotechnologies, il est désormais possible d'analyser simultanément de très grandes quantités d'information génétique. Alors qu'il y a quelques années on devait se satisfaire de l'étude de quelques marqueurs

² <http://web.nrdpfc.ca/landscapegenetics.html>

moléculaires³ dans le génome, ce sont maintenant des milliers de marqueurs qui peuvent être détectés en très peu d'opérations et à moindres frais. Ces grands jeux de données moléculaires peuvent être exploités en analyse spatiale et comparés à des paramètres environnementaux dans le but de déterminer les rôles respectifs des grandes forces évolutives (notamment la dérive génétique, la migration et la sélection) dans le façonnement de la variabilité du génome. On parlera alors de génomique environnementale (*Landscape Genomics*, Joost et al., 2007). Cette approche est présentée dans le présent article, avec comme objectif premier d'identifier les régions du génome soumises à la sélection naturelle.

3. Approche, données et méthode

3.1 Des modèles simples

L'analyse de processus ou de phénomènes dans le cadre des sciences naturelles implique des contraintes dans le choix du type de modélisation envisagé. En effet, les paradigmes en sciences expérimentales sont plus universels et moins sujets aux effets locaux que ceux des sciences naturelles (Rose, 2003). Leur but est de tenter de comprendre des processus élémentaires contrôlés par quelques paramètres. Les problèmes auxquels des disciplines comme la physique sont confrontées sont très difficiles, ont un niveau d'abstraction très élevé, et de plus ils peuvent difficilement être isolés de leur contexte. Au contraire, la complexité en sciences naturelles vient du fait que de nombreux paramètres interviennent et interagissent simultanément, qu'ils ne sont pas stables et que leurs propriétés ne sont pas linéaires. En outre, comme la plupart des propriétés des modèles étudiés ne peuvent pas être quantifiés, il est très difficile de traduire les processus naturels en formules mathématiques. Des tentatives dans le but de quantifier absolument certains phénomènes naturels peuvent d'ailleurs produire des mystifications (Rose, 2003). De toute évidence, il est rarement possible de séparer l'effet d'un paramètre parmi d'autres et de le mesurer en négligeant toutes les interactions dans lesquelles il intervient.

Pour tenter d'identifier des régions du génome potentiellement sélectionnées par des paramètres environnementaux, on pourrait être tenté d'être le plus exhaustif possible et de mettre en évidence la combinaison de tous les facteurs entrant dans un processus de sélection. En mettant en œuvre de tels modèles complexes avec de nombreux paramètres dans le but d'augmenter la part de variance expliquée, on va également produire de l'incertitude. Celle-ci découle de spécificités de l'analyse quantitative en sciences naturelles, et son étude constitue un domaine de recherche en soi. L'incertitude, c'est-à-dire l'erreur, l'imprécision, l'ambiguïté, etc., est traquée et étudiée par beaucoup de personnes dans divers domaines de recherche, dont l'information géographique. L'incertitude est mesurée et ses causes sont identifiées dans le but d'améliorer les techniques et de rédiger des recommandations

³ Fragments spécifiques d'ADN pouvant être étudiés au sein du génome.

pour la réduire. Dans un article sur les limites de la connaissance géographique, Couclelis (2003) montre que l'erreur dans le domaine de la géoinformation est inévitable et ceci pas uniquement en raison d'imprécisions générées par l'homme. Elle ne concentre pas son analyse sur la qualité des données, mais porte son intérêt sur la qualité de la connaissance que cette information permet de produire pour démontrer, en s'appuyant sur différents exemples, qu'il y a beaucoup de choses à propos de l'information que nous ne pouvons pas savoir, et qui ne résultent ni d'imperfections ni de faits empiriques ou de limitations humaines. Ces exemples de limites intrinsèques à la connaissance sont bien connus en sciences expérimentales, notamment en physique. On sait par exemple depuis Newton, puis Poincaré (1910), qu'il n'y a pas de solution analytique aux équations gravitationnelles qui décrivent la dynamique de n corps, quand $n > 2$. Par conséquent, même s'il on investissait davantage de temps et de moyens, on ne parviendrait probablement pas à réduire l'incertitude au-delà d'un certain seuil approximatif. Caloz (2005) montre dans un article dédié à la propagation de l'incertitude en analyse spatiale qu'il est souvent impossible de quantifier cette incertitude, et que l'on devrait plutôt à la place recourir à une évaluation qualitative de la fiabilité des mesures effectuées.

Ces raisons nous ont amenés à mettre en œuvre des modèles simples de manière à mettre en évidence les paramètres environnementaux individuels qui sélectionnent potentiellement les régions analysées du génome. Ceci permet d'une part de pouvoir gérer une incertitude la plus réduite possible et simultanément à laisser ensuite, à partir de cette indication initiale, toute la place à l'expert et à son interprétation (Beven, 2002).

3.2. Données

Deux catégories de données sont nécessaires afin de mettre en œuvre la méthode proposée, soit de l'information génétique et des variables environnementales. Les données moléculaires utilisées sont des fragments spécifiques d'ADN, de séquence nucléotidique connue ou non, qui sont produits par des biotechnologies et appelés *marqueurs* génétiques. Ces marqueurs sont utilisés comme points de repère pour étudier le génome. Ce ne sont pas forcément des gènes, mais ils peuvent être statistiquement associés à des gènes recherchés. A chaque marqueur, tout individu étudié possède un profil génétique particulier appelé le génotype, qui est composé d'un couple d'allèles⁴. Un génotype comportant deux fois le même allèle est dit homozygote, tandis qu'un génotype avec deux allèles différents est dit hétérozygote. Parfois, quand on cherche à déterminer le génotype d'un individu, celui-ci n'est pas accessible car certains allèles cachent la présence d'autres allèles. On a alors accès

⁴ L'analyse d'un fragment d'ADN à un locus donné – un emplacement dans le génome - et dans une population donnée, peut révéler plusieurs variants. Ces différents variants à un locus sont appelés « allèles ».

