Utilisation des marqueurs pour la caractérisation des ressources génétiques

Louis Ollivier, Claude Chevalet, Jean Louis J. L. Foulley

Utilisation des marqueurs pour la caractérisation des ressources génétiques

L. OLLIVIER[1], C. CHEVALET[2], J.-L. FOULLEY[1]

[1] INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy-en-Josas cedex
[2] INRA, Laboratoire de Génétique Cellulaire, BP 27, 31326 Castanet-Tolosan cedex

e-mail : ugenlol@dga2.jouy.inra.fr

Résumé. *Un aperçu des méthodes de caractérisation des ressources génétiques basées sur les marqueurs moléculaires est présenté. La variabilité génétique entre races s'exprime généralement, à partir d'un ensemble de fréquences alléliques, sous la forme d'indices de fixation et de diversité génique, ainsi que de distances génétiques entre races. Ces dernières peuvent être converties en phylogénie, classification ou mesure globale de diversité. La richesse allélique est une information complémentaire intéressante à prendre en compte. La planification expérimentale des études de diversité et leur fiabilité statistique sont brièvement discutées.*

Comme le souligne la Charte Nationale pour la gestion des ressources génétiques (BRG 1999), les ressources génétiques disponibles sont d'autant mieux utilisées qu'elles sont aussi mieux caractérisées. Cette caractérisation peut revêtir plusieurs aspects. Il y a d'abord les inventaires de races, qui rassemblent des informations aujourd'hui mises en base de données – à l'échelle nationale, européenne ou mondiale – et visant à couvrir l'ensemble des populations constituant les ressources génétiques de chaque espèce. Rappelons que ces ressources peuvent être regroupées selon une classification évolutive proposée par Lauvergne (1982), qui distingue quatre catégories de populations, soient des populations sauvages, des populations primaires (ou traditionnelles), des races standardisées et des lignées sélectionnées. Ces catégories correspondent à des statuts domesticatoires différents et reflètent des degrés successifs dans le processus général de domestication. Mais ces inventaires, à part quelques informations succinctes sur les caractères zootechniques des différentes races, ne nous disent pas grand chose sur leurs ressemblances ou dissemblances, et donc sur leur diversité. Les marqueurs génétiques offrent le grand avantage de permettre une évaluation directe de la diversité génétique, et qui soit, par définition, indépendante des effets du milieu. La question est évidemment de savoir si la variabilité génétique des marqueurs reflète bien une variabilité des locus de caractères quantitatifs, par exemple celle des caractères zootechniques ou d'adaptation. Cet aspect, qui déborde le cadre de cet article, a été discuté dans un article récent de Barker (1999), qui rapporte plusieurs études récentes sur des populations naturelles ainsi que chez des plantes montrant que les deux variabilités sont souvent connectées. Le degré des connexions et leur généralité restent encore cependant discutés.

Les marqueurs moléculaires réunissent un ensemble de propriétés qui en font un outil idéal d'évaluation de la variabilité génétique intra-race et entre races. Citons leur polymorphisme, leur ubiquité sur le génome et la possibilité d'automatiser leur identification. D'autres marqueurs génétiques ont également été utilisés, depuis près de 40 ans, dans ce but de caractérisation, mais ils ne cumulent généralement pas tous les avantages mentionnés ci-dessus. On peut sur cet aspect consulter les deux études des races bovines françaises de Grosclaude *et al* (1990) et de Moazami-Goudarzi *et al* (1997) respectivement basées sur des marqueurs biochimiques et moléculaires. Nous nous proposons de donner ici un aperçu succinct des méthodes de caractérisation basées sur les marqueurs génétiques moléculaires, pour lesquels des données de plus en plus nombreuses deviennent maintenant disponibles dans toutes les espèces.

