THE GEOGRAPHICAL DIMENSION OF GENETIC DIVERSITY: A GISCIENCE CONTRIBUTION FOR THE CONSERVATION OF ANIMAL GENETIC RESOURCES

PAR

Stéphane JOOST

Licencié ès lettres de l'Université de Lausanne
et de nationalité suisse et originaire de Langnau im Emmental (BE)

Au moment d'apporter la touche finale à ce document, j'ai une pensée particulière pour mes parents, Drussy et Pierre, disparus alors qu'ils étaient encore beaucoup trop jeunes, au moment où je commençais cette recherche à l'EPFL.

J'ai également une pensée pour Jean-Bernard Favre, le papa de ma belle-soeur Christine, qui nous a quittés récemment et que j'appréciais beaucoup.

ACKNOWLEDGEMENTS

Many people were involved in this research and either helped me to achieve one task, played a game to favour the achievement of this work somehow or other, and encouraged me during these last three years.

L'histoire a commencé un soir de juin 2001, lors de Vivapoly, la fête de l'EPFL. Entre deux verres, Marc Riedo et Abram Pointet m'avaient signalé que la chaire de SIRS (l'actuel Laboratoire de Systèmes d'Information Géographique ou LASIG) était à la recherche d'une personne pour travailler dans le cadre d'un projet européen «à propos d'analyse spatiale et de chèvres». Alors merci Marc et Abram! De plus, vous avez là un argument de choc pour justifier votre présence régulière dans les fêtes...

La semaine suivante déjà, après une première rencontre avec Régis Caloz, puis avec le Prof. François Golay, tout était réglé : j'allais commencer à travailler en septembre dans leur laboratoire ! Un immense merci à tous les deux pour leur confiance, pour la manière dont ils ont combiné leurs compétences afin d'encadrer mon travail. Merci à eux également pour l'ambiance extraordinaire qui existe au sein du LASIG, ceci grâce à leur sensibilité et à l'importance qu'ils accordent à la qualité des relations humaines.

Together with François Golay, Prof. Paolo Ajmone-Marsan (Università Cattolica del Sacro Cuore, Piacenza, Italy) was co-advisor of my Ph.D. thesis. Vorrei ringraziare Paolo per tutto il tempo che ha dedicato ai miei lavori. Questo è stato possibile sicuramente grazie al suo detto : «Dilatiamo il tempo!» (preso ad Einstein...). Grazie anche a Paolo per i suoi consigli, per la sua gentilezza, e per la sua fiducia nel contesto del progetto europeo Econogene del quale era il coordinatore.

La mia gratitudine va anche al Prof. Alessio Valentini (Università della Tuscia, Viterbo, Italia) con chi ho parlato tante volte degli aspetti geografici dell'informazione genetica, e che mi ha invitato a dare una lezione all' «International Summer School of Animal Genomics (ISSAG)» a Tuscania. Alessio mi ha anche dato la possibilità di presentare i miei lavori al secondo congresso dell'Italian Proteome Society (IPSo), ciò che rappresenta una bella dimostrazione della sua fiducia.

Merci infiniment également au Dr Pierre Taberlet - directeur du Laboratoire d'Ecologie Alpine (LECA) à Grenoble - pour ses précieux conseils, pour tout l'intérêt qu'il a porté à mon travail, pour son soutien, ses encouragements et son enthousiasme! His interest in spatial processes to explain genetic diversity is shared by Prof. Michael Bruford (University of Wales, Cardiff) who spent hours with me in order to select the most relevant genetic variables, and who cordially welcomed me for a short stay in his laboratory in Cardiff. Thank you Mike ! I also would like to warmly thank Prof. Godfrey Hewitt (University of East Anglia, Norwich) for the time he spent to explain to me the fundamentals of molecular ecology.

Acknowledgements

Merci également au Prof. Laurent Excoffier et à Gabriela Obexer-Ruff (Université de Berne), aux Profs. Nicole Galland, Jérôme Goudet, Nicolas Perrin et Antoine Guisan (Département d'Ecologie et d'Evolution, Université de Lausanne), à Christophe Randin, ainsi qu'à tout le laboratoire d'Ecologie Spatiale pour les échanges intéressants que nous avons eus ensemble.

Pour terminer, j'aimerais également remercier tous mes amis pour leur soutien et leur présence durant l'élaboration de cette thèse. Vous avez tous dû subir une certaine absence de diversité (!) dans nos sujets de conversation...

Pour les mêmes raisons, merci à toute ma famille ! A vous Cathy, Muriel, Zabo, Daniel, Thierry, Ghita, Pascal, Christine, Julien, Coralie, Anouk, Isabelle, Yvan, Jean-François, Laure, Roland, Corinne, Michèle, Geneviève, Anne-Romaine, Raphaëlle, Dominique, Anne, Emma, et Yves.

Merci à vous aussi Renata, Martin, David, Véro, Gab, Natalia (merci pour l'idée des remerciements multilingues!), Chris, Raymonde, Henriette, Jacqueline et Pierre.

Et par-dessus tout, merci à mon frère Nicolas, à Christine, ainsi qu'à Maxime, Charline et Benjamin. Et à toi Aline, pour les encouragements et le soutien que tu m'as constamment apportés.

Stéphane Joost, février 2006

Acknowledgements

ABSTRACT

. . . .
. .
. .
. .

In its natural framework, genetic information is embedded within a geographic context. Plants and animals are directly influenced by the specific characteristics of their surrounding environment. Therefore, spatial information is a potentially important element to be considered in trying to understand genetic resources.

For many years, Geographical Information Science (GIScience) turned toward environmental modelling, generally to demonstrate how GIS basic features could be efficiently applied to fields related to the natural sciences. However, despite its current predominance in life sciences and its direct application to concerns of public society (health, food), genetics had heretofore remained outside the scope of research by the GIScience community while biologists were developing approaches based on GIS, which conducted to the elaboration of «landscape genetics».

The present GIScience approach to linking genetics and geographics obviously places emphasis on geographic information, unlike most studies in this domain. This perspective is developed through an application to two case studies to assess the potential contribution of GIScience to conservation biology. The main one is provided by Econogene, a European research project aiming at promoting the sustainable conservation of genetic resources in sheep and goats. These small ruminants have considerable economic importance in marginal agrosystems in Europe. The surveying of their genetic resources makes it possible to highlight endangered breeds having high distinctiveness and priority for conservation. The second case study is complementary and supplies wild species data to demonstrate how genetic information is used to assess conservation measures applied to the endangered Scandinavian Brown Bear.

Advanced molecular technologies make it possible to efficiently measure genetic information. In parallel, considerable advancements in computer science have led to the development of sophisticated GIS software and methods. The joining of molecular biology and GIScience enables novel and complementary methods of tackling some of the challenging issues related to evolutionary processes.

This tentative application of diverse facets of GIScience to molecular biology addresses three distinct issues. Firstly, considering the vast quantity of information collected within the Econogene project, exploratory data analysis methods are applied to extract useful information from large spatially explicit genetic data sets. This category of GIS tools facilitates investigations to understand the geographic distribution of genetic diversity among sheep and goat breeds as well as its variation according to environmental parameters.

Secondly, a review of population genetics literature having revealed weaknesses about the way spatial genetic data are generally represented on maps, the semiology

Abstract

of graphics is used to produce improved thematic maps representing patterns of genetic diversity. The recourse to high-performance cartography improves the interpretation of data, and facilitates the transmission of the results as well as their communication between researchers, or between researchers and the general public.

Thirdly, this combination of GIScience with molecular biology mainly leads to the development of a spatial analysis method to detect signatures of natural selection within the genome. The method is adapted to a precise conception of environmental modelling, advocating the implementation of simple models in order to better grasp the functioning of natural processes, and recognizing an inescapable uncertainty. To find out these signatures, spatial analysis takes advantage of the evolution of genetics toward genomics and of the subsequent availability of large data sets generated by large genome scans. This permits to compute many simultaneous univariate logistic regression models and to identify specific regions of the genome which are selected by environmental parameters. These are identical to the ones detected by a standard approach developed in population genomics.

RÉSUMÉ

· ·
· ·
· ·
· ·

Sur la planète, les techniques d'élevage intensif appliquées dans les systèmes agricoles de la plupart des pays ont une influence négative sur la biodiversité des animaux domestiques. Au cours des 15 dernières années, 300 des 6'000 races recensées par la FAO ont disparu. A l'heure actuelle, 1'350 races sont en voie d'extinction, deux races disparaissant en moyenne chaque semaine. Une des mesures mises en oeuvre afin de juguler ce phénomène est la surveillance des ressources génétiques des animaux d'élevage. Cette action a été initiée par les Nations Unies à Rio en 1992 et constitue un volet de la Convention sur la diversité biologique. La mise au point de technologies de pointe en biologie moléculaire a permis de développer des instruments capables de fournir rapidement des indications sur le niveau de diversité génétique de races sous surveillance. Ces informations permettent de repérer celles qui sont en danger d'extinction et de prendre les mesures de conservation qui s'imposent.

La science de l'information géographique peut contribuer à améliorer l'analyse de ces ressources génétiques en exploitant leur dimension géographique. Celle-ci permet d'apprécier comment la diversité génétique varie dans l'espace et peut également mettre en évidence des adéquations en fonction de la nature de l'environnement au sein duquel les races étudiées évoluent.

C'est notamment dans le contexte du projet de recherche européen Econogene, dont le but est justement de fournir des recommandations sur la manière d'assurer la conservation durable de races autochtones de chèvres et de moutons, que des systèmes d'information géographiques (SIG) ont été appliqués à l'analyse spatiale de données génétiques.

L'objectif principal de cette recherche est de montrer que les SIG sont utiles dans le but de fournir des hypothèses de travail alternatives susceptibles d'aider à comprendre le fonctionnement des ressources génétiques animales. Des outils et des méthodes ont pu être appliqués à des problématiques bien spécifiques en analyse de données moléculaires.

Tout d'abord, l'analyse spatiale exploratoire des données permet de traiter de grandes quantités d'informations et d'en révéler les structures spatiales sous-jacentes. Cette approche rend possible l'intégration de nombreux paramètres, de distinguer des pistes nouvelles pour la compréhension des phénomènes de dispersion des ressources génétiques, et d'en extraire des lignes directrices en vue de recherches plus approfondies.

Deuxièmement, sur la base des éléments mis en évidence par l'analyse exploratoire, les règles de la sémiologie graphique ont été appliquées à la représentation cartographique des variables moléculaires. Cette démonstration a pour but de favoriser l'amélioration de la qualité de la production cartographique en géné-

tique des populations, et ainsi de faciliter la lecture, l'interprétation et la compréhension de l'information. A une période où les questions de communication sont primordiales, et plus particulièrement pour des domaines comme la génétique qui touchent de près le grand public via leur rôle dans les questions d'alimentation et de santé, il est essentiel de produire des documents dont le message est clair.

Enfin, c'est l'analyse des pressions environnementales et de l'adaptation génétique qui fait l'objet du troisième volet. Tout en appliquant un principe de modélisation simple des processus naturels, l'analyse spatiale alliée à une méthode statistique est appliquée afin de détecter des signatures de sélection naturelle au sein du génome. Ceci est réalisé en mettant en relation les caractéristiques environnementales des zones dans lesquelles les races étudiées sont élevées, avec des régions précises de leur génome. Les résultats trouvés sont en bon accord avec ceux fournis par une approche standard de génétique des populations et montrent ainsi que la science de l'information géographique offre un moyen complémentaire d'analyse des processus évolutifs.

Appliquée à l'étude de l'ours brun de Scandinavie, la méthode montre également qu'elle peut être appliquée dans le cadre de l'élaboration de cartes d'habitat potentiel, une autre facette de la biologie de la conservation. A cette fin, on utilise les variables environnementales qui ont un effet sur le génome et qui constituent par conséquent des prédicteurs pertinents.

TABLE OF CONTENT

Note about the references to the literature

Most of internet references (URLs) used in this manuscript are mentioned in footnotes only, and do not appear in the literature cited. Full papers which are not published in journals and which are made available through the Internet are mentioned in the literature cited.

Moreover, complete references to literature mentioned in footnotes are indirect sources of information (by way of expert's recommendation), in opposition with the literature cited displayed from page 153 which has been directly consulted.

Glossary

Terms written in **bold italic** refer to the glossary.

GISCIENCE AND GENETIC DIVERSITY

1

INTRODUCTION

Much of the world's information is geographic in nature. People, animals, plants, objects, etc., are dispersed in space and interact in that space. Genetic information being linked to living organisms can therefore be partially characterized by geographic coordinates. The pairing of both genetic and spatial information is very well illustrated by a recent «Genographic» project, launched in part by The National Geographic Society and the IBM Corporation with the goal of collecting and analyzing more than 100'000 samples of *DNA* in order to trace the origins and to map the movements of humans during the last 60'000 years. The idea was first popularized by Luigi Cavalli-Sforza, Paolo Menozzi and Alberto Piazza in «The History and Geography of Human Genes» (Cavalli-Sforza *et al.*, 1993) in which they systematically relied on geographical maps to show how the frequency of human genes is evolving from one population to another across the world.

In this work, through the examination of applied case studies in domestic and wild animal genetic resource conservation, I attempt to :

- bring to the fore the geographic dimension of genetic data;
- demonstrate how the application of Geographic Information Science (GIScience) is relevant for characterizing aspects of *genetic diversity*;
- illustrate that explicit consideration of spatial information may possibly prove to be beneficial for the understanding of evolutionary processes.

Exploiting the geographic dimension of genetic data is not new. Sewall Wright and other cofounders of the field of population genetics considered geographics in their work from the 1930s, as they were studying the distribution of *allele* frequencies under the influence of the four evolutionary forces, namely natural selection, *genetic drift*, *mutation* and migration. Basically, the main use of spatial information was to calculate geographical distances for comparison to genetic distances. Since then, there has been much advance on the notion of geographic distance towards more realistic and sophisticated definitions. In particular, several authors proposed the use of the «Ecological distance» (Michels, 2001; Vuilleumier, 2003; Ray, 2005). Provided largely for context, these specific advances are outside the scope of the present research.

The novel elements of this work first consist of a GIScience approach to linking *genetics* and geographics, obviously placing emphasis on geographic information unlike most studies in this domain. Then, spatial exploratory analysis hints at opportunities associated with the existence of the rapidly increasing number of very large genetic geo-databases. A third innovation is to propose rules for genetic data representation (cartography). And finally, as core of this GIScience contribution to *molecular genetics*, a simple modelling approach with recourse to spatial analysis is applied for the detection of natural selection signatures within the *genome* of sampled animals, independent of all existing models in population genetics.