uniquement au phénotype⁵. Les allèles ayant le pouvoir d'en masquer d'autres sont dits dominants, et ceux qui sont masqués sont dits récessifs. Au sein d'une population, la distribution des génotypes évolue sous l'effet des grandes forces évolutives comme la sélection naturelle. Il existe plusieurs manières de détecter des marqueurs moléculaires au sein du génome (Avisé, 2004). Pour les cas d'étude présentés dans cet article, le génome a été scanné soit avec l'aide de marqueurs AFLPs (Amplified Fragment Length Polymorphisms ; Ajmone Marsan et al., 1997) pour la grenouille rousse et le grand charançon du pin, soit avec des marqueurs microsatellites pour les races de mouton (Avisé, 2004). Les marqueurs AFLP possèdent deux allèles différents, l'allèle « présence de bande » étant dominant sur l'allèle « absence de bande ». Pour ces marqueurs, un phénotype « présence d'une bande » est donc l'expression soit d'un génotype homozygote pour l'allèle « présence de bande », soit d'un génotype hétérozygote et il est impossible de connaître le génotype exact à partir du phénotype sans information supplémentaire. Quant aux marqueurs microsatellites, ils possèdent en général plusieurs allèles sans relation de dominance ou de récessivité entre eux. Le génotype exact peut donc être identifié à partir du phénotype, ce qui fait que les microsatellites sont plus informatifs que les AFLP. Cependant, ce type de marqueurs est généralement disponible en grand nombre à peu de frais, ce qui vient compenser l'information limitée fournie individuellement par chaque marqueur AFLP.

Variable	Description
Altitude	Déterminée avec l'aide du MNA SRTM30 de la NASA
DTR	Moyenne de l'amplitude thermique diurne en °C
FRS	Nombre de jours avec sol gelé
PR	Précipitations en mm/mois
PRCV	Coefficient de variation des précipitations mensuelles en %
REH	Humidité relative (pourcentage)
SUN	Pourcentage de l'ensoleillement maximum possible
TMP	Température moyenne en °C
WET	Nombre de jours par mois avec plus de 0.1 mm de pluie
WIND	Vitesse du vent en m/s, 10 mètres au-dessus du sol

Table 1. Liste des variables environnementales utilisées pour le grand charançon du pin et pour les races de moutons ECONOGENE. Pour chaque variable, on a utilisé la moyenne annuelle (charançon et mouton) ou les valeurs mensuelles (mouton).

Les jeux de données moléculaires utilisés se présentent sous la forme de matrices. Chaque ligne de la matrice correspond à un individu échantillonné, et les

⁵ Le phénotype est la traduction du génotype en caractère observable.

colonnes contiennent les coordonnées géographiques, puis pour chaque marqueur un 1 ou un 0. Pour les marqueurs AFLP, les chiffres 1 et 0 signalent respectivement les phénotypes « présence de bande » et « absence de bande ». Pour les marqueurs microsattellites, les chiffres 1 et 0 signalent respectivement la présence ou l'absence d'un allèle donné au locus.

Les données environnementales sont composées de l'altitude et de données climatiques. L'altitude a été estimée avec l'aide du modèle numérique d'altitude SRTM30 (Shuttle Radar Topography Mission) de la NASA⁶, dont la résolution est de 30 secondes d'arc, excepté pour l'étude sur la grenouille rousse où elle a été enregistrée avec l'aide d'un altimètre. Les données climatiques décrites à la table 1 sont constituées de grilles de 10 minutes de résolution (environ 12 km à la latitude de la Suisse) et contiennent 9 variables mensuelles ainsi qu'une moyenne annuelle. Ces variables caractérisent des régions continentales pour la période allant de 1961 à 1990 (New et al., 2002). Ces données ont été collectées par la Climatic Research Unit (CRU) à Norwich.

Dans le cas du grand charançon du pin, seules les moyennes mensuelles ont été utilisées, alors que dans l'exemple des moutons on a eu recours à l'ensemble des 118 variables, ceci parce que des systèmes de production différents selon les races sont basés sur les périodes ou saisons de mise à bas (Autumn lamb production, Winter lambing, ou Spring lamb production).

3.3 Méthode d'Analyse Spatiale (SAM)

La méthode SAM décrite en détail par Joost et al. (2007) repose sur la coïncidence spatiale, l'un des six concepts d'analyse spatiale distingués par Goodchild (1996). Celle-ci met en relation les caractéristiques génétiques des organismes étudiés avec des valeurs de paramètres environnementaux grâce à des coordonnées géographiques communes. Pour fonctionner, la SAM utilise les deux jeux de données géoréférencés décrits plus haut, l'un qui contient la matrice de 1 et de 0 pour les organismes étudiés, et l'autre les valeurs d'une ou de plusieurs variables environnementales. La régression logistique univariée fournit la mesure du niveau d'association entre la fréquence d'un phénotype AFLP ou d'un génotype microsattellite et les valeurs des paramètres environnementaux. On calcule la significativité des modèles constitués par toutes les paires possibles [marqueur versus paramètre environnemental] afin d'identifier les marqueurs impliqués dans les modèles les plus significatifs : ces marqueurs sont localisés dans des régions du génome qui jouent probablement un rôle dans les processus d'adaptation.