1 / Indices de fixation de Wright et diversité génique de Nei

La méthode la plus classique de caractérisation de populations, et peut-être aussi la plus ancienne, est celle des indices de fixation (F) proposée par Wright en 1943 (cité par Nei 1977), qui définit trois coefficients F entre lesquels existe la relation suivante :

$$(1 - F_{IT}) = (1 - F_{IS})\,(1 - F_{ST}) \qquad (1)$$

Ces indices de fixation généralisent le coefficient de consanguinité en l'étendant au cas des populations subdivisées. Ainsi F_{IT} et F_{IS} sont-ils respectivement les consanguinités (ou corrélations entre gamètes dans l'approche de Wright) des individus dans la population totale (IT) et des individus dans chaque sous-population (IS). Le terme F_{ST}, lui, n'est pas à proprement parler une consanguinité, puisqu'il désigne la corrélation entre deux gamètes pris au

hasard chez deux individus différents dans une même sous-population de la population totale (ST). F_{ST} mesure le degré de différenciation génétique des sous-populations. Ces F peuvent aussi se formuler selon l'approche probabiliste de Malécot (bien que F_{IS} et F_{IT} puissent prendre des valeurs négatives). Notons que ce sont des paramètres statistiques (corrélation, probabilité) qui ne requièrent que des pedigrees et non pas la connaissance de génotypes réels.

Une approche différente a été proposée par Nei en 1973 (cité par Nei 1977), qui consiste à partir des hétérozygoties (H) réelles à différents locus. Nei définit ainsi la notion de diversité génique et montre que la diversité totale (H_T) se décompose en une diversité intra-sous-population (H_S) et une diversité entre sous-populations (D_{ST}). On a $H_T = H_S + D_{ST}$. Notons que les hétérozygoties H_T et H_S sont calculées sous l'hypothèse de l'équilibre de Hardy-Weinberg et ne dépendent donc que des fréquences alléliques observées. Nei définit alors des indices F équivalents à ceux de Wright à partir de moyennes d'hétérozygoties calculées sur plusieurs locus et d'une hétérozygotie moyenne totale observée (H_O), comme suit :

$$F_{IT} = 1 - H_O/ H_T \quad , F_{IS} = 1 - H_O/ H_S \quad \text{et} \quad F_{ST} = 1 - H_S/ H_T \qquad (2)$$

On voit que les F ainsi définis satisfont également la relation (1). Nei définit aussi un coefficient de différenciation génique entre les populations (qu'il appelle G_{ST}) tel que $G_{ST} = D_{ST} /H_T$, qui est donc équivalent – mais pas identique – au F_{ST} de Wright. Ce dernier se place en effet conceptuellement dans la situation d'un seul locus et d'un grand nombre de populations, alors que Nei prend la position inverse qui consiste à considérer plusieurs locus – dans l'idée d'apprécier la diversité du génome – sur un nombre limité de populations réelles.

2 / Distance génétique, phylogénie et diversité

La notion de différenciation génique, exprimée par le coefficient G_{ST} de Nei qui vient d'être défini, s'applique aussi au cas extrême de deux sous-populations et ce coefficient peut alors être considéré comme une mesure de la distance génétique qui les sépare. Une des premières mesures de distance proposée fut en fait la distance F_{ST} de Latter, que l'on retrouve également dans la distance de Reynolds (voir en annexe la définition de ces distances). Compte tenu de la définition donnée précédemment de G_{ST}, fonction seulement des hétérozygoties H_T et H_S à l'équilibre de Hardy-Weinberg, cette distance est une fonction des seules fréquences géniques dans les deux sous-populations. Diverses autres mesures de distance ont par ailleurs été proposées (voir Nei 1987). Les distances les plus usitées, définies en fonction des couples de fréquences géniques notées x_{ij} et y_{ij} figurent en annexe (les références correspondantes sont données par Laval 1997). Remarquons au passage qu'on peut également calculer des distances entre individus sur la base de leurs génotypes, en utilisant une estimation multi-locus des parentés, comme l'a proposé Chevalet (1980).

Si l'on dispose d'un ensemble de N sous-populations, on peut calculer une matrice (N x N) de distances deux à deux. A partir de cette matrice de distances, on peut établir une phylogénie, en partant de l'idée que la distance génétique qui sépare deux sous-populations est proportionnelle au temps qu'elles ont mis à diverger à partir d'une sous-population ancêtre. Dans le cas de distances basées sur des variations de séquence d'ADN – qui s'appliquent surtout à des diversités interspécifiques – l'hypothèse de base est donc la constance de l'évolution des séquences, au rythme de l'« horloge moléculaire ». Dans le cas des fréquences alléliques, qui nous intéresse ici, la divergence est supposée résulter d'un processus de dérive génétique, et l'hypothèse correspondante est donc le rythme uniforme de cette dérive, ce qui implique des sous-populations de même effectif génétique. Les algorithmes permettant de passer de la matrice des distances à un arbre phylogénétique, et la manière de calculer les longueurs des branches sont décrits en détail par Hartl et Clark (1997, p. 368-372).