Before getting to the core, it is necessary that we concern ourselves with the inter-disciplinary nature of this work. As is common in research that lies at the intersection of multiple disciplines, clear and consistent definitions of both concerned disciplines must be adopted.

GISCIENCE AND INTERDISCIPLINARITY

GIScience is inherently interdisciplinary, being a field that provides tools useful through their application to solving problems within other disciplines. Indeed, it has long been applied for a multiplicity of uses in land survey, hydrology, archeology, anthropology, transportation, etc. In this sense, Geographical Information Systems (GIS) are considered to be applications-led technology (Longley *et al.*, 2001). Consequently, Geographic Information (GI) scientists commonly find themselves as «guest» in «host» disciplines in order to best exploit GI analysis tools and methods. Likewise, they are used to adopting an intermediate language which gives the opportunity to be understood by both the GI community and scientists of the investigated discipline. Of course, previous research has focused on how to elaborate connections between languages or how to create dedicated interdisciplinary languages (Paquet, 2001), but it is far from appropriate to develop such a language here. It is probably sufficient to clearly state that this research is not necessarily adopting terminology common to the field of *molecular genetics*.

For many years, GIScience turned toward environmental modelling (Goodchild *et al.*, 1993), generally concerned with explaining basic features of GIS to demonstrate how they could be efficiently applied to fields related to the natural sciences (Caloz and Collet, 1997). However, until now, application to *genetics* has been absent. Despite its current predominance in life sciences, and its direct application to concerns of public society (health, food), genetics had heretofore remained outside the scope of research by the GIScience community. On the other hand, from the end of the 1960s on, biologists gradually appropriated GIS tools, mainly in ecology. Only since the mid-1990s, population geneticists and molecular biologists began to systematically make use of GIS to try to understand how geographical and environmental features were structuring genetic information (see chapter 3).

MOLECULAR GENETICS AS THEMATICS

Molecular biology is the study of biology at a molecular level. It was established in the 1930s, the term being first coined in 1938 by Warren Weaver, director of the natural sciences program for the Rockefeller Foundation. Since the late 1950s and early 1960s, molecular biologists knew how to characterize, isolate, and manipulate the molecular components of cells and organisms. These components include *DNA* (the repository of genetic information), *RNA* (a close relative of DNA whose main function is to serve as a temporary working copy of DNA), and *proteins*, the major structural and *enzymatic* molecule in cells (Morange, 2003).

This discipline chiefly concerns itself with understanding interactions between those elements and how these interactions are regulated. It engendered molecular genetics, a field focused on identifying the structure and function of genes at a molecular level. This subfield of *genetics* is differentiated from others including ecological genetics and population genetics. Molecular genetics targets molecular markers, which are specific genes or DNA sequences that can be used to identify

organisms, species, strains or *phenotypic* traits associated with them. Several types of molecular markers exist, each supplying means of measuring a given aspect of genetic diversity. The information produced can, among other things, be transformed into frequencies. Such information precisely constitutes the data that can be used on an individual or population level for spatial analysis.
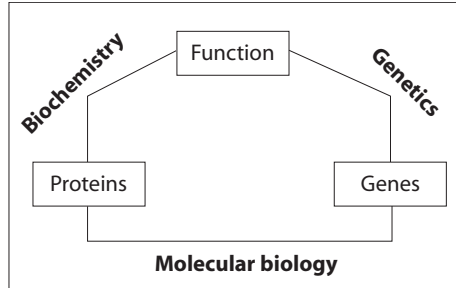
Fig 1.1. Schematic relationship between biochemistry, genetics and molecular biology. Source : Eric Lander, MIT.

Having defined the interdisciplinary context of this work, the following section distinguishes the well-known Geographic Information Systems and the GIScience discipline, which is generally unfamiliar for those not working in GI, including biologists. This is essential to understand the succession of the following chapters and to grasp several GIScience opportunities as discussed throughout this study.

REMOVING AMBIGUITY REGARDING GIS

Geographical information analysis may be one of the few fields for which sophisticated tools were developed and existed about 30 years before the science justifying its development was founded. As a matter of fact, this must have been the case for most of former sciences too, but the relative youth of this discipline makes it possible for us to live during the period (or an important part at least) in which papers about the realization of the field are being written in GIS journals.

Indeed, researchers began realizing that dealing with spatial information should involve much more then merely constituting an interest group of system designers, the Geographical Information Systems community. It is only in his founding paper of 1992 (Goodchild, 1992), in the International Journal of Geographical Information Systems, that Michael Goodchild formalized keynotes addressed by Roger Tomlinson in 1984 and by himself in 1990 at the Symposium on Spatial Data Handling, to pave the way for the Geographical Information Science.

GIScience consists of a two-sided field made up of its own technology driven research and development aspects closely related to Computer Science, the GIS part (software, *topology*, databases, standards, formats, etc.), and of a collection of methods and models that explicitly use the spatial referencing of each data case, the spatial analysis (Goodchild and Haining, 2004)[1]. The whole constitutes a science which is studying the fundamental issues arising from GI (Longley *et al.*, 2001).

---

1. These two sides can also be presented as two levels with data management on the one hand and data exploitation on the other hand.

Because of this late reflection on what constitutes GI research, we are challenged by a lack of integration of GIS and spatial analysis (Goodchild, 1992). This results in a gap between a trend of spatial data management for which geography is a mechanism for accessing information and whose works are technology-oriented, and a movement of spatial analysis interested in functionality and models for which geography has a fundamental role. The information management aspect is much more visible than the analysis one, probably because of the technology driven early stages of this discipline, as well as its business aspects. The development of technologies naturally led to a GIS industry (software producers) narrowly involved together with academic GIS users. While normal, it perpetuates vagueness about GI - where is science, where is business ? - and gives concrete expression to the present difficulty GIScience meets with being recognized as a full scientific discipline in academia. A perfect illustration of this is embodied by the different geoinformation magazines which are often talking about research issues but are always dealing with them through a commercial perspective. Another illustration is the so called «Environmental Science Research Institute (ESRI)» of Redlands California, today a pure business enterprise that conscientiously preserves the ambiguity expressed not only by its name, but also by its language which advocates to «Think spatially» and to concentrate on spatial questions instead of focusing on softwares and functions. Mike Phoenix, an ESRI Education Solutions Manager, suggested to me one of the best definitions of GIScience I know, which is «Towards spatial analysis...». Excellent, but it sounds a bit odd when stated by one of the world GIS software leaders. This ambiguity is obviously positive for ESRI's sales but simultaneously counterproductive for the clarity of what GIScience is (see for instance the paper of Dangermond, 1993). And at the same time, relative obscurity surrounds spatial analysis even if this is where the real intellectual core of GIScience lies. GI research needs to free itself from software production, and should depend less on the developments of this industry to define its progress. The emergence of GIS open source applications is likely to advance the present situation.

The point here is not to philosophize about the possible erring ways of GIScience, but as this research is concerning people in biological sciences too, it is necessary to make it clear that this work falls within the scope of a wider discipline than the application of specialized geographical functions. While scientific revolutions have arisen from the invention of new instruments (Galileo and his telescop, Franklin, Watson and Crick for the discovery of the double helix structure of DNA thanks to X-ray diffraction), in other cases, the discovery of new concepts produce considerable scientific advance (Wegener and the plate tectonics theory). The progression of science requires both aspects (Dyson, 1999). In this context, GIS are not only tools : their use belongs to a wider group of specific knowledges which have spatial information in common and are unified within GIScience.

Specifically, the present work reveals a GIScience angle on particular aspects of molecular genetics. It falls within the discipline of «GIScience» because GIS tools have been involved in the context of a scientific approach carried out together with biologists to assess their potential usefulness in discovering genetic diversity patterns and in bearing out hypotheses suggested by population geneticists.

EXPLOITING THE GEOGRAPHIC DIMENSION OF GENETIC DATA

Chapter outline

In its natural framework, genetic information is embedded within a geographic context. Plants and animals are directly influenced by the specific characteristics of their surrounding environment. Therefore, spatial information is a potentially important element to be considered in trying to understand genetic resources. This perspective is developed here through a presentation of two case studies drawn from the field of conservation biology which will illustrate how GIScience is likely to support and complement population genetics approaches as conducted in this research. For each case study, the general scientific challenge is presented followed by development of the overall approach.

THE SPATIAL DIMENSION OF GENETIC DATA

During his travel on the Beagle, Darwin noticed a gradation in the size of the beaks of the chaffinch birds living on different islands in the Galapagos (Darwin, 1989). This led him to the conclusion that the birds evolved to fit the environment in which they were living. Chaffinches' beaks differ because geographic isolation conducted to evolution towards a fit to various features of the different islands. As has been documented in many cases, geographic division can trigger evolutionary divergence and corresponding speciation (de Duve, 2005a).
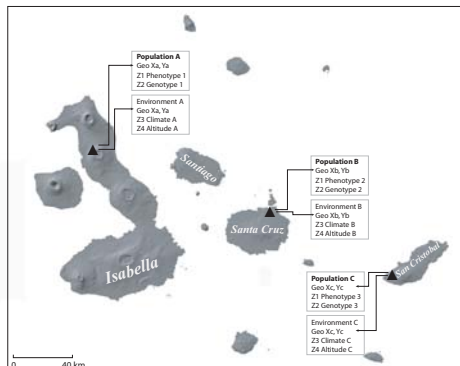
Fig 2.1. Darwin's chaffinch birds example in the Galapagos Islands : populations are differentiated by two-dimensional geographic coordinates (X, Y), but also by a third dimension characterizing the environment like altitude, land use, temperature, etc. (Z3, Z4) or the organisms located at this place (Z1, Z2). Sources : Marc Souris, IRD.

This famous example emphasizes the role of geography in discriminating phenotypic traits. A few years later, in 1866, Mendel conducted his foundational research of pea *genetics*, followed notably in 1910 by Thomas Morgan's documentation of a link between expressed traits and genetic material, from study of Drosophila. In 1918, Ronald Fisher initiated the modern evolutionary synthesis by proposing a genetic model that showed that continuous variation among characters could be the result of Mendelian inheritance. From that time on, the field of population genetics developed at Sewall Wright's, Ronald Fisher's and John Haldane's behest, attaching direct importance to spatial information, with the clear recognition that genetic differentiation is influenced by geography (Epperson, 2003).

Those landmarks in the history of genetics stress the fact that geography fulfils an important function in the context of the analysis of genetic information of living populations. At the same time, Darwin's observations on chaffinches' beaks highlight the decisive significance of environmental characteristics.

Presently, advanced molecular technologies make it possible to efficiently measure genetic information. As for geographic information, considerable advancements in computer science have led to the development of sophisticated software (GIS), in parallel with the elaboration of a wide variety of spatial analysis methods, making it possible to extract information from any environmental profile. The intersection of molecular biology and GIScience may enable novel and complementary methods of tackling some of the challenging issues related to evolutionary processes.

I chose to investigate two distinct contexts to assess the potential contribution of GIScience to *conservation biology* applied to animals. One is about the management of the genetic resources of livestock species which constitute the main data source of the research. The second case study is complementary and supplies *wild* species data, in comparison with domestic species data, in order to appraise the spatial analysis method proposed in chapter 7, and to demonstrate how genetic information is used to assess conservation measures applied to an endangered species in Scandinavia.

CONSERVATION OF LIVESTOCK GENETIC RESOURCES

Within the branch of agriculture, the livestock sector is currently facing stress due to issues with modern farming technique, and amplified by global climate change effects. All breeds developed over the past 8'000-10'000 years had valuable traits useful for adaptation to harsh conditions (drought, poor quality feed), or to tolerating parasitic and infectious diseases (Bruford, 2003). These traits are being gradually replaced by a few high production breeds which require specific inputs, skilled management and comparatively benign environments (Thrupp, 1998). Biodiversity is threatened because artificial selection and controlled reproduction, combined with natural selection, *gene flow* (particularly by cross-breeding), and *genetic drift* processes in populations of decreasing size, gradually lead to a general loss of genetic diversity (genetic erosion) within species and breeds, which

may potentially cause damaging effects like loss of breeding stock (extinction, new diseases) (Bruford *et al.*, 2003). The consequences are of global concern, and sustainable methods must be found to optimally conserve livestock genetic resources and diversity as we are confronted with an accelerating extinction crisis (Luikart *et al.*, 2003).

The Food and Agriculture Organization (FAO, United Nations) estimates that the world looses at least one breed of traditional livestock every week (Thrupp, 1998). As farmers focus on new breeds, many traditional breeds have disappeared. Of the over 3'800 breeds of cattle, water buffalo, goats, pigs, sheep, horses, and donkeys believed to have existed in this century, 16% have become extinct, and a further 15% are rare (Thrupp, 1998). These losses weaken the potential of breeding programs that could improve hardiness of livestock. A Convention on Biological Diversity elaborated by the United Nations Environment Programme (UNEP) charges nations to identify and monitor their biodiversity, to maintain, organize and share the resulting data, and to integrate the conservation and sustainable use of biological resources into national decision-making.

To face the threat, FAO initiated a global strategy for the management of Farm Animal Genetic Resources (*AnGR*) whose general purpose is to propose the introduction and adaptation of policies aimed at conserving livestock biodiversity. Among their goals, the FAO recommends the development and use of more AnGR (ensure food security, environmental development, and meet market demands), to identify and understand the genetic resources of each farm animal species, and to prioritize and conserve unique AnGR.

Different means have been developed to survey livestock breeds, among which molecular methods[1]. Since the beginning of the 1990s, the development of biotechnologies led to the elaboration of an array of different molecular techniques able to measure diversity at the DNA level (Karp *et al.*, 1997) and molecular approaches have been progressively recognized to be appropriate tools to measure, monitor and manage genetic diversity (Bruford *et al.*, 2003). To fit this challenge of recognition, a *genomics* research platform[2] applied to livestock was created to discuss what European research had to offer in the important field of genomics and to introduce the findings of ten projects funded under the European Union's Fifth Framework Programme (FP5) for research aiming at improving the sustainability of European agriculture. One of the projects presented is Econogene[3], conceived to promote the sustainable conservation of genetic resources in sheep and goats (see list of breeds in appendices 8 and 9), and provider of most of the analyzed data in the present work. From 2001, specialists in genetics, socioeconomics and GIScience have combined efforts and expertise to accomplish this multidisciplinary task. The study focused on small ruminants autochthonous breeds which have considerable economic importance in marginal[4] agrosystems

---

1. Other methods consist notably in developing demographic indices characterizing populations of breeds. The New Number of Females (NFN) is an example described here : http://www.tiho-hannover.de/einricht/zucht/eaap/nfn.htm (17.11.2005). There are also other factors to assess the status of endangerment of breeds which are described on this page : http://www.tiho-hannover.de/einricht/zucht/eaap/factors.htm (17.11.2005).