Deux tests statistiques ont été appliqués afin de vérifier la significativité des modèles, c'est-à-dire de vérifier si un modèle qui inclut une variable environnementale explique plus de variance qu'un modèle avec une constante

⁶ <http://www2.jpl.nasa.gov/srtm/>

seulement. Nous avons suivi Hosmer et Lemeshow (2000) qui recommandent l'utilisation du test du taux de vraisemblance (Likelihood ratio ou statistique G) et du test de Wald.

a) la statistique G, soit
$$G = -2 \ln \frac{L}{L'}$$

où L est la vraisemblance du modèle initial avec une constante uniquement, et L' la vraisemblance du nouveau modèle qui inclut la variable étudiée. Si le paramètre ajouté est égal à zéro, cette statistique sera conforme à une distribution du Khi carré, avec un nombre de degrés de liberté égal au nombre de paramètres ajoutés (Hosmer et Lemeshow 2000).

b) la statistique de Wald, soit
$$W = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}$$

où β est le maximum de vraisemblance pour le paramètre β_i ; $\hat{\beta}_i$ est l'estimation du maximum de vraisemblance pour le paramètre β_i , et $\sigma(\hat{\beta}_i)$ est une estimation de son écart-type. Si l'hypothèse nulle est vérifiée, le rapport calculé suit une distribution normale standard, et la méthode qui permet de calculer la variance est conforme à la théorie du maximum de vraisemblance (Hosmer et Lemeshow, 2000). Pour les deux tests statistiques décrits ci-dessus, l'hypothèse nulle stipule qu'un modèle avec la variable analysée ne permet pas d'expliquer plus de variance qu'un modèle composé d'une constante uniquement.

Dans les analyses présentées, un modèle est considéré comme significatif uniquement si les deux tests rejettent l'hypothèse nulle. Cette précaution a été prise en raison de l'existence de conclusions contradictoires sur la fiabilité des tests statistiques en général en régression logistique (Hosmer et Lemeshow, 2000) et sur celle des deux tests utilisés (Joost et al., 2007).

Comme les jeux de données moléculaires peuvent contenir beaucoup de marqueurs et que de nombreux paramètres environnementaux peuvent être utilisés pour caractériser les sites d'échantillonnage, un nombre important de modèles univariés est calculé simultanément dans le but de détecter les marqueurs potentiellement soumis à la sélection naturelle (plus de 80'000 dans le cas du mouton). Dans ce contexte de vérification d'hypothèses multiples, la probabilité de rejeter au moins une des hypothèses nulles pour un seuil de significativité α alors qu'elle est vraie (erreur de type I) est bien plus élevée que α . Parmi les différentes méthodes qui permettent de tenir compte de ce risque et de le corriger, nous avons choisi d'appliquer la correction de Bonferroni (Shaffer, 1995). Elle consiste simplement à diviser le seuil de significativité désiré α par le nombre de comparaisons effectuées (en l'occurrence, le nombre de modèles qui sont calculés simultanément). Cette correction est connue pour être très conservatrice (Narum, 2006), ce qui nous permet dans le cas présent de limiter le nombre de modèles significatifs de manière à restreindre notre analyse aux associations les plus robustes.

Le processus de calcul a été automatisé dans un programme Matlab® appelé « SAM » qui recourt à la fonction GLMfit - Generalized Linear Model Fitting (MacCullagh et Nelder, 1989). La procédure résout les équations de vraisemblance afin d'estimer les paramètres, calcule les p valeurs associées aux tests statistiques (Wald et G), génère les graphes des courbes de probabilité, et exporte les résultats. Une macro Excel complémentaire a été développée afin de traiter les matrices de résultats dans des tables dynamiques qui permettent d'ajuster facilement le niveau de significativité et de sélectionner les marqueurs intéressants (tables de réjection).

3.4 Génomique des populations

Afin de permettre une comparaison et de valider les résultats produits par la méthode SAM, une approche théorique en génomique des populations a été systématiquement appliquée aux mêmes jeux de données.

Diverses méthodes ont été mises au point dans le but de mettre en évidence les régions du génome potentiellement soumises à la sélection naturelle (voir Bonin, 2006a, et les références qui y sont mentionnées). Certaines d'entre elles appartiennent à l'approche appelée « Gène candidat » qui considère un locus spécifique comme point de départ et cherche à déterminer s'il est impliqué dans un processus de sélection ou non (Phillips, 2005) en utilisant divers tests de neutralité (Nielsen, 2005). Une autre catégorie de méthodes a pour but d'identifier les régions chromosomiques (QTL⁷) impliquées dans le déterminisme génétique des caractères phénotypiques complexes comme la couleur du pelage, la taille ou tout autre trait adaptatif, en mesurant le degré d'association entre le génotype à un marqueur moléculaire et la valeur phénotypique du caractère en question (Mackay, 2001).

Mais il existe actuellement une stratégie alternative qui permet de préciser les bases génétiques de l'adaptation locale et qui combine les avantages des deux approches décrites plus haut. La génomique des populations consiste en l'étude simultanée de nombreux marqueurs génétiques afin de préciser les rôles respectifs des grandes forces évolutives dans le façonnage de la variabilité du génome (Luikart et al., 2003). Elle ne requiert pas de connaissance approfondie sur le génome de l'organisme étudié. Les forces évolutives qui s'appliquent sur le génome sont de deux types (Luikart et al., 2003). Certaines, comme la dérive génétique ou la migration, influencent tous les locus du génome sans exception, neutres ou non-neutres. D'autres, comme la sélection naturelle, n'agissent que sur certains locus en particulier. La plupart des locus étant neutres, il existe donc dans le génome une variabilité génétique de base imprimée par les forces évolutives globales, tandis que quelques locus ont une variabilité atypique due aux forces spécifiques (Black et al., 2001). Il est alors possible de dégager ces locus singuliers par rapport au reste du génome par comparaison des niveaux de diversité génétique. C'est sur ce principe que s'appuie la génomique des populations quand elle cherche à élucider les bases

⁷ Pour Quantitative Trait Locus.

génétiques de l'adaptation locale (Storz, 2005). En pratique, la génomique des populations a donc recours à des criblages multilocus, c'est-à-dire au génotypage⁸ de nombreux marqueurs dans le génome de plusieurs individus provenant de populations différentes.