Le nombre des arbres, ou topologies, possibles a priori augmente très rapidement avec N et, pour N=10 sous-populations, le nombre des arbres possibles dépasse déjà 34×10^6. La topologie obtenue à partir d'un ensemble de distances est seulement la plus vraisemblable dans une vaste quantité de possibles. La question du degré de confiance à lui accorder se pose donc. La technique généralement adoptée est celle de la « languette de botte(1) » (bootstrap en anglais), qui consiste à rééchantillonner les données – par tirage au hasard non exhaustif des 2n allèles présents à chaque locus dans un échantillon de n individus, ou par tirage des L locus – et, sur un grand nombre de rééchantillonnages (1000 par exemple), à calculer la fréquence d'apparition de chacun des embranchements obtenus sur l'arbre initial (Felsenstein 1985). On pourra alors se fixer un seuil de signification, à 5 % par exemple si l'embranchement apparaît dans 95 % des cas. Notons au passage que la validité statistique de cette technique repose sur une hypothèse forte d'homogénéité des unités expérimentales, qui n'est pas forcément vérifiée en particulier pour des locus différents. Un exemple d'arbre phylogénétique de races bovines françaises (Moazami-Goudarzi *et al* 1997) est donné dans la figure 1. On voit que cet arbre est peu fiable, puisqu'aucun embranchement n'atteint le seuil de 5 %.

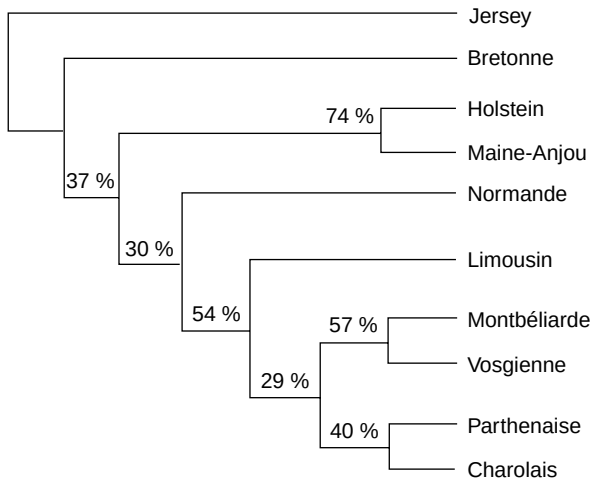
La matrice des distances génétiques peut également servir à évaluer la diversité globale d'un ensemble S de races, selon l'approche proposée par Weitzman (1993). Celui-ci définit une fonction de diversité V(S), qui est le maximum sur toutes les races de l'ensemble S, de la distance d_i entre la race i et la race la plus proche dans S, plus la diversité du sous-ensemble S_i obtenu en éliminant la race considérée, soit :

$$V(S) = \text{maximum sur S de } [d_i + V(S_i)] \qquad (3)$$

Il s'agit, on le voit, d'une définition récursive, puisque $V(S_i)$ doit être aussi calculé selon la même procédure, soit $V(S_i)$= maximum sur S_i de $[d_{ij} + V(S_{ij})]$, d_{ij} étant ici défini comme la distance entre la race j et le sous-ensemble S_{ij} obtenu en éliminant i et j, et ainsi de suite.

(1) Selon Robert (1992), ce terme fait allusion à une histoire contée par Cyrano de Bergerac (Histoire Comique Contenant les Estats et Empires de la Lune, 1657), dans laquelle le héros atteint la lune en tirant sur ses languettes de botte. Mais l'auteur de la méthode semblait lui-même ignorer Cyrano et il a choisi une appellation qui faisait référence à une aventure similaire contée par un auteur allemand du 18è siècle (D. Laloë, communication personnelle).