2. http://europa.eu.int/comm/research/agriculture/events/genomics_en.html (03.11.2005)

3. http://lasig.epfl.ch/projets/econogene/ (03.11.2005)

4. Marginal areas have experienced trends of abandonment and depopulation, as they provide fewer work opportunities and are endowed with relatively little transport and industrial infrastructure and fewer urban areas (Bertaglia, 2004).

in Europe. The surveying of their genetic resources makes it possible to highlight endangered breeds having high distinctiveness and priority for conservation, and to bring them to the attention of authorities so that tailored conservation measures are taken. Moreover an optimal exploitation of these local genetic resources may improve regional socio-economic situations and promote the maintenance of rural communities.

GIS have been involved in a central and strategic way in the Econogene project, with a structural and federative role to play on the one hand (see chapter 4), and an analytical role on the other. GIS are possessing the notable advantage of enabling the interdisciplinary connection of genetic, socio-economic and environmental information levels. Among other aspects, the project argued that GIScience was likely to provide adequate methods and tools to contribute to the understanding of genetic resources, mainly by taking into account its spatial variability, and giving the opportunity to characterize the surrounding environment in which studied organisms were raised, and to which they are adapted. Indeed, both aspects were expected to supply political decision-makers with additional indicators of when genetic resources management policies need to be defined in order to prioritize breed populations for conservation.

CONSERVATION OF ENDANGERED WILD SPECIES

To validate and extend analyses carried out on domestic species, a conservation genetics issue applied to a wild species was also considered. Thanks to a collaboration with the Laboratoire d'Ecologie Alpine in Grenoble, it was possible to work on Scandinavian Brown Bear data[1]. *Ursus arctos* has a holarctic geographic distribution from Spain to the United States. Since the 1800s, human activity caused a drastic reduction of its habitat range (Taberlet *et al.*, 1995). In Western Europe, its range is extremely patchy and several populations are facing the threat of extinction. Among them, the Scandinavian Brown Bear population was estimated at 5'000 individuals in the mid-19th century, of which 65% were living in Norway and 35% in Sweden. Their number rapidly declined due to extermination programs and in the early 1930s, the surviving population consisted of approximately 130 individuals. The Swedish policy changed by the end of the 19th century to save the bear from extinction, the most effective measure being to reduce and then to suppress incentive allowances for people to kill bears. Norway didn't make the same choice and bounties were suppressed only in 1930. Moreover, there was no full protection for the brown bear before 1972 in this country. The consequence was that the Norwegian population disappeared and the Swedish recovered (Waits *et al.*, 2000). Presently, the Scandinavian Brown Bear population numbers about 1'000 individuals and is still expanding (Waits *et al.*, 2000 and references therein). These bears are distributed in three (Manel *et al.*, 2004) or four (Taberlet et al, 1995; Waits *et al.*, 2000) geographical regions defined as female concentration areas where bears were continuously present during the population bottleneck. These areas are thought to represent surviving relict populations maintained separately because of strong philopatry of females (Waits *et al.*, 2000 and references

---

1. DNA provided by Jon Swenson, Norvegian University of Life Sciences, Aas.

therein). The definition of these zones was deducted from the analysis of genetic data (Taberlet *et al.*, 1995; Waits *et al.*, 2000); identified lineages and their spatial distribution made it possible to pick out potential units of conservation. Conservation units consist of areas in which sets of populations share a common genetic lineage and can be managed effectively by virtue of their common productivity and vulnerability. When required (extinction risk), animals can be transferred from similar evolutionary units (Avise, 1992 cited by Taberlet *et al.*, 1995).

In this context, geographical information is used to assess the pressure of the natural environment to identify regions of the ***genome*** of the Scandinavian Brown Bear possibly under selection. In addition, the results will be used in order to build maps of potential habitat, the purpose of such an approach being to refine the delimitation of conservation units according to the observed effect of diverse environmental influences on the respective populations.

GISCIENCE CONTRIBUTION TO THE UNDERSTANDING OF GENETIC RESOURCES

To reach the conservation goals which are described above, one has reason to analyze the dramatically growing amount of genetic data gradually produced by molecular techniques in the context of many research projects either on genome sequencing (human, chicken, cattle, etc.), or in population genetics and ***conservation biology***. These efforts are collecting a huge quantity of biological samples, most of which is related to living organisms and therefore spatially located within a geographic context.

In parallel with molecular approaches, it is highly desirable to apply a diversity of interdisciplinary approaches to understand such complex information. GIScience holds promise for being one of the appropriate ways to investigate genetic data from a point of view which is somewhat unique to the traditional field of life sciences. The geographic attributes of molecular data are worthy of attention and consist of an alternative means of studying the variation of genetic diversity and of analyzing natural selection processes.

Combining GIScience with ***molecular genetics*** technologies will increase the power of the latter by exploiting the spatial dimension of the information they provide, proposing an alternative perspective that may lead to improved understanding of genomic functions. The visualization (exploratory spatial analysis) and the representation (cartography)[1] of spatially distributed genetic data are likely to highlight patterns of diversity and thus offer additional concrete support for interpretation. Furthermore, spatial analysis may allow the discovery of relationships between genome regions and properties of the environmental surroundings for the examined populations of animals.

---

1. «Visualization» is the fact of showing a signal and to make it comparable to possible pre-existing mental images of the readers. «Representation» is the action of building the signal with the help of rules mainly defined by the semiology of graphics and of a given cultural context.

In the following chapters, I will describe how geographic information methods and tools were used until now to better understand genetic resources, and how it is presently possible to make use of GIScience in order to :

- Extract useful information from large, spatially-explicit genetic databases;
- Make spatial patterns of genetic diversity stand out;
- Produce improved thematic maps representing patterns of genetic diversity and showing other molecular genetics variables by applying the semiology of graphics;
- Detect precise regions within the *genome* which are under natural selection[1] thanks to the application of a simple modelling approach.

The fulfilment of these objectives will contribute to answer the following questions about genetic information :

- Does the exploration of the spatial dimension of molecular data usefully and efficiently contribute to facilitating the understanding of the evolutionary history and of the patterns of diversity of the analyzed breeds ?
- Is it possible to make use of spatial analysis to detect parts of the genome likely to be under natural selection ?
- If yes, do these results confirm the outcome of standard approaches developed and applied in population genetics ?

About geographic information :

- Is it also possible to enrich GIScience methods and tools through the combination with molecular genetics ? If yes, to what extent ?
- Does GIScience has a role to play in the challenge of understanding evolution and life as a profound process ?

And about modelling in natural sciences :

- Can a simple modelling approach be applied in order to obtain significant results ?

IMMERSING INTO MOLECULAR GENETICS

Undertaking such an interdisciplinary task with solely a GIScience and general natural sciences background implies a rapid and efficient engagement with molecular genetics to learn about the field's technologies and the information they produce. Moreover, this phase was important because it was necessary as well to know more precisely to which extent one could expect geographical information to be useful when applied to genetic data. In parallel with the many references to the literature, the first orientation and impulse was generated during early detailed discussions in particular with Dr Pierre Taberlet (LECA, University Joseph-Fourier in Grenoble), Profs. Laurent Excoffier (CMPG, University of Bern), Nicole Galland (DEE, University of Lausanne), Paolo Ajmone-Marsan (IZ, Catholic University of Piacenza), Godfrey Hewitt (SBS, University of East Anglia Nor-

---

1. This will be named «detection of signatures of natural selection».

wich), Jérôme Goudet and Dr Alexandre Hirzel (DEE, University of Lausanne). In addition, frequent and continuous research exchanges with Christian Parisod, Prof. Nicole Galland's Ph.D. student, provided increased exposure to the diverse knowledge required for undertaking this work.

From subsequent regular meetings with scientific partners involved in the Econogene project (see chapter 4), some primary issues became apparent, including the very large amount of genetic data of various kinds to be analyzed, and corresponding logical expectations about the potential findings to be revealed by visualization techniques. Also related to that subject, it emerged that only a few very general working hypotheses were existing regarding the anticipated behavior and the spatial distribution of genetic data. Finally, consideration of the spatial dimension of genetic data was found not to be completely unfamiliar, however most partners were unaware of the nature of geographic information, of how projection systems are used, and especially of the inherent constraints of a coherent spatial database needed to manage this considerable amount of data.

Therefore, these observations have led to a principal methodological choice, that is to resort to exploratory analysis to investigate genetic spatial information. With chapter 4 being dedicated to the description of molecular data and their management in a spatial context, chapter 5 focuses on the exploratory spatial analysis of genetic data (ESDA) and means by which knowledge can be extracted from large geographic molecular databases. Chapter 6 shows how to display on geographic maps the promising patterns ESDA revealed to observe, and thus addresses cartography and aspects of representing genetic data. Review of population genetics literature revealed weaknesses on that subject, particularly about adapted geographic contextual information which is likely to facilitate the interpretation of patterns of genetic diversity. Exploratory facets are also present in chapter 7 in which we move forward with spatial analysis, but for which this aspect remains in the background of a new context of modelling approaches within natural sciences, and of the search for selection signatures within the genome. Paradoxically, the exploratory aspects of the method are applied in a purely deterministic fashion in which we look for explanatory environmental factors corresponding to the existence of particular *molecular markers*.

TOWARDS LANDSCAPE GENETICS

3

. . . .
                                            .
                                            .
                                            .

Chapter outline

After having depicted the way GIS and biological sciences gradually came closer together through environmental modelling and biodiversity analysis contexts, this chapter describes the different works and researches which have been carried out at the intersection between GIScience and Genetics since the generalization of the use of molecular techniques.

First of all, and by way of preamble, it is interesting to note that until now, all existing research works and project references involving GIS and genetic data have been undertaken on biologists' initiative. I mean that all sources mentioned hereunder have been spread by biology, genetics, forestry or ecology journals, and that - with the exception of one or two references - these works have all been led by people with a biological background, though geoinformation specialists may have been involved. Apparently, genetic information didn't represent significative interest for geoinformation specialists, despite the fact that genetics and biotechnologies are representing fields of growing importance since a few years, and despite the obvious spatial nature of genetic information.

From a GIScience point of view, molecular biology lies amongst environmental processes. This is the first element we will review to make it come out that GI scientists were able to offer a technological context to approach the biodiversity issue, an important notion which can be considered according to genetic criteria or not.

GISCIENCE AND ENVIRONMENTAL PROCESSES
. . . .

For many years, a GIScience current was directed towards environmental modelling, with the constant concern of explaining GIS basic features to show how they could be efficiently applied to natural sciences related fields (Caloz and Collet, 1997). «How best to use GIS to change the way environmental modelling is being done?» to refer to the way Maidment (1996) addressed this issue. GIS have a very important role to play in environmental monitoring as considered by Larsen (1999), as long as it is not restricted to the creation of «pretty maps». Nevertheless, cartography in all likelihood constituted the first achieved integration of geographical information and any environmental issue.

Twelve years ago, in a book dedicated to «Environmental modeling with GIS» (Goodchild *et al.*, 1993), several authors acknowledged that GIS and Environmental modelling were well established methods, but that their integration was still an emerging field (Fedra, 1993; Parks, 1993). Today, talking of environmental model-

ling may even imply the fact that geographical information technologies are involved, this because since the 1990s the requested integration has been massively realized in forestry, hydrology, atmospheric sciences, risk and hazards management, geomorphology, agriculture, etc. This had also been the case in ecology, but in another way, as ecologists gradually appropriated GIScience technologies since the first steps of ecological modelling in the late 1960s. This process is well explained in an excellent review of GIS involvement in ecological systems written by Hunsaker *et al.* (1993).

Ecology and habitat modelling

Carrying out an original evolutionary biology research on a cricket species for which they notably had recourse to GIScience methods to interpolate morphological traits (Kidd and Ritchie, 2000), David Kidd and Michael Ritchie wrote a very informative introduction about GIS use in biological sciences. They efficiently related different GI utilization purposes, from biodiversity hotspots location to habitat connectivity quantification and metapopulation viability modelling (see references in Kidd and Ritchie, 2000). Moreover, they highlighted studies for which researchers specifically resorted to GIS as a working tool and not only for one dedicated task, like cartography as often as not. In particular they cite Jaquet, Lachavanne and Lehmann (Lehmann and Lachavanne, 1997; Lehmann *et al.*, 1997) who examined a community structure by quantifying the relationship between the distribution, the biomass, and the growth form of aquatic plant species with water depth and sediment characteristics.

But as Kidd and Ritchie pointed it out, much researches using GIS are led in spatial ecology, especially in habitat modelling where considerable developments are made in interaction with advanced spatial statistics. Presently, habitat modelling probably represents the main use of GIS in biological sciences with applications in biogeography, climate change research, ***conservation biology*** for habitat or species management. This last issue is the first for which people in biology appropriated GIS tools. Antoine Guisan and Niklaus Zimmermann wrote an exhaustive review of the different existing modelling approaches, some of them implying the use of GIS (Guisan and Zimmermann, 2000). Moreover, it is interesting to note that in this paper, Guisan and Zimmermann are corroborating observations made in the first chapter about the lack of integration between GIS and spatial analysis, by picking out that GIS, despite being widely used in different types of spatially explicit studies, are still deficient in integrating generic statistical procedures for predictive purposes. «This is a serious flaw because not all statistically derived models are similarly easy to implement in a GIS environment». The same observation is also made by N.Tait (Centre for Epidemiology and risk assessment, University of Lancaster, UK) who regrets that «(...) this plea [for integrating sophisticated spatial analytical functionality] has not been heeded by the GIS software companies, which consider «spatial analysis» to be little more than advanced data manipulation» (Tait *et al.*, 2004). This observation made by scientists working in a thematic discipline resorting to GIS shows that Michael Goodchild's claim for moving «from system to science, to establish GIS as the intersection between

group of disciplines with common interests, supported by a toolbox of technology,...» (Goodchild, 1992) has still not totally paid dividends...