Afin d'évaluer les résultats produits par la méthode SAM, les mêmes jeux de données ont été analysés par DFDIST et FDIST2⁹, deux logiciels développés sur la base de l'approche décrite ci-dessus, et basés sur le programme FDIST de Beaumont et Nichols (1996). La méthode FDIST s'appuie sur le principe de différenciation génétique entre populations qui est habituellement exprimé par l'indice FST (Neigel, 2002). Dans un modèle comprenant un nombre infini de populations de tailles constantes échangeant des migrants avec un taux constant (modèle infini en îles), le FST introduit par Wright (1951) correspond à la probabilité que deux allèles tirés au hasard dans une population proviennent d'un même allèle ancestral issu lui-même de cette population (Beaumont, 2005). L'approche FDIST est une variante du test de Lewontin-Krakauer (Lewontin et Krakauer, 1973) qui repose sur l'idée que dans une population, tous les locus du génome ont des FST comparables parce qu'ils partagent la même histoire démographique. Seuls les locus sous sélection dérogent à cette règle, car la sélection naturelle modifie leur FST (Beaumont, 2005).

4. Cas d'étude

La nouvelle méthode d'analyse spatiale SAM a été appliquée à un insecte, à une plante, ainsi qu'à des espèces animales sauvages et domestiques dans des contextes de conservation des ressources génétiques. Ci-dessous, nous présentons un cas d'étude consacré à la grenouille rousse pour lequel les données analysées ont été produites par le Laboratoire d'Ecologie Alpine (LECA) à l'Université Joseph Fourier de Grenoble. C'est le même laboratoire qui a fourni les données pour le deuxième exemple, le grand charançon du pin. Et dans le cas des espèces domestiques, le projet de recherche européen ECONOGENE (www.econogene.eu) a fourni des données relatives à des races de moutons échantillonnées dans une vaste surface géographique comprise entre le Portugal à l'Ouest, l'Ecosse au Nord, l'Égypte au Sud, et la Turquie orientale à l'Est (Ajmone Marsan, 2005; Joost, 2006).

La grenouille rousse présente plusieurs caractéristiques la rendant particulièrement susceptible de développer une différenciation adaptative. Ainsi, elle est soumise à des conditions environnementales très diverses à travers son aire de répartition très vaste, et les populations sont généralement assez structurées localement. Par conséquent, il s'agit d'un bon modèle pour étudier l'adaptation locale (Bonin, 2006a). Dans le cas présent, c'est l'adaptation de l'espèce à l'altitude qui a été analysée. Des populations ont été échantillonnées dans les Alpes du Nord

⁸ Détermination du génotype d'un individu à un ou plusieurs marqueurs moléculaires.

⁹ DFDIST est conçu pour analyser des marqueurs de type dominant comme les AFLP, et FDIST pour analyser des marqueurs codominants comme les microsatellites.

en France à trois altitudes différentes: basse (à Saint-Rémy-de-Maurienne et Cognin, environ 400 m), moyenne (Col de Plainpalais et Les Tines, environ 1100 m) et haute (Lac des Aiguillettes, Lac des Tempêtes, environ 2100 m). Au total, 190 individus ont été échantillonnés et 392 marqueurs AFLP ont été développés (Bonin et al., 2005), parmi lesquels 364 ont été utilisés avec la méthode SAM (Bonin, 2006a).

Dans le deuxième exemple, c'est la relation entre un insecte phytophage avec sa plante hôte qui est étudiée. De fortes pressions de sélection s'exercent sur l'herbivore et le conduisent à se spécialiser. L'insecte développe entre autres des adaptations pour localiser, atteindre et exploiter la ressource végétale. Le grand charançon du pin (*Hylobius abietis* L.) fait partie des insectes dont la biologie et la dynamique de population sont déterminées par un facteur qui tend à n'exister qu'en quantité limitée dans l'espace et dans le temps puisque ses larves se développent sur des conifères mourants. Mais depuis quelques temps, les méthodes modernes d'exploitation forestière ont changé ces conditions à une large échelle en Europe en accroissant le nombre de sites de ponte et de nourriture. Plusieurs facteurs susceptibles de rendre compte de l'adaptation du grand charançon du pin à son environnement forestier ont été testés par Conord (2006a) et Conord et al. (2006b), soit l'influence de la géographie et de la plante hôte sur la structuration des populations. Elles ont été complétées par les analyses effectuées avec la méthode SAM que nous présentons ci-dessous dans le but de déterminer les facteurs environnementaux qui représentent des forces potentielles de sélection agissant sur le génome à une large échelle géographique (Joost et al., 2007). En effet, la température, les précipitations, le nombre de jours de gel, ou la vitesse du vent peuvent avoir un impact direct sur la survie des larves et des adultes, et un impact moindre sur les qualités de la plante hôte. Les charançons (larves et adultes) ont été échantillonnés dans vingt forêts situées en Estonie, en Finlande, en Pologne, en France et en Irlande. Quatre-vingt-trois marqueurs AFLP ont été utilisés (Conord et al., 2006b).

Enfin, le troisième exemple illustre l'application de la méthode SAM aux animaux d'élevage. Les techniques d'élevage intensif appliquées dans les systèmes agricoles ont une influence néfaste sur la biodiversité des animaux domestiques et entraînent des effets génétiques négatifs (dérive génétique, consanguinité). Une des mesures mise en œuvre afin de juguler ce phénomène d'érosion génétique est la conservation des ressources génétiques des animaux d'élevage : on va chercher à déterminer quelles sont les races les plus rares et les plus menacées afin de les protéger, quelles sont les races les mieux adaptées à leur environnement et pourquoi, etc. Chez le mouton, la recherche de signatures de sélection est susceptible de favoriser la découverte de régions génomiques impliquées dans des processus d'adaptation comme la résistance à des parasites ou à des maladies (Joost et al., 2007). Dans ce cadre, le projet de recherche européen ECONOGENE a échantillonné 1748 moutons appartenant à 57 races (56 autochtones, 1 cosmopolite) dans divers pays d'Europe et du Moyen-Orient de manière à produire différents jeux de données moléculaires. Trente et un marqueurs de type microsatellite, pour lesquels on a pu distinguer 744 allèles différents (Peter et al., 2007) ont été utilisés.