Figure 1. *Phylogénie de 10 races bovines françaises (d'après Moazami-Goudarzi* et al *1997).*
L'arbre est construit à partir des distances calculées sur la base de 17 marqueurs moléculaires (microsatellites). Les chiffres indiqués sont les pourcentages d'apparition de chaque embranchement sur 500 rééchantillonnages.

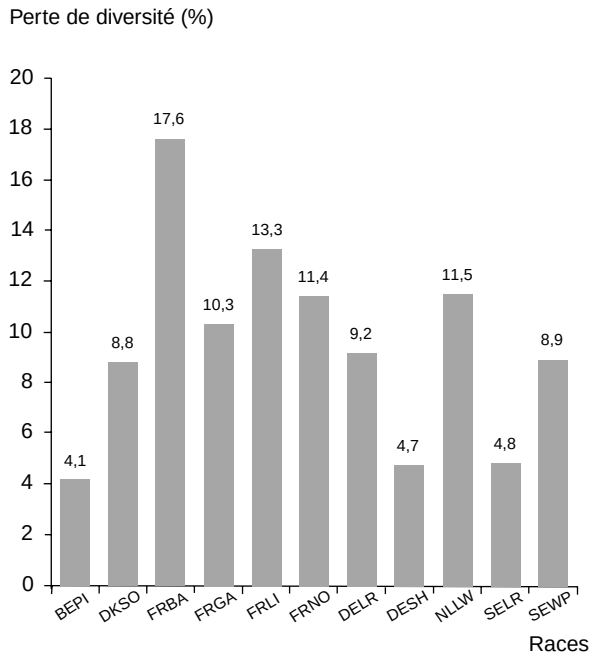


Les propriétés intéressantes de cette fonction et ses applications dans le contexte de la diversité des races animales domestiques ont été décrites par Thaon d'Arnoldi *et al* (1998). Notons aussi que la solution de (3) génère un arbre phylogénétique qui a la propriété de maximiser la vraisemblance de la diversité observée, et dont la fiabilité peut être évaluée statistiquement, sans faire appel à des rééchantillonnages. De plus, la longueur totale des branches de l'arbre est égale à la fonction V définie en (3), et la longueur de chaque branche raciale mesure approximativement la contribution de la race correspondante à la diversité totale V. Un exemple d'application à des races porcines européennes est donné dans la figure 2.

Contrairement aux arbres obtenus par les algorithmes classiques (voir Hartl et Clark 1997), la longueur des branches de l'arbre de Weitzman ne repose sur aucune hypothèse évolutive. Comme le souligne Weitzman (cité par Thaon d'Arnoldi *et al* 1998), une diversité peut être évaluée et un arbre phylogénétique construit, même si les éléments de l'ensemble S ne résultent d'aucun processus évolutif.

Quelle que soit la méthode utilisée pour construire des arbres, ceux-ci permettent une classification des races, avec des regroupements en sous-ensembles plus ou moins distincts. Cette classification est évidemment dépendante des hypothèses qui ont présidé à la construction des arbres correspondants, et sa fiabilité est limitée par la fiabilité de ces derniers. D'autres méthodes de classification sont applicables, comme, par exemple, l'analyse des correspondances appliquée à des tableaux de contingence race x allèles. Cette méthode permet de représenter un ensemble de races dans un espace euclidien, et de mesurer la diversité par l'inertie du système qui représente cet ensemble (voir Laloë *et al* 1999). Notons que ce procédé de classification est affranchi de toute hypothèse concernant l'évolution qui a conduit à la diversité observée.

Figure 2. *Pertes marginales de diversité associées à onze races porcines européennes (d'après Laval* et al *2000).*
Les pourcentages indiquent de combien est réduite, en valeur relative, la diversité de l'ensemble des onze races lorsque la race correspondante est soustraite de l'ensemble. Races : Piétrain = BEPI, Sortbroget (porc local danois) = DKSO, Basque = FRBA, Gascon = FRGA, Limousin = FRLI, Normand = FRNO, Landrace allemand = DELR, Schwäbisch-Hällisches (porc souabe) = DESH, Grand Yorkshire (Large White) = NLLW, Landrace Suédois = SELR, Sanglier européen = SEWP.