But people concretely acted at the intersection between GIS and ecology. Alexandre Hirzel (Department of Ecology and Evolution, University of Lausanne) developed Biomapper, a software conceived to produce predictive habitat maps based on the Ecological Niche Factor Analysis (ENFA) (Hirzel, 2001; Hirzel *et al.*, 2001; Hirzel *et al.*, 2002), and also to map dispersal barriers and corridors. This package is linked to Idrisi[1], an existing widely distributed GIS. This work is particularly interesting in the sense that Hirzel, as a biologist, really entered into a GIS technology functioning to complete and improve it, in order to meet his needs. It is rare, and one rather observe GI scientists embark themselves on biological matters.

These ecological works aiming at modelling habitat are a part of landscape conservation. This discipline encompasses the planning and the management of wildlife and scenic resources in geographical and ecological systems (Aspinall, 1999). The two main concepts that have been adopted to guide and focus conservation actions are sustainability and biological diversity (Grehan, 1993 cited by Aspinall, 1999). Biodiversity is precisely a concept which was first based on descriptive works (morphology, traits or landscape units inventories) and in a second time integrated genetic information. The next section shows how this major development occurred since the late 1980s.

Biodiversity

Save ecological applications, biology set a second foot in GIScience after the concept of biodiversity, coined by W.G. Rosen, was brought to the attention of a wide field of scientists on the occasion of the «National Forum on Biodiversity» held in Washington D.C. in September 1986. This is the period when an important development at the intersection of GIS and ecology was made between 1987 and 1989 when the Gap Analysis concept and later Program (GAP) was invented by J.M. Scott (Idaho Cooperative Fish and Wildlife Research Unit), to develop predictive information that can be used to manage biological diversity so that ordinary plant and animal species will not become threatened with extinction. The product contains - it is still distributed - a wide range of tools and procedures, standards for classifying natural vegetative communities, and satellite images. It had an important impact and was widely used in conservation studies in the 1990s (Scott, 1993[2]).

Maintaining the maximum degree of biodiversity is one objective in nature conservation, and GIS have been involved to develop, manage, maintain and analyze the information base that supports strategies and actions in biodiversity conservation (Aspinall, 1999, and references therein). Analyses often consist in evaluating geographical patterns of diversity (biodiversity maps) generated from biological variables such as vegetation, vertebrate distributions, etc. (McKendry & Machlis, 1991), or in habitat modelling (Jones *et al.*, 1997). Papers with promising titles such

---

1. http://www.clarklabs.org/ (07.11.2005)
2. A GAP bibliography with more than 600 citations : http://www.gap.uidaho.edu/Literature/Bibliography.htm (04.11.2005)

as «The role of GIS and environmental modelling in the conservation of biodiversity» (Mackey, 1996) or «The use of geographical information systems in biodiversity exploration and conservation» (Jones, 1997) don't show a drastic evolution of GIS use in biodiversity topics. This is in fact the notion of biodiversity that evolved with the integration of genetic data and genetic diversity (the sum of genetic information contained in the genes of individual plants, animals, and micro-organisms) to complement species diversity, ecosystem diversity and cultural diversity which is determining how people interact with nature[1]. This new dimension of biodiversity possibly reinforced the role of GIScience, and especially the one of spatial analysis in the sense it multiplied in a phenomenal way the number of organisms' informative elements to be tested in relation to geographic and environmental information. Next section provides insights about the main stages of how the geography of genetic information was gradually taken into account.

EXPLOITING THE SPATIAL COMPONENT OF GENETIC INFORMATION

Population genetics

Considering genetics only, the study of spatial structures is existing since a long time. Indeed, in 1931 Sewall Wright developed adaptation and evolution models which were incorporating spatial distribution and distance considerations (Epperson, 2003). Distance between populations or habitats remains a central issue in spatial genetics as the main reference models in this discipline directly refer to, or are constrained by it (genetic isolation by distance, stepping-stone model and infinite-island model) (Epperson, 2003; MacArthur & Wilson, 2001). A lot of different statistics in which distance is playing a role were developed within geographical genetics (Epperson, 2003). For instance, the well known *Mantel test*, developed in 1967, allows to test the association of one set of pairwise measures with another. This was applied to compare geographical with genetic distances (Epperson, 2003, and references therein) to find out if distance from a source was likely to explain genetic diversity gradients. These aspects will be developed in the «Landscape genetics applications» section (page 27).

Towards landscape genetics

According to Luigi Cavalli-Sforza, Arthur Mourant was the first to have the idea of making geographic maps of *gene* frequencies and to use them extensively (Cavalli-Sforza *et al.*, 1994). Mourant led original works on blood groups and their hereditary clinical, social, and geographic patterns he published in «The Distribution of the Human Blood Groups» (Mourant, 1954) which long was regarded as a pioneering work. He notably discovered that Basques were the more ancient

---

1. http://www.fieldmuseum.org/biodiversity/index.html (04.11.2005), Department of Environment, City of Chicago.

inhabitants in Europe, this being revealed by the highest frequency of the RH-blood group in comparison with the other populations.

In the 1950s Luigi Cavalli-Sforza had the idea that one could map the worldwide geographic distribution of the genes. The application of this idea led to the writing of «The History and Geography of Human Genes», with his colleagues Paolo Menozzi and Alberto Piazza, which proposes an interpretation of the understanding of how humans left Africa and populated the rest of the world, and also to the detecting of ancient migrations, as for example the migration of Neolithic farmers from the Middle East towards Europe. This book is really important because it is bringing concrete arguments against dubious theories developed by Charles Murray and Richard Herrnstein in the famous «The Bell Curve» published in 1994, which claims in particular that race and class differences are largely caused by genetic factors. Stephen Jay Gould notably fought this biodeterminism with relentlessness and firmly argued in an edition of the New Yorker (november 28, 1994) that the current evidence showing heritability of IQ (based on one study only in «The Bell Curve») did not indicate a genetic origin to group differences in intelligence. Moreover, Gould maintained that «The Bell Curve» was not published at this moment by chance and that the period was particularly favorable to support «conservative ideology»[1]. This is in this particular context that the book of Cavalli-Sforza and his colleagues was published, and the new information it brought allowed to shed light on the delicate issue of heritability and relationships between populations. «The evidence it contains may carry enough weight to flatten Murray's thesis once and for all» (Subramanian, 1995). Indeed, the main conclusions of «the first genetic atlas of the world» (according to the Time magazine of january 16, 1995) are that if the genes for surface traits such as coloration are left out of account, the human «races» appear to be noteworthy analogous under the skin, and the genetic variation among individuals is much more important than the differences among groups. The diversity among individuals is even so enormous that the whole concept of race becomes meaningless at the genetic level. Finally, Cavalli-Sforza and his colleagues stated there was «no scientific basis» for theories touting the genetic superiority of any one population over another. But how did they work to come to these conclusions ? The famous collection of *gene* frequency maps stemmed from the idea to test the hypothesis that early farmers came to Europe from the Middle East (Pringle, 1998; Zeder and Hesse, 2000; Luikart *et al.*, 2001; MacHugh and Bradley, 2001). The migration of farmers would have generated circular gradients of gene frequencies around the region of origin (Menozzi, 1978, cited by Cavalli-Sforza *et al.*, 1994). To construct the maps, the authors used genetic information accumulated during fifty years and examined over 110 different inherited traits (blood types, proteins, DNA markers), in over eighteen hundred primarily aboriginal populations. Each gene and its diverse expressions (the *alleles*) were considered separately : the element chosen to be spatially represented was the proportion of a given allele in the different populations. Very roughly described, the applied standard procedure was to represent the frequency of the alleles on maps according to the locations where the studied populations were sampled, and the points of equal gene frequencies

---

1. «The Bell Curve» publication coincided with an unprecedented period of restrictions in social expenses during previous decades in the USA. This was also the moment when the famous and controversial republican representative Newt Gringrich was elected (Gould, 1996).

were connected by lines or «isogenic curves». Two reasonings are proposed, first that the mapping of alleles is useful to understand facts about the allele itself, in particular about its evolutionary history and effects of evolutionary factors like natural selection or **mutation**. Second, the correlation of allele frequencies with environmental parameters can be determining to discover specific genetic adaptations. We will look again, more in details, into both approaches in chapter 6 about sheep and goat genetic diversity cartography, and in chapter 7 about natural selection signatures. Anyhow, Cavalli-Sforza's works do represent the main achieved ones in the domain this research is related to. Only the geographical information system is a notion which is not evoked in this book. But spatial analysis methods are, as variography for instance when a choice had to be made between interpolating or smoothing surfaces between sampling points.

Considering other works, variography precisely was a method used to define specific genetic diversity zones in several studies. For instance, Hoffmann *et al.* (2003) centred their effort on the use of «The GIS technology Kriging» they applied on a differentiation index based on **nucleotide** diversity. This was not made in a predictive way, but to define *Arabidopsis thaliana* areas of similar diversity across Europe. This research is interesting because GIS aspects are well investigated and detailed, and not used just to represent information as it is often the case. On the other hand, a negative side is that the representation of the results on maps are not effective, and inadequate cares were brought to the choice of representation scales. This is typically the kind of unfinished cartography which led to the different proposals developed in chapter 6.

Variography was also probed by Gabriele Bucci and Giovanni Vendramin (Bucci and Vendramin, 2000) to delineate genetically homogeneous regions, to predict **haplotype** frequencies and to construct a «continental-scale availability map» of the European Norway spruce. The same year, Hamann and colleagues of the Department of Forest Science of the University of British Columbia in Vancouver exploited ordinary kriging to predict performance of seed sources at unsampled locations (Hamann *et al.*, 2000). They produced maps with dispersal probabilities of red alder in order to predict the location of top performing seed sources, the final goal being to develop seed-transfer guidelines and to delineate seed-procurement zones for forestry. Moreover, the authors suggested to explore the composition of the environment constituting the dispersal zones (using temperatures and precipitations) to test if the genetic differentiation would fit the ecological one. This **gene** ecology perspective, directly related to the content of chapter 7, was exploited by Skøt *et al.* (2002) in their investigation of the interaction between environmental characteristics of a forage grass (*Lolium perenne*) and its molecular information. Marker-assisted selection has the potential to increase the efficiency of breeding according to a given interesting property : this paper focused on cold resistance and the aim of the research was to understand the ability of *Lolium perenne* to survive and grow at low temperature, to acclimate to cold, to tolerate wind, snow cover and ice encasement. Six AFLP molecular markers that correlated in frequency with cold tolerance could be identified to be potentially involved in this process. In this context, the role of GIS was to display plants locations and to retrieve corresponding environmental variable values available on separate data layers. Applying the same reasoning, AFLP markers were also used to show association with salt tolerance in wild barley (Pakniyat *et al.*, 1997).

The last three mentioned papers made use of the «genecology» term, to refer of course to the contraction of «Gene» or «Genetic» and «Ecology». According to their web site[1], the establishment of «The professional field of Gene Ecology» was initiated by GenØk, the Norsk Institutt for Genøkologi in Tromsø. This more or less felicitous naming is including the study of the interaction between genetic information and environmental characteristics with the objective to understand how gene-modified organisms can affect ecological systems. It emerged because there was a lack of certainty about the consequences of manipulations aiming at improving characteristics of plants, animals and human beings for specific purposes. «How best to judge the potential environmental danger of a genetically modified organism» (Cyranoski, 2004) is one (biotechnological) way to consider gene-environment interaction.

The other is to examine the influence of the environment on the **genome** and try to understand how geographical and environmental features do structure genetic information. In this context, David Galbraith of the Royal Botanical Gardens proposed «to *place genetic diversity information into a spatial framework*»[2]. This approach was named «landscape genetics» and widely adopted by The Natural Resources DNA Profiling and Forensic Centre in Peterborough (Canada) where David Galbraith was working in 1995. This centre is to my knowledge the first to have systematically used GIS to analyze the geographical distribution of different genetic markers (see definition on page 43). The works of David Galbraith and Bradley White in particular, started from the observation that spatial issues in genetics traditionally solely dealt with geographical structuring and effects on populations of reduced **gene flow**, this being due to fragmentation of the landscape. Much more was required to be known about genetic processes in landscapes. To this end, GIS were used in order to try to separate the effects of natural selection from the ones of **genetic drift**, and many landscape genetics studies were applied to the Canadian's fauna since the second part of the 1990s.

In fact, during the last 10 years, several studies corresponding to the definition proposed by Galbraith have been achieved, but without referring to the *landscape genetics* designation. This is possibly due to the fact that the DNA Profiling and Forensic Centre of the Trent University didn't communicate immediately on this specific topic (the method), rather stressing on the concrete problematics they had to solve. Another explanation is perhaps the fact that the researchers of Peterborough were rather active in animal genetics than in plants (although they published about plants too) and that the communication between both fields was not really effective. For instance, when reading a contribution of Escudero, Iriondo and Torres (Escudero *et al.*, 2003), it appears that during this period, the latent landscape genetics still looked like a jigsaw in researcher's minds. About the analysis of plant genetic variation in an explicit spatial context, they stressed the fact that, as geneticists had recognized the importance of the interaction between genome and environment to better understand evolution (Berry, 1989) - that is to say to connect genetics and ecology (Jelinski, 1997) - it was urgent to materialize this synergy by integrating the respective knowledges, the meeting point between

---

1. http://www.genok.org/ (04.11.2005)
2. http://www.nrdpfc.ca/landscapegenetics.html. Papers written by researchers at the Natural Resources DNA Profiling and Forensic Centre are listed here : http://www.nrdpfc.ca/pubs.html (04.11.2005)

those approaches being obviously spatial analysis, in order to finally formulate management strategies and to be able to exploit this new tool in conservation biology. Berry's presidential address to the British Ecological Society in 1988 (Berry, 1989) was really important to stress the importance of gene-environment interaction, especially for spatial analysis, but we will go back over it in chapter 7. Let's see how the naming of the integration of ecology, genetics and GIScience was finally brought to an end.

This is in 2003, with the paper «Landscape genetics: combining landscape ecology and population genetics» written by Stéphanie Manel, Michael Schwartz, Gordon Luikart and Pierre Taberlet (Manel *et al.*, 2003) that the application field of the discipline was clearly defined. Making its roots date from de Candolle (1778–1841) precursory works, and Wallace's (1823–1913) early observations he made in the Malay Archipelago, they described landscape genetics as a new approach being a combination of landscape ecology and population genetics, likely «to facilitate our understanding of how geographical and environmental features structure genetic variation at both the population and individual levels, and [which] has implications for ecology, evolution and conservation biology».