5. Résultats

Pour la grenouille rousse, parmi les 364 marqueurs AFLP analysés, la méthode SAM en a identifié 46 comme étant significativement associés avec l'altitude à un seuil de significativité de 2.74×10^{-5} , soit un niveau de confiance (NC) de 99%, correction de Bonferroni incluse, ceci avec au moins l'un des deux tests. Au même NC, seuls 20 marqueurs sont significatifs avec les deux tests statistiques. Le marqueur **301** ressort très fortement puisque tous les tests sont très significatifs (NC= 2.74×10^{-11}). Ensuite, dans un ordre dégressif de significativité, viennent le 320 (NC= 2.74×10^{-10}), le 214 (NC= 2.74×10^{-9}), le 337 et le 357 (NC= 2.74×10^{-8}), le 180, le **84**, le 328 et le 179 (NC= 2.74×10^{-7}), le 385, le 233 et le 354 (NC= 2.74×10^{-6}), le 62, le 390, le 58, le 248, le 3, le 271, le 228 et le 265 (NC= 2.74×10^{-5}).

Les analyses avec DFDIST ont été effectuées sur les données de populations regroupées par catégorie d'altitude. Seuls deux marqueurs, le **301** et le **84**, ont été détectés à un niveau de confiance de 99%. Tous deux ont une probabilité élevée d'être sélectionnés par l'altitude (figure 1) puisqu'ils sont également détectés par la méthode SAM. Avec un NC de 95%, 4 autres marqueurs (97, 228, 250 et 388) sont mis en évidence par DFDIST. Ces derniers sont aussi repérés par la SAM ; toutefois les marqueurs 97 et 388 ne le sont que par le test G (figure 4).

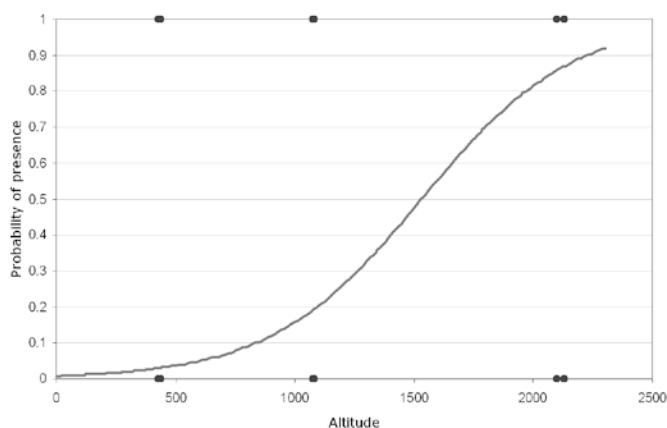


Figure 1. La probabilité du phénotype « présence de bande » au marqueur 301 augmente avec l'altitude pour les grenouilles rousses échantillonnées.

En ce qui concerne le grand charançon du pin, la méthode SAM a identifié 11 marqueurs significativement associés avec des paramètres environnementaux pour les deux tests, avec un seuil de significativité de 1.20×10^{-5} , ce qui correspond à un niveau de confiance de 99%, correction de Bonferroni incluse. Trois marqueurs (**38**, **52** and **63**) sont encore très significativement associés pour un seuil de significativité de 1.20×10^{-13} . Le marqueur 38 est associé avec l'amplitude diurne de température (DTR), le pourcentage d'ensoleillement maximum (SUN), et le nombre jours

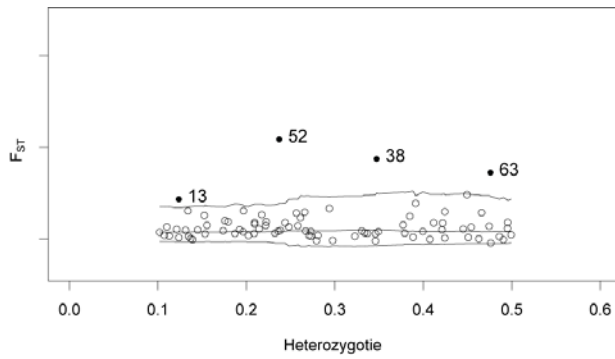


Figure 2. Chez le grand charançon du pin, les quatre marqueurs situés au-dessus de la ligne – limite supérieure des 95% d'une distribution de marqueurs neutres simulée – sont détectés et potentiellement sous sélection.

Chez le mouton, les associations les plus significatives mises en évidence par la SAM, avec les deux tests et pour un seuil de significativité de 1.13×10^{-17} (le niveau de confiance de 99% correspond à 1.13×10^{-7}), impliquent 5 allèles localisés sur 4 marqueurs : SRCRSP9 (2 allèles), DYMS1 (1 allèle), ILSTS28 (1 allèle) et OARFCB304 (1 allèle). Seul le dernier a également été détecté par FDIST2 à un niveau de confiance de 99% (figure 3). Enfin, à un niveau de significativité inférieur pour la SAM (1.13×10^{-15}), le locus OARJMP29 est détecté par les deux méthodes.

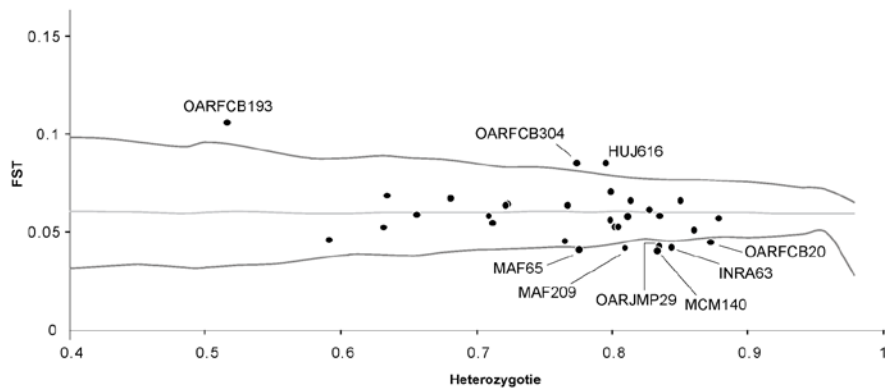


Figure 3. Résultats de FDIST2 pour les races de moutons ECONOGENE. Les marqueurs situés au-dessus et au-dessous des lignes gris foncé - qui délimitent 99% d'une distribution neutre simulée – sont détectés et potentiellement soumis à la sélection naturelle.