3 / Variabilité génétique intra-race et richesse allélique

Si les mesures de différenciation des sous-populations que nous venons de voir sont très importantes pour caractériser la diversité, il faut envisager également l'utilité de mesures de variabilité intra-sous-population. De ce point de vue, si le critère classique est l'hétérozygotie, de nombreuses raisons (voir Barker 1999) militent en faveur d'un critère complémentaire qui est le nombre des allèles recensés par locus dans chaque sous-population. Dans une étude française récente (Petit, El Mousadik et Pons 1998, cités par Barker 1999), le concept de richesse allélique a été étendu au cas des populations fractionnées, selon une approche parallèle à celle de Nei pour les hétérozygoties. Petit et al. calculent la contribution de chaque sous-population à l'hétérozygotie comme la différence entre l'hétérozygotie calculée sur la population dans son ensemble et celle calculée en excluant la sous-population en question. La contribution de chaque sous-population à la richesse allélique totale (C_T) se calcule selon le même principe et se décompose aussi en une contribution liée à la richesse intra-sous-population (C_S) et une contribution liée à sa divergence relativement aux autres sous-populations (C_D). L'exemple des arganiers marocains de Petit *et al* que reprend Barker (1999) montre que les contributions C_T et C_D sont très corrélées entre elles (0,86), alors qu'elles ne sont pas en corrélation significative avec les contributions à la différenciation génique G_{ST} , comme l'indique le tableau 1. On voit aussi que, parmi les contributions à la richesse

allélique, seule la contribution C_D est significativement liée à la perte relative de la fonction de diversité de Weitzman (D_W), et que celle-ci est par ailleurs significativement corrélée à G_{ST}.

Tableau 1. *Corrélations entre plusieurs mesures de diversité (d'après Barker 1999).*
Corrélations obtenues sur 12 sous-populations d'arganiers marocains, les corrélations significatives sont indiquées en gras. Les quantités G_{ST},C_T, C_S, C_D et D_W sont définies comme les contributions de chaque sous-population d'arganiers aux mesures de diversité suivantes : G_{ST} = coefficient de différenciation génique de Nei (défini dans le texte) ; C_T, C_S et C_D = richesse allélique totale (T), intra-(S) et entre sous-populations (D) ; D_W = fonction de diversité de Weitzman.

	G_{ST}	C_T	C_S	C_D
C_T	-0,05			
C_S	**-0,77**	0,47		
C_D	0,39	**0,86**	-0,05	
D_W	**0,66**	0,44	-0,21	**0,62**

Cet exemple indique que les diversités inter et intra ne sont pas forcément liées et qu'il y aurait donc lieu de les prendre en compte l'une et l'autre. On voit aussi que la richesse allélique entre sous-populations peut aider à mesurer leur différenciation.

Notons que ce critère de richesse allélique, encore peu utilisé pour les races animales, est fortement lié à la taille des échantillons. On peut en effet facilement prévoir que, toutes choses égales par ailleurs, le nombre d'allèles trouvés dans une race augmentera avec la taille de l'échantillon étudié. La procédure proposée par Petit *et al* pour éliminer ce biais consiste à ramener la richesse allélique observée à sa valeur attendue pour la plus petite taille d'échantillon rencontrée sur l'ensemble des sous-populations à comparer (voir Barker 1999). Il n'est pas sûr que cette sorte de nivellement par le bas utilise au mieux la totalité de l'information disponible.

Discussion et conclusions

Au terme de cet aperçu général des méthodes de caractérisation, il apparaît que ce champ d'investigation connaît depuis quelques années un regain d'intérêt, mais que les méthodes disponibles sont si nombreuses et variées qu'il est souvent difficile d'y voir clair. A la multiplicité des mesures de distance s'ajoutent celle des méthodes de construction d'arbres phylogénétiques, ainsi que les incertitudes qui entourent l'interprétation de ces derniers. Ce n'est qu'assez récemment qu'il a été reconnu que les processus phylogénétiques à l'origine des races animales domestiques n'ont pas grand chose à voir avec ceux qui prévalent dans l'évolution des espèces ou sous-espèces sauvages, comme le notent Grosclaude *et al* (1990) ainsi que Thaon d'Arnoldi *et al* (1998). Il est donc hasardeux de vouloir transposer aux races animales des méthodes qui ne s'appliquent qu'aux populations naturelles, et on peut se demander si l'arbre (phylogénétique) n'a pas jusqu'à présent trop souvent caché la forêt (de la diversité animale).