An important point in this paper is that an explanation is emerging about the apparent hesitations - no diffusion of the appellation in the literature - around the landscape genetics field until the 2000s. According to this important paper, the spatial genetics approach had been limited by the *unavailability of enough molecular markers* to examine biogeography at a fine spatial scale. Moreover, this is only the possibility to combine elaborated spatial statistics, GIS and several types of numerous molecular markers «which enabled the amalgamation of ecological biogeography with molecular ecology to better understand population biology and evolution», and thus which finally allowed to give birth to *landscape genetics*.

Probably those elements had to be clarified because since this 2003 paper, landscape genetics is becoming a widespread designation to include all research about genetic data and exploiting their geographic dimension, this being confirmed by recent papers (Hirao and Kudo, 2004; Watts *et al.*, 2004; Spear *et al.*, 2005). This last article is particularly interesting in the sense that it perfectly illustrates the potential landscape genetics offers to researchers : it's up to anyone's imagination or inventiveness ! Moreover, this is one of the rare papers attributing so much importance to GIS tools. Working on 8 *microsatellite loci* of the blotched tiger salamander, Spear and colleagues tested whether landscape variables (elevation, wetland likelihood, cover type and number of river and stream crossings) were correlated with genetic subdivision (F_{ST}, an index of genetic structure) and then compared five hypothetical dispersal routes with a straight-line distance using *Mantel tests*. The results showed that environmental variables significantly influenced genetic differentiation in addition to distance : all models incorporating environmental variables even explained a higher proportion of variation in F_{ST} than distance alone. It allowed to raise important questions relative to salamander conservation and management «which would not have emerged without the inclusion of GIS landscape analyses in our evaluation of genetic structure». They also emphasized GIS role when saying that «however, geographical information systems (GIS) provide the framework in which high-resolution analyses of habitat variables can be performed» and that «GIS-based analyses should add to the

emerging field of landscape genetics (...) to examine the extent to which landscape features influence genetic structure via dispersal».

Manel *et al.* (2003) and Spear *et al.* (2005) works constitute a full and direct recognition of GIS tools and methods' role in the framework of the analysis of genetic information in a spatial context. The importance here conferred to the management and the analysis of geographical information makes GIS a compulsory component of landscape genetics, together with *molecular genetics* and ecology.

LANDSCAPE GENETICS APPLICATIONS

In addition to the case studies presented in chapter two, the following section gives concrete expression to how the combination of GIScience and genetics can be applied to solve real life problems. The first example is halfway between fundamental research and applied issues in conservation biology, and constitutes an appropriate transition before talking about applications only.

Dispersal modelling and ecological distance

In their paper, Spear and colleagues' research highlights a specific ecological movement, when resorting to dispersal routes across landscape (Spear *et al.*, 2005). Indeed, GIS were used to exploit statistical instruments developed by population geneticists in order to study how animal or plants were moving - or dispersing - between habitat patches, and to assess the role of the environment on the spatial genetic structure of these populations. The idea is that the information contained in spatial patterns may be captured in pairwise measures of genetic correlation as a function of physical distance (Epperson, 2003). For a long time, euclidian distances were used and compared to genetic ones by applying the *Mantel test* and its derivatives. For instance, Olivier Hardy developed a software for Spatial Pattern Analysis of Genetic Diversity (SPAGeDI) designed to characterize the spatial genetic structure of mapped individuals or populations using *genotype* data (Hardy and Vekemans, 2002). But with the progressive coming of GIS in ecology, the notion of distance was improved and became more realistic. Séverine Vuilleumier developed a sophisticated model to simulate the dispersion of the greater white-toothed shrew in a highly fragmented landscape. She used a digital landscape data model to define animal moving rules and programmed various behavior possibilities to obtain different dispersal scenarios (Vuilleumier, 2003; Vuilleumier and Perrin, in press; Vuilleumier and Metzger, in press). Alexandre Hirzel (2001) also developed such a kind of model to simulate the spreading of an invading species (the alpine Ibex) across a complex landscape. His HexaSpace software is based on a cellular automaton approach, the cells being hexagonal and characterized by a few properties like a carrying capacity, impermeability rates quantifying exchanges between adjacent cells, and population density. The spreading simulations are based on local population dynamics models and habitat suitability maps. Jean-François Arnaud (2003) also incorporated information about the structure of the landscape to obtain more realistic distances regarding the movement of individuals in heterogeneous environments. Nicolas Ray, collab-

orator of Laurent Excoffier at the Computational and Molecular Population Genetics Lab in Bern, developed an application (Ray, 2005) which is incorporating information about the structure of the landscape in order to obtain effective distances to be compared to genetic ones.

This kind of reasoning demonstrates, following the example of Michels *et al.* (2001) studying the dispersal of zooplankton in a system of interconnected ponds, that an effective distance - here calculated according to the water flows - provides a better approximation of the true rates of genetic exchange among populations than a mere euclidean geographical distance. In human-dominated landscapes, when habitat is gradually fragmented or even destructed, such approaches are applied in the context of landuse planning for example, to determine corridors for species that have to be preserved and protected in zones that are close by urban areas (Patthey, 2003; Vuilleumier, 2003).

Conservation genetics

Conservation genetics is a mixture of ecology, molecular biology, population genetics, mathematical modelling and evolutionary systematics (construction of family relationships). It is first a basic science because genetic relationships of the organisms under study have to be understood before management techniques can be applied by wildlife managers to preserve biological diversity. Organisms can be endangered because habitat destruction put a population at risk, or when the total population of a species becomes too small, making it more susceptible to stochastic events (natural catastrophes, environmental changes or *mutations*). This kind of event can cause sudden decreases in population size, and further reduction of their remaining numbers can sharply reduce genetic diversity. The example presented in chapter 2 about the Scandinavian Brown Bear perfectly illustrates a conservation genetics case (Taberlet *et al.*, 1995; Waits *et al.*, 2000; Manel *et al.*, 2004). It shows how genetic information can be used to identify and delimit geographical areas in which populations are homogeneous, facilitating transfers of individuals whenever required. Nowadays, conservation genetics principles seem to be globally applied all over the world without any problem, but it has apparently not always been so.

Wei Ji and Paul Leberg wrote a paper about GIS and conservation genetics in which they explained a paradox that took place when the first biochemical techniques were applied to the study of genetic variation in the 1960s (Ji and Leberg, 2002). In fact, much of the literature about molecular techniques were inaccessible to natural conservation agencies «that typically do not include population geneticists on their staffs». Moreover, the journals in which publications about molecular genetics were made did not have a resource management focus. So consequently, conservation scientists and planners were not aware of the existence of genetic data pertinent to issues they were addressing. On the other hand, Ji and Leberg observed that genetic information was of little use unless it was analyzed in relation to habitat and land use across ecosystems. And no tools were available to manipulate site-specific genetic data for understanding relationships between genetic variability and environmental factors. Amazingly, the credit of GIS during its coming in conservation genetics - aside from its technical and scientific contri-

bution - was to build a bridge (Clinton, 1996) between wildlife managers and molecular biology as they made it possible to give genetic data a sense by localizing it and by allowing to characterize the surrounding landscape of concerned organisms. In their own study which resorts to a simple use of GIS features, Ji and Leberg mainly stressed the fact that just as well the understanding of the distribution of both genetic diversity and areas managed for biodiversity preservation are required to realize a correct assessment of the conservation status of genetic diversity. In addition to this integration of genetics and environmental variation, they insisted on the possibility provided by GIS to work on a regional scale in opposition with smaller scales[1] usually found in ecological studies.

Other applied studies in conservation genetics don't take further thoughts about all aspects of the GIS contribution to the discipline, excepted the very practical ones about management facilities. The Natural Resources DNA Profiling and Forensic Centre in Peterborough made use of landscape genetics to incorporate moose DNA profiles of functional and neutral loci into GIS databases. This was helpful to help to decipher the effect different habitat types, hunting regimes and anthropogenic pressures were having on the diversity and fitness of local moose populations (Wilson *et al.*, 2003). The same research group led different applied studies of the same kind on bear, wolf, swift fox, caribou, frog, fish, and whale (see page 25, note 2).

The Department of Fish and Wildlife Resources in Moscow (University of Idaho, USA) is active in achieving studies and providing management decisions notices about threatened species. Adams *et al.* (2003) developed a method to detect hybrids among red wolves (endangered) and coyotes populations in North Carolina. They combined genetics using ***mitochondrial DNA*** sequences with GIS technology to produce maps of the spatial distributions of hybrids compared to the respective original populations.

In fishery, GIS and genetics are evenly used to define essential fish habitat (EFH) (Valavanis *et al.*, 2004), or to monitor species dispersal (maps) through chemical tags and genetic comparisons in order to supervise population movements and to measure the spread of species (Palumbi *et al.*, 2003).

It is impossible to be exhaustive in mentioning all works applying the combination of GIS and molecular genetics in animals and plants, this approach being now widely used in conservation biology (Apps *et al.*, 1994; Greene *et al.*, 2004; Spear *et al.*, 2005; Manel *et al.*, 2004). Another reason is that an important proportion of these studies are published or distributed from specialized agencies in the form of professional reports. Anyway, the main observation is that GIS technologies use is increasing in conservation issues, but the way they are exploited still remains elementary (cartography) in comparison with the possibilities described and tested in landscape genetics researches.

The function of the next two sections is merely designed to inform about the exploitation of GIScience and genetics combination in close but different domains.

---

1. According to the cartographic definition which takes into account the ratio between the real size of an object on earth and the size of its representation on a map, in opposition with the geographic definition of scale which refers to the spatial extent of the study area (Bian, 1997).

The few information provided hereunder should give a basic but sufficient insight about how this geogenetic synergy can be applied to other real life issues.

Plant genetic resources and agrobiodiversity

The International Plant Genetic Resources Institute (IPGRI)[1] or the International Center for Tropical Agriculture[2], are research institutes with a mandate to advance the conservation and use of genetic diversity in plants for the well-being of present and future generations. Genetic resource management is a complex process that includes a number of mutually dependent stages, from the identification of a target *gene* pool for conservation to the use of genetic resources. Many of these activities not only generate but also require georeferenced data. The application of GIS to analyze these data make the process more efficient, for instance in merging genetic diversity information with population density, climate, topography or soils data, adding value to genetic resources. The use of GIS and spatial modelling techniques permits of course to study genetic diversity, to monitor genetic erosion, to select potential collecting sites, to develop conservation strategies, with a global goal which is to enhance the use of genetic resources (Guarino *et al.*, 2002).

Molecular epidemiology

In june 2001, the Virginia Bioinformatics Institute at Virginia Tech organized a conference on «Applications of GIS to Bioinformatics» (Garon, 2001). The interest of the event was that researchers from both fields were present, and not anybody as Michael Goodchild was invited to give his advice on bioinformatics and GIScience cooperation. Geoffrey Jacquez, scientist working for a private biotechnology company, led the discussion on a crucial issue when he mentioned the potential enormous breakthrough by joining *genomics* and proteomics with GIS and spatial epidemiology. The contribution of this combination can be really important to study disease patterns in human populations, to determine how variations in genes combine with environmental and other factors to induce cancer particularly. Jacquez emphasized the fact that this new dialog was permitting to gain a link between individual-level processes tracked by genomics and proteomics, and population-level outcomes tracked by GIS and epidemiology. This is precisely what allows to improve monitoring, quantifying, and predicting human-health consequences associated with the environment (Jacquez, 2004). He added that «After years of using GIS data to track diseases in populations in terms of the what, where, and when, the integration of bioinformatics data, genomics and proteomics data telling the story of what takes place at the cell and sub-cell level in individual diseased organisms will soon enable epidemiologists using GIS to capture the how of disease outbreaks.»

The promising perspectives of GIS and bioinformatics combination stimulated the achievement of several studies - even before the Virginia Tech conference natu-

---

1. http://www.ipgri.cgiar.org/ (07.11.2005)
2. http://www.ciat.cgiar.org/biotechnology/ (07.11.2005)

rally - which ended up on the publication of many papers about disease transmission or causes like tuberculosis for instance (Richardson *et al.*, 2002; Munch *et al.*, 2003; Moonan *et al.*, 2004), or malaria (Schellenberg *et al.*, 1998) and cancer of course (Sokal, 2000; Jacquez, 2004). In his conclusion about current practices in the spatial analysis of cancer, the latter speaks for the development and the application of a «higher dimensional GIS» (Loytonen, 1998) or a Space-Time Information System (STIS) to better represent space-time dynamics and «to provide a rich framework for the generation and the evaluation of epidemiologic hypotheses founded on the exploration of space-time disease patterns in relation to their putative causes and covariates» (Jacquez *et al.*, in press). This would indeed allow GIScience to take an important part in this challenge for spatial analysis of cancer and other diseases, which is to fully exploit the temporal dimension as this information will become more easily available.

SUMMARY

It clearly appears that landscape genetics has only been developed, used and improved on biologists' initiative. Geographical information scientists participated as supporting actors, early contributing to lay the foundations of this field when working with ecologists.

From a GIScience point of view, the few elements available in literature about the management and the spatial analysis of genetic data consist in reviews of how GIS have been used in conservation biology (Aspinall, 1999), as a tiny part of the global and classical environmental modelling category, in which landscape genetics didn't find its place yet.

Therefore, it is important to contribute to landscape genetics research by turning a GIScience critical gaze on this field, and to propose improvements on the aspects for which GI experts assessment only is likely to make things evolve (see chapter 6).

Managing Molecular Geodata *4*

. . . .
. . . .
. . . .
. . . .

Chapter outline

This chapter is about the way the information used for farther landscape genetics analyses is prepared, collected and managed. It is also about the nature of genetic data to make it comprehensible what is the meaning of the presented investigations.

It starts with explanations to understand GIS community interest for genetics. Then fundamentals of geodata modelling are contextually exposed, references being made to the Econogene project and to the geographic objects constituting its reality.

Finally, a brief introduction to the variety of molecular markers exploited in the Econogene project is preceding explanations about concepts and techniques at the roof of most of molecular markers discovery.

GISCIENCE INTEREST IN GENETIC DATA

As explained in the previous chapter, the GIS community was involved in natural sciences through diverse topics in environmental modelling since about 40 years, but never directly embarked upon genetics. It is now time to go beyond the step, first because on the one hand some specific characteristics of genetic data do constitute interesting problems to treat in GIScience (notably spatio-temporal issues raised by different *mutation* rates of molecular markers, to mention an aspect which is not treated in this research), and on the other hand because genetics related fields nowadays constitute a scientific key-theme which can now significantly progress also in the context of interdisciplinary approaches. As mentioned in the first chapter, GIScience has presently almost no general public image and would therefore fully benefit from increasing the value of its contribution by joining together with leading scientific fields. In fact, there is an additional argument which is to *take initiatives* in collaborating with biological sciences researchers to reinforce the works the latter have launched upon landscape genetics since 15 years, and to take advantage of this experience to raise new research issues profitable for both GIScience and genetics scientific communities.