En plus de ces premiers enseignements sur la sensibilité de la méthode, les résultats obtenus avec la régression logistique renforcent les bons candidats déterminés avec les tests classiques de génétique des populations. Cette analyse montre que le comportement atypique de certains marqueurs peut être établi par deux méthodes indépendantes fondées sur des hypothèses radicalement différentes.

Dans le deuxième exemple, les résultats obtenus pour le grand charançon du pin montrent une exacte correspondance entre les deux méthodes, avec un seuil de significativité de 95% pour DFDIST. Ils mettent surtout en évidence deux marqueurs (38 et 52) dont la fréquence varie de manière opposée par rapport à l'amplitude thermique diurne (DTR). Une grande amplitude thermique signale généralement un climat continental sévère qui semble exercer une pression de sélection sur deux régions indépendantes du génome du charançon. De plus, la fréquence du marqueur 38 est également significativement associée avec le nombre de jours de gel du sol (FRS), et elle diminue avec l'augmentation des précipitations. Ces informations viennent confirmer des travaux qui démontrent que la tolérance au froid et à la sécheresse chez les insectes est contrôlée par des processus physiologiques similaires (Sinclair et al., 2003), avec, en plus, l'avantage de pointer sur les régions génomiques impliquées.

Chez le mouton, grâce notamment à l'utilisation des microsatellites qui sont utilisés comme marqueurs moléculaires depuis plus longtemps que les AFLP, et qui sont mieux documentés, les résultats obtenus sont plus spectaculaires. Par exemple, OARJMP29 est un locus, détecté par les deux méthodes, qui est associé au nombre de jours de pluie (WET). Il avait déjà été identifié comme étant lié à la résistance à un nématode gastro-intestinal (Beh et al., 2002). D'autre part, parmi les marqueurs détectés par la SAM uniquement, DYMS1 – également associé avec le nombre de jours de pluie – est connu pour être impliqué dans un processus de résistance à un parasite chez la Scottish Blackface (Buitkamp et al., 1996). Cette race vit justement dans l'un des environnements les plus pluvieux observés dans cette étude, mais cela ne permet pas encore de tirer de conclusion sur la nature du lien entre humidité du climat et résistance à un parasite.

De manière générale, il est important de garder en mémoire que la SAM détecte de simples associations, ce qui n'indique en rien que le ou les paramètres environnementaux mis en lumière influencent réellement le ou les marqueurs concernés, et encore moins que leur éventuelle action est véritablement de nature sélective.

Les résultats obtenus chez le mouton, déjà documentés, permettent de valider la nouvelle approche au-delà de la comparaison avec les résultats fournis par la méthode de génomique des populations. Les indications qu'elle fournit sur des régions du génome et leur association avec des paramètres environnementaux peuvent dès lors être utilisés comme point de départ pour des investigations plus poussées. On suggère par exemple d'utiliser des marqueurs qui présentent une signature de sélection pour évaluer la valeur adaptative de populations en biologie de la conservation (Bonin et al., 2007). Dans d'autres domaines, il est également

possible d'imaginer de déterminer ou de confirmer l'origine géographique d'échantillons biologiques ou d'individus (Luikart, 2003).

6.2. Particularités de la méthode SAM

La SAM est une méthode exploratoire. Elle permet d'effectuer un criblage de toutes les associations potentielles entre un jeu de données moléculaires et un jeu de paramètres environnementaux. Par rapport à une démarche hypothético-déductive classique, elle permet d'éviter de devoir faire des choix sur un certain nombre de variables qui correspondent à l'hypothèse. Elle affranchit également le chercheur du risque de passer à côté de telle ou telle association importante, dans la mesure où il est possible d'être exhaustif sans que le traitement des données ne représente une tâche démesurée : les tables de réjection permettent de mettre en évidence rapidement les associations les plus significatives.

Contrairement aux approches théoriques en génomique des populations, pour fonctionner la SAM recourt à deux familles de données : les données génétiques, mais également les données environnementales. Même si l'utilisation de la SAM offre des avantages, la recherche ainsi que la connaissance des caractéristiques des jeux de données environnementaux constitue une contrainte pour les biologistes et les généticiens, à moins qu'ils ne fassent partie de réseaux interdisciplinaires au sein desquels il leur sera possible de trouver les compétences nécessaires. Ces dernières sont particulièrement précieuses lorsqu'il s'agit de choisir une résolution de données appropriée en fonction de l'étendue d'une zone d'étude. En outre, il faut signaler qu'il est toujours possible de collecter certaines données environnementales en faisant des mesures sur le terrain lors des campagnes d'échantillonnage pendant lesquelles les coordonnées géographiques sont également acquises.

En ce qui concerne l'échantillonnage, l'emploi de la SAM implique des règles différentes de celles que l'on a l'habitude de suivre en génétique des populations. Le but est d'obtenir un nombre statistiquement représentatif d'individus par type d'environnement, indépendamment de leur appartenance à une population. Il est, en effet, plus informatif d'échantillonner parmi une grande diversité d'écosystèmes de manière à pouvoir mettre en évidence un nombre potentiel plus grand d'associations entre des régions du génome et des paramètres environnementaux. Les différentes espèces, races ou populations sont prises en compte plus tard dans l'analyse.