Pendant longtemps, l'optimisation du plan d'expérience à mettre en œuvre pour caractériser un ensemble de races n'a guère retenu l'attention. Les recommandations d'un groupe de travail réuni par la FAO en 1993 (cité par Barker 1999) ont donc été particulièrement bienvenues. Ce groupe de travail a choisi de baser ses recommandations sur une mesure de distance (D) ayant une variance d'échantillonnage V(D) particulièrement simple. En fonction du nombre n d'individus échantillonnés par race, du nombre L de locus, du nombre k de distances indépendantes calculées par locus, correspondant à k+1 allèles, et de la distance vraie d, cette variance s'écrit :

$$V(D)= 2\,[d+(1/n)]^2 / Lk \qquad (4)$$

La démonstration formelle de cette formule a été présentée par Foulley et Hill (1999), sur la base de la distance de Sanghvi, elle-même proche de celle de Reynolds *et al* (voir respectivement les références Balakrishnan et Sanghvi 1968 et Reynolds *et al* 1983 de l'annexe). Cette formule intègre bien les deux sources d'aléa que sont l'échantillonnage des individus (n) dans la population et l'échantillonnage des locus marqueurs (L) dans le génome. L'écriture des variances des autres mesures de distances est généralement plus complexe (voir, par exemple, Laval 1997).

La formule (4) a permis au groupe de travail FAO de recommander n=25 et Lk=50 pour évaluer avec une précision suffisante des distances entre des races étroitement apparentées, c'est-à-dire avec des valeurs de la distance de Sanghvi proches de d = 0,05. La formule (4) montre que le nombre L de locus (ou d'allèles indépendants Lk) est plus important que le nombre n d'individus par race pour obtenir une distance précise. Remarquons aussi que la fonction de diversité de Weitzman étant une somme de distances, sa précision peut se calculer à partir de (4). On peut ainsi raisonner à la fois le nombre des races et la taille des échantillons par race.

Mentionnons pour terminer que la fiabilité statistique des distances, si elle paraît nécessaire pour obtenir une phylogénie significative, ne peut pas la garantir dans tous les cas. L'exemple d'une étude récente de diversité dans l'espèce porcine (Laval *et al* 2000) tend à le montrer. Les distances calculées entre les onze races considérées sont en effet toutes significatives sur la base de la formule (4), alors qu'aucune phylogénie fiable n'est obtenue pour neuf de ces races. A l'inverse, si des phylogénies douteuses sont obtenues, comme dans le cas de la figure 1, il paraît hasardeux de conclure à une insuffisance du nombre des locus marqueurs, sans avoir au préalable établi les variances d'échantillonnage des distances estimées et vérifié leur degré de signification.

Remerciements

Les auteurs remercient F. Grosclaude (INRA, Génétique animale, Jouy-en-Josas) pour les commentaires qu'il leur a fournis lors de la préparation de cet article.

Références

Barker J.S.F., 1999. Conservation – are trees and animals different? In : Symposium in honour of Dr. G. Namkoong, Unifying Perspectives of Evolution. Conservation and Breeding, Vancouver, 22-24 juillet 1999.

BRG, 1999. Charte Nationale pour la gestion des ressources génétiques. BRG, Paris.

Chevalet C., 1980. Calcul des coefficients d'identité, inégalités et distances génétiques. In : J.M. Legay, J.P. Masson, R. Tomassone (eds), Biométrie et Génétique, 42-49. Société Française de Biométrie, INRA, Paris.

Felsenstein J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39, 783-791.

Foulley J.L., Hill W.G., 1999. On the precision of estimation of genetic distance. Genet. Sel. Evol., 31, 457-464.

Grosclaude F., Aupetit R.Y., Lefebvre J., Mériaux J.C., 1990. Essai d'analyse des relations génétiques entre les races bovines françaises à l'aide du polymorphisme biochimique. Génét. Sél. Evol., 22, 317-338.