GEOGRAPHIC DATA MODELLING

The present chapter is mainly dedicated to the description of the genetic information which is investigated farther, this in order to fully grasp the meaning and issues held by the maps and analysis proposed in the next three chapters. Important notions are presented to understand what is genetic information, how it is structured, how it is possible to measure it, and especially how digital data sets can be produced to be subsequently processed by GIS. With this aim in view,

before paying attention to the nature of this information, and focusing on the tools GIScience can get down to work, it is firstly necessary to state - or remind - some elementary knowledge about GIS implementation, as they are the instruments we will utilize to manipulate genetic data. Of course, it is not the point to compile a series of recommendations of GIS methodological aspects applied to natural sciences, this very important task having been achieved for a long time (Caloz and Collet and references therein, 1997; Burrough, 1998), but to emphasize the essentials. On the basis of these indications, the geographic context within which data we are interested in were collected will be described, as well as the way sampling was performed and the information stored.

Two key-observations do allow to fulfil this task. First, unlike what is usually perceived, GIS not only are applications developed to create maps or to analyze data, but actually «systems designed to acquire, to manage and to process spatial information» (Caloz and Collet, 1997). On one of its web pages[1], the International Plant Genetic Resources Institute supplies a good example to illustrate this typical underestimate of GIS role. Indeed, «Managing genetic resources with Geographic Information Systems» is a promising heading, but despite the employed vocabulary (managing information !), IPGRI only refers to data analysis, and not to the way genetic information has to be managed upstream.

Secondly, it is not possible to reconstitute the multiple details of the reality of an analyzed natural process within a GIS, thus it is necessary to resort to a simplification, in fact to the modelling of the geographic space which is considered in the context of a study. The «map» or rather the graphic representation of information constitutes the interface between the user and the data which is stored in a *geographic database*, the real - but at first sight invisible - heart of GIS (Longley *et al.*, 2001). This is crucial because it emphasizes the fact that spatial information is not an element that appears on demand, displayed on a screen when available in a pop-up menu, as if it was a part of the software's programming. Reachable through a GIS, the raw visual representation of spatial data is the result of choices made before and during an acquisition phase, and then captured into a database according to strict rules in order to get comparable information. This preparation stage «requires from the applied field to interrogate about the nature and content of relevant information, as well as about the transfer of classical investigation and analysis methodologies» (Caloz and Collet, 1997). It is made up of several classical components appearing on figure 4.1 and constituting a continuous cycle of geographic information.

Geographic reality is naturally too complex to be completely represented within a computer system. It means that objects of the real world have to be selected to simplify the reality. This transition from reality to a model takes part in the geographic information life cycle whose 6 main steps are described hereunder.

1. The structuring and possible hierarchical organisation of the elements constituting a reality and of their relationships permits to generate a *conceptual model*. When confronted to mere realities, the conceptualizing phase may be self-evident, doesn't even require a particular effort and often remains mental.

---

1. http://www.ipgri.cgiar.org/regions/Americas/programmes/gisforpgr.htm (07.11.2005)

2. The passage to a type of spatial data model requires the identification of the nature of the spatial objects and of the way they are fixed in the landscape (*raster* or *vector* data; see McNoleg, 2003).

3. Once the spatial model defined, data can be acquired. Aside from sampling that will be treated farther, data acquisition confronts system's designers and users to a necessary, rigorous and sometimes inconvenient ***standardisation*** stage so that the whole information be uniform and comparable. The strictness of standardisation will condition an optimal implementation of data and the quality of the forthcoming analysis. This step is not as trivial as it seems to be : from an external point of view, which often is the one of the applied discipline, it is difficult to realise the importance of standardisation and especially of the possible consequences of a lack of rigour at this moment of the project's development; indeed, it looks like many consider that work begins when analysis starts. It was particularly striking in the context of the Econogene project : during work plenary sessions, it was really a challenge to lay down rules to avoid subsequent problems, for instance about the capture of geographic coordinates. Users think it is a waste of time to suffer constraints like protocol application when acquiring information, but no doubt it is a necessary evil.
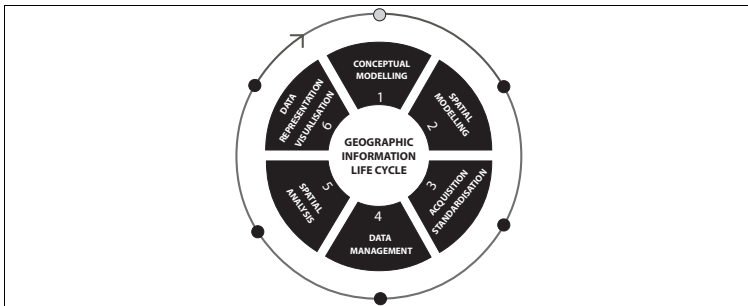
Fig 4.1. Geographic information cycles from conceptual modelling to results visualization. Sources : Pointet, A. *et al.*, 2004, CLAN Cultural Land Use Analysis Methodology Inter American Development Bank Ed., modified.

4. Data management consists in storing and treating spatial data, in fact making it available in a most adequate and comfortable way to allow spatial analysis;

5. Spatial analysis : as it will be illustrated in chapter 5, there often is a continuous *va-et-vient* between data analysis and representation, generated views allowing to adapt parameters or to produce temporary new working hypotheses in the context of interaction processes.

6. Data representation (see chapter 6).

These different steps are important to observe when designing a GIS, but it is necessary to pay attention to the fact that the system *is not an end in itself*. It is a basic infrastructure which is allowing an increase in value of data only through analysis, visualization and interpretation.

Before broaching the nature of genetic data, next section will go through the different components of the Econogene reality and explain the corresponding spatial data model.

DATA MODELLING IN THE ECONOGENE CONTEXT

The Econogene geographic reality is rather simple and didn't require the elaboration of any advanced conceptual model. It was mainly depending on projects objectives and related requirements.

Breed choice and implications on sampling

According to the Econogene context (described on page 13) and to its goals, it was necessary to get biological samples from sheep and goats spread in areas where better exploitation of local genetic resources and further diversification of the rural economy was likely to improve the socio-economic situation, and promote the maintenance of rural communities. Moreover, the choice had to focus on breeds not previously investigated and typical for marginal rural areas. The aim of this choice was to achieve a fuller representation of autochthonous breeds adapted to the extensive farming management of lesser developed regions in Western and Eastern Europe and in the Mediterranean area. To this end, breeds from the Central and Eastern European Countries (CEEC) as well as from North Africa and Middle East were also sampled in order to involve experienced scientists in these countries, and to ensure the availability of samples from relevant areas, some of which situated near original sites of domestication or near routes of migration into Europe. Indeed, Olivier Hanotte stipulated in a FAO conference about the Role of Biotechnology for the (...) Conservation of (...) Animal (...) Genetic Resources, that it was important «to ensure that breeds selected for conservation include populations from the geographic areas representing the different domestication centres where we would expect to find large genetic diversity and genetically differentiated populations. Animals and populations present at the geographic area of a centre of domestication will also be expected to be very distinct from the ones found at other centres of domestication. Also, the understanding of the geographic pattern of livestock migration from a centre of origin will allow the identification of populations present at the end of a migration route. It is expected that these populations will be genetically distinct from the populations present at the ends of other migration routes as a result of random ***genetic drift*** and/or the effect of local selection pressures.» (Hanotte and Jianlin, 2005). In addition to autochthonous ones, a few cosmopolitan breeds have also been included in the sampling, as reference material.

It appears that a lot of constraints strongly determined the geography of sampling. In fact, this breed-oriented sampling choice was not compulsory. Pierre Taberlet mentioned that a solution would have been to free the project from the breed notion and to sample animals within an initially defined regular grid throughout the Econogene area (personal communication, 2005). This would have granted a greater weight to the geographic distribution and ensured an homogeneous density of samples. This proposal is the expression of a current considering that the notion of breed is evolutional in essence[1], in opposition to the competing vision

---

1. Semestral review of the Société d'Ethnozootechnie, number 29, «Le concept de race en zootechnie», with notably a contribution of Raymond Laurans «L'évolution du concept de race en zootechnie». http://www.ethnozootechnie.asso.educagri.fr/default.cfm (07.11.2005)

tending to congeal the concept of breed according to observed fixed criteria. For Raymond Laurens, the breed is inevitably a changing notion because of the numerous parameters it is depending on and also because there is no discontinuity between breeds[1] [2]. According to this reasoning, a breed is not an entity to study and the interest rather lies in understanding how a given *phenotype* is obtained from genes, how genes do convey information (Pellegrini, 2005). Adherents of the breed as a static entity will probably always look for a better adequacy of animals to the notion of breed while defenders of the evolutional and morphological approach will tend to adapt its definition to the needs and constraints of breeding (Pellegrini, 2005).

Presently, the traditional concept of breed tend to be progressively replaced by functional morphology criteria. An animal is evaluated on the basis of milk production performances for instance, among which the quantity is important but also other criteria like fat or *protein* rates. This apparent dilution of the breed notion and the gradual recourse to the evolutional notion of functional morphology (Pellegrini, 2005) is an indication showing that the breed-oriented choice for sampling design was not the unique relevant solution. From the concurrent point of view, one must admit that, because of the socio-economic context, the adopted strategy was more reliable to assess the importance of given autochthonous breeds in specific marginal areas. A geo-centred sampling would have possibly caused the omission of certain situations not to be neglected.

Animals sampled

Between 2001 and 2003, three animals per flock from 11 farms spread over the traditional rearing area of each breed were sampled. The biological material collected was blood and hair (and even tissue when possible). The 33 sampled animals had to be unrelated. In the absence of reliable information on kinship, direct descents were in any case excluded from sampling. About one third of the animals had to be males, to permit *Y chromosome* investigation. A total of 3401 animals has been collected, belonging to 57 sheep and 47 goats populations (see complete lists of breeds in appendices 5 and 6). Among these are appearing 52 sheep and 43 goats autochthonous breeds. In addition, cosmopolitan Merino sheep and Alpine goats were double sampled in their site of origin (Spain and Switzerland respectively) and sampled in multiple other locations (Spain, Germany, Hungary, Poland and Romania for Merino sheep, and Switzerland, Germany, Italy and France for Alpine goat). A few difficulties were encountered because of blue tongue epidemics in Southern Italy, and Foot and Mouth disease in UK.

1. Raymond Laurans, 1989, «Le concept de race : approche ethnozootechnique, approche biologique», proceedings of the colloquium «La gestion des ressources génétiques des espèces animales domestiques», Paris, Bureau des ressources génétiques.
2. It is to be noted that despite the mentioned genetic continuity, publications show that it is possible to identify breeds thanks to the use of molecular markers in cattle (Blott *et al.*, 1999, and references therein).

Farms

Each selected breed is raised in a farm whose longitude and latitude had to be captured in the WGS84 *geodetic reference system* (decimal degrees) by partners in the respective concerned countries. The latter either used a *GPS* device with a dedicated protocol, or derived the location from local maps. Moreover, the altitude of the farms was also measured on the field with either an altimeter, or a GPS device. All data were validated with the 30 *arc second* digital elevation model (DEM) of the Shuttle Radar Topography Mission (SRTM30[1], NASA). As for the 3 arc second model[2], it was used to determine altitude when it had not been recorded on the field. The use of both *DEMs* turned out to be efficient in respectively detecting anomalies (farms located at an altitude of 6200 meters...) and assigning coherent altitudes in case of missing data.

The 885 farms where sampling has been carried out are constituting the basic geographic units for environmental data acquisition. Their attributes consist in climatic and topographical data, and in the aggregation of 1 to 3 animal molecular data, depending on the number of effectively sampled animals[3].
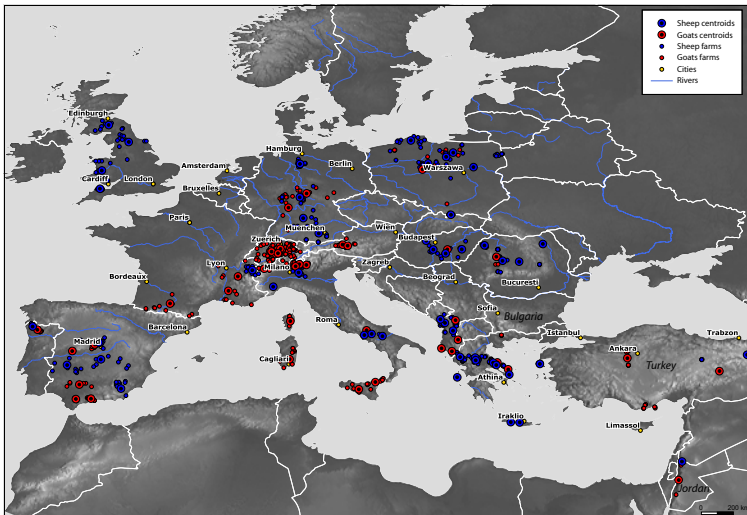
Fig 4.2. Spatial distribution of sheep and goat farms and breed centroids across EU, CEEC, North African and Middle Eastern countries. Sources : Econogene Consortium, SRTM90 - NASA 2000, Eurostat 2002.

Animals and farms data were locally collected by partners on *ad hoc* paper questionnaires, and possibly completed in laboratories. The information was then collected and centralized by the mean of an online database located on a project

---

1. SRTM30 succeeded to GTOPO30.
2. The resolution of SRTM3 is about 90 meters, and the resolution of SRTM30 about 1000 meters at the equator.
3. The corresponding protocol stipulated 3 animals, but in some cases more than 3 were sampled.

server. The structure of the web interface and the programmed rules gave concrete expression to the standardization recommendations (see page 35).