6.3. Des apports importants pour l'écologie moléculaire

Comme nous venons de le voir, la méthode SAM se libère de toute notion de population et procède au niveau individuel. Elle présente donc un triple avantage. Premièrement, elle est plus facilement applicable chez des organismes où il n'est pas aisé de définir les limites géographiques d'une population. Deuxièmement, elle ne fait pas d'hypothèse préalable sur la structure génotypique des populations. Ceci est

un atout immense dans le cadre des criblages génomiques basés sur des marqueurs dominants comme les AFLP, pour lesquels les méthodes classiques dépendent beaucoup de modèles théoriques en génétique des populations (Bonin, 2006a). Finalement, cette nouvelle méthode offre l'avantage de, non seulement détecter des locus potentiellement adaptatifs, mais également de formuler des hypothèses plus précises concernant les facteurs environnementaux potentiellement responsables de la pression de sélection. A plus long terme, la méthode SAM ouvre donc une voie vers la caractérisation fonctionnelle des gènes impliqués dans l'adaptation locale des organismes à leur environnement.

La méthode proposée est parfaitement adaptée pour exploiter les données produites en génomique des populations et disponibles pour beaucoup d'espèces (Luikart et al., 2003). La mise en relation des données environnementales avec des données génétiques issues de larges balayages du génome¹⁰ permet de compléter les apports de la génétique environnementale (Manel et al., 2003) et de constituer ce que l'on peut appeler la génomique environnementale (Joost et al., 2007). Elle met à disposition de la communauté scientifique un outil capable de fournir des points de départ pour le repérage de gènes et la compréhension de leur(s) fonction(s).

6.4. Un domaine d'application pour la Science de l'information géographique

Dans son plaidoyer « A new biology for a new century », Carl Woese (2004) affirme qu'il est souhaitable que différents types d'approches soient exploités de manière à aborder sous différents angles cette information si complexe que sont les données moléculaires, et à procurer des moyens complémentaires pour la comprendre. De même, Michel Morange explique dans « Les secrets du vivant, contre la pensée unique en biologie » (2005) que de nouvelles approches ainsi que des innovations importantes sont requises en biologie moléculaire de manière à ce qu'« une nouvelle lumière » émerge. Pour lui, ces innovations ne pourront voir le jour que dans le contexte d'efforts interdisciplinaires.

La science de l'information géographique est justement une discipline qui aborde l'information génétique d'un point de vue original susceptible de fournir des méthodes inédites dans le but d'aborder quelques-uns des défis liés à la compréhension des ressources génétiques et des processus évolutifs. De plus, il est certain que cette discipline pourra nourrir ses propres champs de recherche en s'intéressant à l'information génétique puisque cette dernière peut être utilement exploitée en analyse spatiale exploratoire (ESDA, GVIS), et en cartographie thématique (Joost, 2006). De plus, la très grande variabilité dans le temps de marqueurs moléculaires comme les microsatellites, couplée à un contexte spatial omniprésent en biologie de la conservation, représente un contexte d'étude idéal

¹⁰ Dans le cas du mouton avec des marqueurs de type microsatellite, c'est le niveau d'association entre 118 paramètres environnementaux et 744 allèles qui a été calculé une opération, soit plus de 80'000 modèles.

pour la modélisation de données spatio-temporelles. Enfin, les résultats présentés dans cet article illustrent parfaitement les possibilités offertes par l'analyse spatiale.

Nos plus vifs remerciements vont à Laurence Després, UJF Grenoble, et Cyrille Conord, WSL Zurich, pour la mise à disposition des données sur le grand charançon du pin, ainsi qu'à Paolo Ajmone Marsan, UCSC Piacenza, et au Consortium Econogene (www.econogene.eu) pour les données sur le mouton.

7. Bibliographie

- Ajmone-Marsan P., Valentini, A., Cassandro, M., Vecchiotti-Antaldi, G., Bertoni, G., Kuiper, M., « AFLP (TM) markers for DNA fingerprinting in cattle », *Animal Genetics*, vol. 28, n° 6, 1997, p. 418-426.
- Ajmone-Marsan P., « Overview of Econogene, a European project that integrates genetics, socio-economics and geo-statistics for the sustainable conservation of sheep and goat genetic resources », *International Workshop on the role of biotechnology for the characterisation and conservation of crop, forestry, animal and fishery genetic resources*, Turin, 2-5 mars 2005, Rome, FAO, p. 89-96.
- Avise J.C., *Molecular Markers, Natural History, and Evolution*, Sunderland, Sinauer, 2004.
- Beaumont M.A., « Adaptation and speciation: what can FST tell us? », *Trends in Ecology & Evolution*, n° 20, 2005, p. 435-440.
- Beaumont M.A., Nichols R.A., « Evaluating loci for use in the genetic analysis of population structure », *Proceedings of the Royal Society of London Series B-Biological Sciences*, vol. 263, n° 1377, 1996, p. 1619-1626.
- Beh K.J., Hulme D.J., Callaghan M.J. et al., « A genome scan for quantitative trait loci affecting resistance to *Trichostrongylus colubriformis* in sheep », *Animal Genetics*, n° 33, 2002, p. 97-106.
- Berry R.J., « Ecology : where genes and geography meet », *Journal of Animal Ecology*, n° 58, 1989, p. 733-759.
- Beven K.J., « Towards a coherent philosophy for environmental modelling », *Proceedings of the Royal Society A*, n° 458, 2002, p. 2465-2484.
- Black W.C., Baer C.F., Antolin M.F., DuTeau N.M., « Population genomics: genome-wide sampling of insect populations », *Annual Review of Entomology*, n° 46, 2001, p. 441-469.