Hartl D.L., Clark A.G., 1997. Principles of Population Genetics (3rd ed). Sinauer Associates, Sunderland, Massachussetts, USA.

Laloë D., Moazami-Goudarzi K., Souvenir Zafindrajaona P., 1999. Analyse des correspondances et biodiversité dans les races domestiques. Société Française de Biométrie, 20 mai 1999, Grenoble, 5p.

Lauvergne J.J., 1982. Genética en poblaciones animales despuès de la domesticacion : consecuencias para la conservacion de las razas. Proc. World Cong. Genet. Appl. Livestock Prod., 6, 77-87.

Laval G., 1997. Modélisation et mesure de la différenciation génétique des races animales à l'aide de marqueurs microsatellites. DEA de Biologie des Populations, Génétique et Ecoéthologie, Université de Tours, 1997, 25p.

Laval G., Iannuccelli N., Legault C., Milan D., Groenen M.A.M., Andersson L., Fredholm M., Geldermann H., Foulley J.L., Chevalet C., Ollivier L., 2000. Genetic diversity of eleven European pig breeds. Genet. Sel. Evol, 32, 187-203.

Moazami-Goudarzi K., Laloë D., Furet J.P., Grosclaude F., 1997. Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites. Anim. Genet., 28, 338-345.

Nei M., 1977. F- Statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet., 41, 221-233.

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

Robert C., 1992. L'analyse statistique bayésienne. Economica, Paris.

Thaon d'Arnoldi C., Foulley J.L., Ollivier L., 1998. An overview of the Weitzman approach to diversity. Genet. Sel. Evol., 30, 149-161.

Annexe. *Distances utilisées en génétique.*

Notation : i=1,2,..I locus ; j=1,2,.., J_i allèles par locus,
de fréquences x_{ij} et y_{ij} dans les deux populations considérées

Rogers (1972) (euclidienne) :

$$D_R = \frac{1}{I} \sum_{i=1}^{I} \left[\frac{1}{2} \sum_{j=1}^{J_i} (x_{ij} - y_{ij})^2 \right]$$

Prevosti *et al* (1975) ou Gregorius (Manhattan) :

$$C_P = \frac{1}{2I} \left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} |x_{ij} - y_{ij}| \right)$$

Cavalli-Sforza et Edwards (1967) (longueur d'une corde) :

$$D_C = \frac{2}{\pi I} \sum_{i=1}^{I} \left[2 \left(1 - \sum_{j=1}^{J_i} x_{ij} y_{ij}\right) \right]^{1/2}$$

Balakrishnan et Sanghvi (1968) (Khi-deux) :

$$X^2 = \frac{1}{J_+ - I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{2(x_{ij} - y_{ij})^2}{x_{ij} + y_{ij}}$$

Reynolds *et al* (1983) ou F_{ST} de Latter (1972) :

$$F_{ST} = \frac{1}{2} \frac{\sum_{i=1}^{I} \sum_{j=1}^{J_i} (x_{ij} - y_{ij})^2}{\sum_{i=1}^{I} \left(1 - \sum_{j=1}^{J_i} x_{ij} y_{ij}\right)}$$

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

$$J_{XY} = I^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J_i} x_{ij} y_{ij} \; ; \; J_X = I^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J_i} x_{ij}^2$$

Nei minimum : $D_m = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left(x_{ij} - y_{ij}\right)^2 = (J_X + J_Y) / 2 - J_{XY}$

Nei standard : $D = - \log_e \frac{J_{XY}}{(J_X J_Y)^{1/2}}$

Nei D_A : $D_A = 1 - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left(x_{ij} y_{ij}\right)^{1/2}$

A noter que F_{ST} peut se mettre aussi sous la forme :

$$F_{ST} = \frac{1}{2} \frac{(J_X + J_Y) / 2 - J_{XY}}{1 - J_{XY}}$$

Attention également au biais qui affecte les distances dans ces formulations classiques : voir par exemple Laval (1997) pour le F_{ST} , ainsi que Foulley et Hill (1999) dans le cas de la distance de Sanghvi.