Breed centroids

As breeds are constituting the main elements to be analyzed, and as these breeds are scattered among several farms, it was necessary to create a virtual geographic entity to allow breeds data aggregation and to represent this information on geographic maps. This virtual geographic object has been defined as the centroid of the scatter of points representing all farms where the corresponding breed is raised. All breeds will thus be represented according to a geometric definition (see figure 4.3)
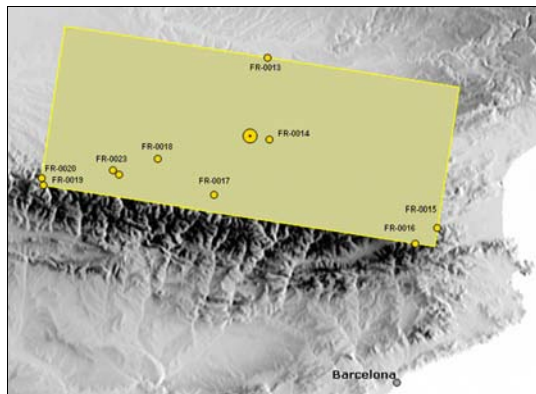
Fig 4.3. Construction of breeds centroids. In this example, all farms where goats of the Pyrenean breed are raised have been selected. The centroid of the common enclosing rectangle of the farms is calculated. It will represent the location of the Pyrenean breed on geographic maps. Sources : Econogene Consortium and SRTM30 - NASA 2000.

This location is showing a mean geographic location of breeds which is associated with the Econogene sampling. It is to be well distinguished from breeds historical location of origin which can be found in livestock genetic databases[1] for example. This historical location of origin was only used to clearly represent goat breeds in Switzerland. Indeed the spatial distribution of breeds is totally mixed up in this country, showing no geographic homogeneity : having recourse to the location of origin allowed to clearly attribute a region to sampled breeds, to avoid any confusion.

Organization levels in Econogene data

There are three kinds of geographical entities in Econogene data which are animals, farms and breeds. The animal is the basic and unique entity for acquiring genetic information. As we consider domestic species, the geographic reference

1. For instance the EAAP-Animal Genetic Data Bank : http://www.tiho-hannover.de/einricht/zucht/eaap/ (European Association for Animal Production) (08.11.2005)

for animals is a farm and not an individual location. Molecular data attached to animals are thus aggregated to the farm level. Then, farms data are aggregated at the breed level (see figure 4.4).
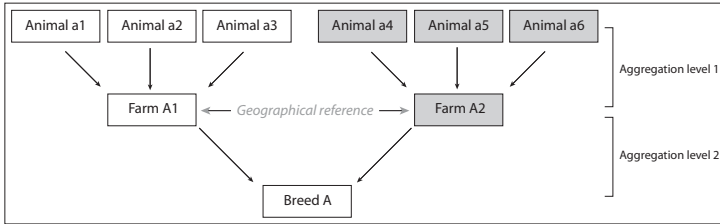
Fig 4.4. Levels of organization and corresponding data aggregation in the Econogene project.

The farm level was not employed in the present study[1] as the main object of investigation is the breed. In chapter 5 and 6, all analyses have been carried out at the breed level and all representations show the whole Econogene area extent. In chapter 7, the analysis is led at the animal level[2].

Environmental data

Environmental information was necessary in order to allow the characterization of the natural environment of farms and breeds sampling areas. Topographic data were described in the farm section. As for climatic data, it consists of latitude/longitude grids with a resolution of 10 minutes[3] containing 9 monthly variables over global land areas shown in table 4.1. This climatology was produced by the Climatic Research Unit of Norwich (CRU)[4] and takes into account the period 1961 to 1990. The way those climatic data were produced is described in details in New *et al.* (2002).

Original CRU data consist of one text file per variable. These files contain geographical coordinates (vector points as grid nodes) constituting the 10 minutes grid, and the attributes. This information was transferred to breeds centroids in accordance with the spatial coincidence concept (Goodchild, 1996) by :

1. generating **Voronoi polygons** to create continuous climatic grids (adjacent squares);
2. overlaying farms to the climatic continuous grid;
3. carrying out a spatial overlay to characterize farms' surrounding environment (see figure 4.5).

---

1. There is one exception with two maps representing an index of sustainable activity for the farms (see figure 6.13 on page 102 and figure 6.14 on page 103).
2. A unique scale was used to carry out the whole study, corresponding to the farm level. This is the geographical entity to which genetic data is made available, and also the geographical entity characterized by topographic and climatic data. The term «level» used in this section points out levels of organization of data and not scalar relationships (Allen, 1998).
3. This approximately corresponds to a resolution of 12 kilometers at the latitude of Switzerland.
4. http://www.cru.uea.ac.uk/ (10.11.2005)

.

Variable	Description
altitude	altitudes recorded on the field and completed with SRTM3 when required
wnd	monthly values of windspeed in m/s, 10 meters above the ground
dtr	monthly values of mean diurnal temperature range in deg C
frs	monthly values of number of days with ground-frost
pr	monthly values of precipitations in mm/month
prcv	monthly values of the coefficient of variation of monthly precipitation in percent
tmp	monthly values of mean temperature in deg C
rdo	monthly values of wet-days (number of days with >0.1mm rain per month)
reh	monthly values of relative humidity in percent
sun	monthly values of percent of maximum possible sunshine (percent of day length)

Table 4.1. Name of the topo-climatic variables, their description and abbreviation. Altitude was either recorded on the field with GPS devices or calculated with the SRTM3. All other variables have been calculated by the Climate Research Unit of Norwich (CRU).

It is to be noted that those climatic data are used to discriminate regions between them and to provide general climatic indicators through the overall sampling area. This is important because the sampling of genetic information was made after the 1961 to 1990 period, and as climate continuously evolves, these data are not necessarily expressing the present climate. In spite of this gap between the period described by climatic data and the moment when sampling was carried out, the analysis of the relationship between animals and the environment they are living in (exploited in chapter 7) remains relevant. First because the interval characterized by climatic data is long and therefore smoothed enough to keep significance, and then because the surrounding environment of farms is also smoothed given the resolution of the grids (about 12 kilometers on figure 4.5) : we are properly working on indicators.
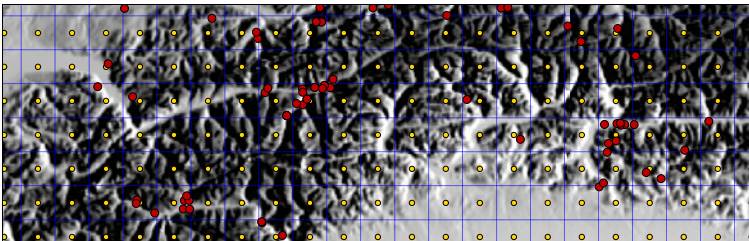
Fig 4.5. Generation of a climatic continuous grid by mean of Voronoi polygons. The farms are represented in red. Yellow points are coordinates provided by the CRU (grid nodes) to which are attached climatic data. Squares in blue are Voronoi polygons forming a continuous grid and making it possible the transfer of climatic information to the farms layer. Sources : Econogene Consortium, 2005; CRU, 2002; and SRTM30-NASA, 2000.

With altitude, a total of 118 topo-climatic variables[1] were made available for analysis. Monthly values have to be considered because there exist several management and production systems based on lambing (sheep) or kidding (goats) periods (seasons), for example fall lamb production, winter lambing, or spring lamb production in sheep. The influence of monthly climatic conditions are well illustrated by the following description provided by the College of Agriculture and Home Economics of the New Mexico University : «The ovulation rate in sheep is

---

1. The yearly mean of each variable was also calculated.

normally at its peak in late september through november. Temperatures at this time are typically not high enough to decrease ram fertility or to cause embryo loss. Normally, spring temperatures are mild and death loss associated with weather conditions is minimal. But newborn lambs must be offered some protection from spring winds. In this type of system, ewes are bred when ovulation rates should be high, so that flushing, teasing, or control of environmental conditions has less effect on conception rate or length of lambing season»[1].

Annual means of those different variables have also been computed in order to provide an indicative environmental profile of the breeds. These informations are appearing in appendix 1.

Molecular data sets

The description of the geoenvironmental components being completed, we will move straight onto genetic information. In addition to the fact that the Econogene project is among the first to yield complementary data on population and evolutionary genetics, on animal husbandry practices, and also including socio-economics over a large geographic extent (Bruford, 2005), a very original aspect is that it is based on the simultaneous investigation of five different kinds of molecular markers. This abundance of means to measure genetic diversity generates different information sources whose cross-checking is likely to make researchers gain insight into molecular markers respective contribution, and consequently of course into the way sheep and goats genetic resources should be optimally managed.

The concerted efforts of the Econogene project in sheep and goat diversity allowed for the first time to compare *microsatellites*, AFLP (Amplified Fragment Length Polymorphisms), genomic SNPs (Single *Nucleotide* Polymorphisms), Y chromosome SNPs, and *mitochondrial DNA* (Lenstra, 2005). Data for all of them were made available as well on the farm as on the breed level and have been used for exploratory data analysis (chapter 5) and molecular data representation (chapter 6).

DNA extraction was carried out in local laboratories to avoid problems of biological sample transportation, shipping and related sanitary or conservation issues. Then all DNA was centralized in a collecting and distribution site. This laboratory received 3165 samples, aliquoted them in batches of 2 µg of DNA each and distributed them to the different partners in charge of molecular analyses. These laboratories produced excel sheets data sets which were then uploaded on a project server and made available to all project's partners. Each DNA sample was labeled with a unique identifier allowing the animal information to be linked to its origin farm and to the related geographic coordinates.

The different types of genetic markers which have been produced may differ with respect to important features, such as their abundance in the **genome**, their level of detected polymorphism, **locus** specificity (see next section for *polymorphism* and *locus*), reproducibility, or technical requirements (Karp *et al.*, 1996). That is the rea-

---

1. http://www.cahe.nmsu.edu/pubs/_b/100B15.html (09.11.2005)

son why several types of molecular markers were used within the Econogene project, their own respective particularities allowing the investigation of different scientific goals. However, none of the available techniques is superior to others. The key question rather being to know which marker to use in which situation. But the choice of the most appropriate *genetic marker* may also depend on financial limitations (according to the type of equipment required) or time constraints. This last element conditioned the use of *microsatellites* and AFLPs only, in chapter 7, when looking for markers possibly under natural selection, despite the subsequent availability of other kinds of molecular markers. Both are neutral[1] and their nature and structure will be detailed in the following section, accompanied by a few essential molecular biology notions. So as to make it plain, the simplified technical elements appearing in the next section are intended to any potential GI scientist involved in landscape genetics. For a rigorous description of the implied processes and techniques, please refer to the mentioned literature.

MOLECULAR MARKERS

Markers are DNA fragments acting as reference points to follow the transmission of chromosome segments from one generation to another. They can be used to look for genes involved in animals zootechnic interesting characters (parasite resistance for example), but also to survey and maintain genetic diversity in threatened populations.

Before entering into the details about two kinds of molecular markers, here used in conjunction with the attentive study of their geographical attributes, there is still a need to give a few explanations about important concepts and techniques at the roof of most of molecular markers discovery.

DNA

Genetic information is stored in cells as long deoxyribonucleic acid (DNA) molecules, constituting the genome. DNA is made of a common structure constituted by a 5 atoms carbon sugar (pentose), and a phosphate. Nucleotides[2] are attached to the pentose-phosphate backbone. Four types of nucleotides are existing and often are designated by their initial as abbreviation : A(denine), C(ytosine), G(uanine) and T(hymine). Nucleotide's sequences on both DNA strands are complementary (C bases G, and T with A, see figure 4.6). DNA is organized in chromosomes. For humans, together with animal domestic species, genetic information is contained in 2n chromosomes distributed among n pairs of homologous chromosomes, each chromosomes pair being composed of one coming from the father, and one from the mother.

---

1. See reflections about the neutral theory proposed by Kimura (1968, 1991) in chapter 7. It asserts that most of evolutionary changes at the molecular level are not caused by Darwinian selection.
2. Nucleotides are subunits of DNA. Each nucleotide is divided into three parts : 1. a nitrogenous base (A, C, G, or T); 2. a phosphate molecule; 3. a sugar molecule (deoxyribose).

On the basis of the way information is stored in DNA, two important notions are existing to characterize it : *polymorphism* and *linkage* which are involved in the use of molecular markers.
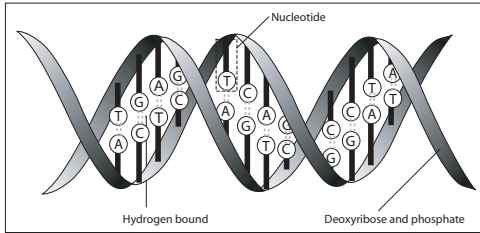
Fig 4.6. Representation of a DNA molecule. Each strand is constituted of sequences of the four A, C, G and T nucleotides. A gene is only a portion of the DNA molecule. Genes do correspond with the coding part of the genome which is subsequently translated in proteins by the mean of amino acids assembling. But this only corresponds to 5 to 10% of the whole genome. Sources : DNA strands adapted from M.-H. Farce, INRA, 1998.

Polymorphism

Considering a gene or not, the analysis of a sequence of nucleotides at a given location on the genome (a locus, figure 4.7), and in a given population, can show changes. Alternative nucleotides sequences at a given locus are called *alleles* and the existence of several possible alleles at a given locus is defining genetic polymorphism (figure 4.8). An individual with two identical alleles at a given locus is defined as *homozygote*, and an individual with two different alleles at a given locus is *heterozygote*. This polymorphism is allowing to establish the parental origin of an allele at a given locus, as it is possible to distinguish chromosomes received from the father and from the mother.
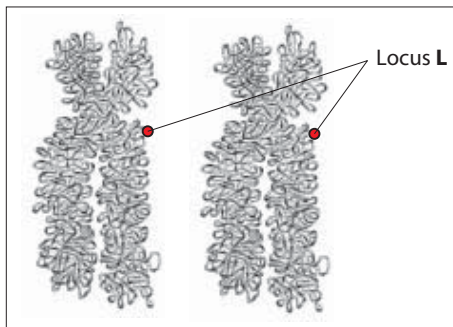
Fig 4.7. A pair of homologous chromosomes. DNA is organised in chromosomes, all chromosomes of a considered organism constituting its genome. A locus is a precise location on a given chromosome. Sources : chromosomes extracted from M.-H. Farce, INRA, 1998.

An allele is likely to play a marker role only if it can be distinguished from other alleles. Moreover, within a population, a marker is likely to be useful only if the breeder is heterozygote at the location of this marker. This of course because for a homozygote breeder, the marker provides no information to distinguish two

types of descendants. And even in the case for which the father, the mother and the offspring are heterozygote (A/a), the marker is not providing information.