- Bonin A., Pompanon F., Taberlet P. « Use of Amplified Fragment Length Polymorphism (AFLP) markers in surveys of vertebrate diversity », in *Molecular Evolution: Producing the Biochemical Data, Part B* (Ed. Zimmer E.A., Roalson E.), p. 145-161, Academic Press, New York City, 2005.
- Bonin A., Génomique des populations et adaptation locale : exemple de la grenouille rousse (*Rana temporaria*) le long d'un gradient d'altitude, Thèse de doctorat, Université Joseph Fourier – Grenoble I, 2006a.
- Bonin A., Miaud C., Taberlet P., Pompanon F., « Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*) », *Molecular Biology and Evolution*, n° 23, 2006b, p. 773-783.
- Bonin A., Nicole F., Pompanon F., Miaud C., Taberlet P., « Population Adaptive Index: a new method to help measure intraspecific genetic diversity and prioritize populations for conservation », *Conservation Biology*, vol. 21, n° 3, 2007, p. 697-708.
- Buitkamp J., Filmether P., Stear M.J., Epplen J.T., « Class I and class II major histocompatibility complex alleles are associated with faecal egg counts following natural, predominantly *Ostertagia circumcincta* infection », *Parasitology Research*, n° 82, 1996, p. 693-696.
- Caloz R., « Réflexions sur les incertitudes et leurs propagations en analyse spatiale », *Revue Internationale de Géomatique*, vol.15, n° 3, 2005, p. 303-319.
- Cavalli-Sforza L., Menozzi P., Piazza A., *The history and geography of human genes*, Princeton, Princeton University Press, 1994.
- Conord C., Ecologie, génétique et symbiose bactérienne chez le grand charançon du pin, *Hylobius abietis* : adaptation d'un insecte ravageur à son environnement forestier, Thèse de doctorat, Université Joseph Fourier – Grenoble I, 2006a.
- Conord C., Lempérière G., Taberlet P., Després L., « Genetic structure of the forest pest *Hylobius abietis* on conifer plantations at different spatial scales in Europe », *Heredity*, n° 97, 2006b, p.46–55.
- Couclelis H., « The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge », *Transactions in GIS*, vol. 7, n° 2, 2003, p. 165-175.
- Darwin C., *The voyage of the Beagle, Charles Darwin's journal of researches*, London, Penguin Books, 1989 (1839).
- de Duve C., *A l'écoute du vivant*, Odile Jacob, Paris, 2005.
- Epperson K.E., *Geographical Genetics*, Princeton, Princeton University Press, 2003.

- Goodchild M.F., « Geographic Information Systems and spatial analysis in the social sciences », in *Anthropology, space, and Geographic Information Systems*, (Ed) Aldenderfer M., Maschner H.D.G., Oxford University Press, New York, p. 214-250, 1996.
- Hosmer D.W., Lemeshow S., *Applied logistic regression*, New York, John Wiley & Sons, 2000.
- Jelinski D., « On genes and geography: a landscape perspective on genetic variation in natural plant populations », *Landscape and Urban Planning*, vol. 39, n° 1, 1997, p. 11-23.
- Joost S., The geographical dimension of genetic diversity: a GIScience contribution for the conservation of animal genetic resources, Thèse de doctorat n° 3454, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, 2006.
- Joost S., Bonin A., Bruford M.W., Després L., Conord C., Erhardt G., Taberlet P., « A Spatial Analysis Method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation », *Molecular Ecology*, 18:3955-3969.
- Lewontin R.C., Krakauer J., « Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms », *Genetics*, n° 74, 1973, p. 175-195.
- Luikart G., England P.R., Tallmon D., Jordan S., Taberlet P., « The power and promise of population genomics: from genotyping to genome typing », *Nature Reviews Genetics*, n° 4, 2003, p. 981-994.
- MacArthur R.H., Wilson E.O., *The theory of island biogeography*, Princeton, Princeton University Press, 2001.
- MacCullagh P., Nelder J.A., *Generalized Linear Models*, London, Chapman & Hall/CRC, 1989.
- Mackay T.F., « The genetic architecture of quantitative traits », *Annual Review of Genetics*, vol. 35, 2001, p. 303-339.
- Manel S., Schwartz M., Luikart G., Taberlet P., « Landscape genetics: combining landscape ecology and population genetics », *Trends in Ecology & Evolution*, vol. 18, n° 4, 2003, p. 189-197.
- Mourant A., *The distribution of the human blood groups*, Oxford, Blackwell Scientific, 1954.
- Morange M., *Les secrets du vivant, contre la pensée unique en biologie*, Paris, La Découverte, 2005.
- Narum S.R., « Beyond Bonferroni: Less conservative analyses for conservation genetics », *Conservation Genetics*, n° 7, 2006, p. 783-787.

- Neigel J.E., « Is Fst obsolete? », *Conservation Genetics*, n° 3, 2002, p. 167-173.
- New M., Lister D., Hulme M., Makin I., « A high-resolution dataset of surface climate over global land areas », *Climate Research*, n° 21, 2002, p. 1-25.
- Nielsen R., « Molecular signatures of natural selection », *Annual Review of Genetics*, n° 39, 2005, p. 197-218.
- Peter C., Bruford M., Perez T., Dalamitra S., Hewitt G., Erhardt G., ECONOGENE Consortium, « Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds », *Animal Genetics*, n° 38, p. 37-44, 2007.
- Phillips P.C., « Testing hypotheses regarding the genetics of adaptation », *Genetica*, vol. 123, 2005, p. 15-24.
- Poincaré H., *Leçons de Mécanique céleste I*, Paris, Jacques Gabay, 2005 (1905).
- Rose S., *Lifelines : life beyond the genes*, New York, Oxford University Press, 2003.
- Shaffer J.P., « Multiple Hypothesis Testing », *Annual Review of Psychology*, n° 46, 1995, p. 561-584.
- Sinclair B.J., Vernon P., Klok C.J., Chown S.L., « Insects at low temperatures: An ecological perspective », *Trends in Ecology and Evolution*, n° 18, p. 257-262, 2003.
- Storz J.F., «Using genome scans of DNA polymorphism to infer adaptive population divergence», *Molecular Ecology*, n° 14, 2005, p. p. 671-688.
- Woese C.R., « A new biology for a new century », *Microbiology and Molecular Biology Reviews*, vol. 68, n° 2, 2004, p. 173-186.
- Wright S., « The genetical structure of populations », *Annals of Eugenics*, n° 15, 1951, p. 323-354.