The efficiency of a marker is assessed according to its unambiguous ability to distinguish two descendants groups according to a marker allele. A *codominant* marker is a marker for which all alleles can be merely deducted from the observation of the **phenotype**. It is providing more information than a *dominant* marker whose recessive allele can be observed only when homozygote. A marker is providing the more information when the number of alleles is high and their frequencies are balanced. This is why highly polymorphic codominant markers are checked for. A system to increase the information provided is to consider a group of narrowly bound markers as a unique marker called **haplotype**, and whose polymorphism is the result of the allelic combination of each basic marker (Crow, 1986; Suzuki, 1991).
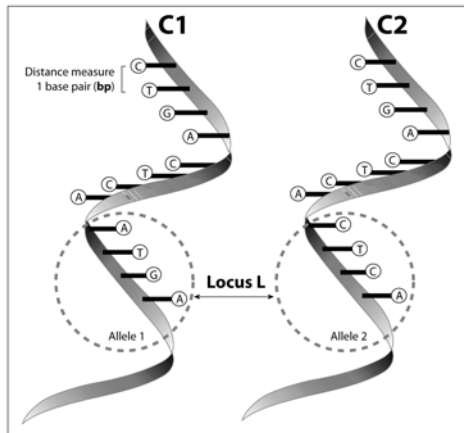
Fig 4.8. Representation of two homologous chromosomes, or two identical chromosomes of two individuals. For each chromosome (C1 and C2), only one brand of DNA is figured. At locus L, the sequence of nucleotides is varying. Sources : DNA strands adapted from M.-H. Farce, INRA, 1998.

Linkage

To constitute a **gamete**'s genome during **meiosis**, one chromosome is randomly extracted from each homologous pair of the parents. Therefore, two genes located on chromosomes of distinct pairs are independently transmitted. When located on the same chromosome, two genes are generally transmitted both at once. This rule is not absolute and we have to take into account that material can be exchanged between homologous chromosomes. In this *crossing-over* case, two genes initially located on the same chromosome are likely to find themselves back on both homologous chromosomes (figure 4.9), and *recombination* occurs. Recombination frequency is a function of the distance between loci. Loci close from each other are not recombining very much and tend to be transmitted both at once.

Genetic linkage may allow to generalize observations made on a given locus to the whole DNA segment surrounding this locus. In case there is no recombination

between the observed locus and the surrounding segment, this locus becomes a marker of this segment and of all genes present in it. The efficiency of a marker is greater when the considered DNA segment is shorter as it is limiting the recombination rate between this marker and a gene. Generally, the distance between a marker and a gene cannot be greater than 20 *centimorgans* (cM)[1] (Crow, 1986; Suzuki, 1991).
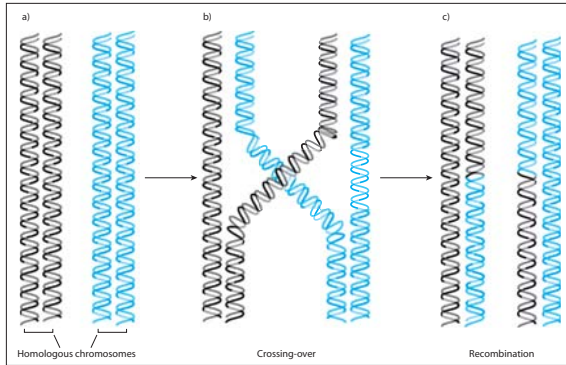
Fig 4.9. A molecular recombination event represented by two DNA double helices, breaking under polarity constraints (Holliday model). Free extremities then combine with the complementary homologous helix. Sources : Suzuki *et al.*, 1991.

Linkage (des)equilibrium

Within all *gametes* of a considered population, there is linkage equilibrium between two loci when for each X and each Y allele at this locus, the frequency of the XY *haplotype* equals the product of X and Y alleles frequencies. In this case, it is not possible to predict the allele at the second locus when knowing the allele at the first one. On the contrary, there is linkage desequilibrium when preferential associations are existing between alleles at both loci. Linkage desequilibrium is often due to the fact that a genetic link is existing, but the reverse is not true and the existence of a genetic link doesn't imply linkage desequilibrium. Linkage equilibrium is generally admitted as working hypothesis when considering a large closed population. Indeed, linkage desequilibrium generally occurs through selection, migration, *mutation*, or *genetic drift*, and is gradually replaced by successive recombinations in the course of generations. Consequently, each global linkage desequilibrium within a population is not stable and is existing only in the case of recent evolutionary processes (selection, *mutation*, migration, drift) or if loci are physically very close to one another. In this case, markers efficiency would be weak as the association between two alleles at two loci detected on a population's sample could not be generalized on the level of the whole population.

----

1. One *centimorgan* equals about 1 million base pairs. The size of the a whole animal genome is about 3'000 cM.

Polymerase Chain Reaction

The polymerase chain reaction (PCR) is a technique used to amplify small segments of DNA. This molecular biology method was developed in 1985 by Kary Mullis (chemistry Nobel prize in 1993). Small single-stranded segments of DNA made of 20-30 nucleotide bases (oligonucleotides[1]) are synthesized in vitro in order to be correctly bound to opposite strands of the DNA segment it is wished to replicate. At the points of contact an added *enzyme* (DNA polymerase[2]) can start to read off the nucleotide sequence and, through bases complementarity, synthesizes a new sequence until two new double strands of DNA are formed. The sample is then heated, which makes the strands separate so that they can be read off again. The procedure is continuously repeated, doubling at each step the number of copies of the desired DNA segment[3]. Through such repetitive cycles, it is possible to obtain millions of copies of the desired DNA segment within a few hours (see appendix 2). According to the usual approach, nucleotides provided to start the reaction are radioactive to make it possible to distinguish the different alleles by autoradiography[4] after electrophoresis[5] (figure 4.7). Since a few years, radioactivity is progressively replaced by fluorescent labelling.

The PCR technique is presently used in numerous molecular genetics applications and it is notably essential to reveal *microsatellites* and AFLP polymorphisms.

Microsatellite markers

Microsatellites are stretches of DNA that consist of tandem repeats of sequences of *mono*, *di* or *tri* nucleotides which are repeated between 10 and 20 times (for example, the frequent TG di nucleotide repeated 15 times in succession) and have no known coding function. These sequences are numerous, regularly distributed over the genome and characterized by an important polymorphism due to the variation of the number of repeats from an allele to the other. Using *PCR*, these

---

1. A molecule usually composed of 25 or fewer nucleotides used as a DNA primer. A primer is a nucleic acid strand (or related molecule) that serves as a starting point for DNA replication. A primer is required because most DNA polymerases (*enzyme*s that catalyze the replication of DNA) cannot begin synthesizing a new DNA strand from scratch, but can only add to an existing strand of nucleotides.

2. DNA polymerase is due to thermophil bacteria resisting to very high temperatures. For instance Thermus aquaticus (Taq polymerase).

3. «One Friday night I was driving, as was my custom, from Berkeley up to Mendocino where I had a cabin far away from everything off in the woods. My girlfriend, Jennifer Barnett, was asleep. I was thinking. Since oligonucleotides were not that hard to make anymore, wouldn't it be simple enough to put two of them into the reaction instead of only one such that one of them would bind to the upper strand and the other to the lower strand with their three prime ends adjacent to the opposing bases of the base pair in question (...) But what if the oligonucleotides in the original extension reaction had been extended so far they could now hybridize to unextended oligonucleotides of the opposite polarity in this second round ? The sequence which they had been extended into would permit that. What would happen? EUREKA!!!! The result would be exactly the same only the signal strength would be doubled. EUREKA again!!!! I could do it intentionally, adding my own deoxynucleoside triphosphates, which were quite soluble in water and legal in California. And again, EUREKA!!!! I could do it over and over again. Every time I did it I would double the signal. For those of you who got lost, we're back! I stopped the car at mile marker 46,7 on Highway 128. In the glove compartment I found some paper and a pen. I confirmed that two to the tenth power was about a thousand and that two to the twentieth power was about a million, and that two to the thirtieth power was around a billion, close to the number of base pairs in the human genome (...)». Taken from the Nobel Lecture Kary Mullis gave in december 1993 in Stockholm. (http://nobelprize.org/chemistry/laureates/1993/mullis-lecture.html) (10.11.2005)

4. A technique using X-ray film to visualize molecules or fragments of molecules that have been radioactively labeled.

5. Fragments of DNA are placed in a semi porous gel, and an electrical field is turned on. The fragments move in response to the field, with smaller fragments generally moving faster. After a time, the fragments have separated enough to form a series of separated lines like a bar code that characterizes the DNA.

repeats can be easily amplified. The number of repeat units that an individual has at a given locus can be resolved using a polyacrylamide[1] gel whose high resolution permits a distinction of alleles whose size is one base pair different. From the gels, it is generally possible to perceive two genetic marks (alleles) for individuals as each one is inheriting one length of nucleotide repeats from his mother and one from his father and are thus considered co-dominant. Individuals with only one band have in fact received the same allele from both their mother and father. A *sine qua non* condition to use microsatellites in an efficient way is to make sure that the considered locus is unique. To check for it, flanking sequences on both sides of the locus have to be the same.
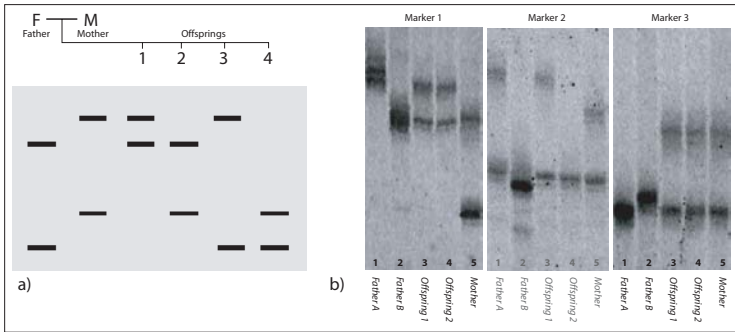
Fig 4.10. In part a), each bar represents one allele of the microsatellite for which there is a particular number of repeats. On this figure, each individual has two alleles (only one bar would be visible in case of homozygote individuals). Each offspring has inherited one allele from its mother and one from its father. In part b), autoradiography reveals 3 panels (3 markers) of DNA samples constituted of 5 lanes. Lane 1 and 2 show alleles present in two possible fathers, lanes 3 and 4 show alleles for offsprings, and lane 5 shows alleles for the mother. The 3 markers are pointing out that lane 1 correspond to the «real» father. Sources : adaptation of http://www.liv.ac.uk/~kempsj/fingerprint.html (09.11.2005)

Microsatellites are highly variable. In a population, many alleles of a single microsatellite locus, different in the number of repeats, may exist (up to 70 at a single locus). Moreover, microsatellite alleles change rapidly over time (Smith and Gaffney, 2000), evolving over time, from generation to generation. That is a reason why they are used to detect recent changes in population like effects of population fragmentation. Microsatellites are also useful for the identification of incipient differentiation of populations.

Different loci have been selected to be investigated in the context of the Econogene project. Their selection was made with the aim of covering most chromosomes, maximizing the overlap between the Econogene list and markers employed in other large scale projects[2], optimizing the experimental effort, and finally satisfying partners preferences and experience. It is to note that the Econogene list of microsatellite markers has become the FAO recommended list for the investigation of biodiversity in sheep and goats.

---

1. A polymeric thickener.
2. Biotech project on sheep and goat diversity; the Nordic, Baltic and ILRI sheep diversity projects.

AFLP markers

Amplified Fragment Length Polymorphism (AFLP) is a highly sensitive method for detecting polymorphisms in DNA which is allowing the detection of polymorphisms of genomic restriction fragments by **PCR** amplification.

Firstly, the analyzed genomic DNA is digested by restriction **enzyme**s (endonuclease) which is a class of bacterial enzymes that cut DNA at specific sites. Then, adapters fragments of known sequence are specifically added to the ends of the fragments generated by the used restriction enzymes (see Mueller and Wolfenbarger, 1999; see appendix 3). Those fragments are amplified by PCR, using an oligonucleotide as a primer which is complementary to the sequence of the adapters. This sequence is extended into the unknown part of the fragments - usually one to three arbitrarily chosen nucleotides beyond the restriction site. To achieve a selective amplification of a subset of these fragments, only the ones with complementary bases of the arbitrary nucleotides are amplified, the others are not and will not appear during next stage where the amplified fragments are separated on a sequencing gel and visualized by autoradiography or fluorescent sequencing (figure 4.11).
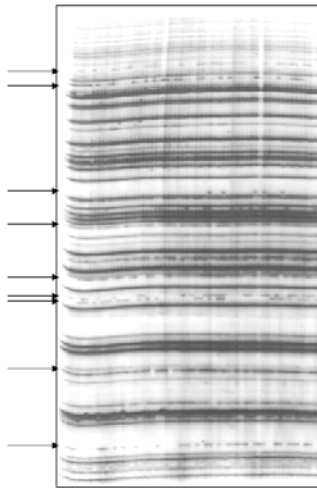
Fig 4.11. After electrophoresis and autoradiography, AFLP EcoRI/TaqI combination profiles in goat. The arrows indicate AFLP polymorphisms. Source : Istituto di Zootecnica, Facoltà di Agraria, Catholic University of Sacro Cuore, Piacenza.

The enzyme-adapter combination is able to reveal polymorphism (for instance **mutation** in the restriction site, mutation in the adapter flanking region, or large insertion between adapter sites) between individuals and constitutes an AFLP marker. The identification of a marker is depending on the sequence of the location of the restriction enzyme and on the arbitrary bases. An important number of enzyme-primer combinations are existing as about ten restriction enzymes are routinely used, and as there are numerous amplification primers to be combined with three arbitrary bases.

	International code	Selective nucleotides
Goat	E32/T38	AAC/ACT
	E43/T33	ATA/AAG
	E45/T32	ATG/AAC
Sheep	E35/T38	ACA/ACT
	E35/T32	ACA/AAC
	E45/T38	ATG/ACT

Table 4.2. Most informative AFLP primer pairs combinations identified for goats and sheep in the Econogene project.

In the framework of Econogene, AFLP primer pairs were screened in goats and sheep to identify the most informative combinations to use AFLP analysis on a large scale (table 4.2). Following the testing of 64 primer pairs in goats and 48 primer pairs in sheep, three primer pairs per species were chosen and produced 104 AFLP markers in goats and 98 in sheep.

However, and to conclude, let's mention a disadvantage of the AFLP method which generates dominant rather than codominant markers. It means that the markers are scored as present or absent and thus do not allow the identification of homologous alleles. This renders the method less easy to use for studies that require precise assignment of allelic states, such as heterozygosity analysis. However, because of the rapidity and ease with which reliable markers can be generated, AFLPs are emerging as a powerful tool among other molecular markers for studying genetic diversity (Mueller and Wolfenbarger, 1999; Ajmone-Marsan, *et al.*, 1997).

FROM GENOME TO GEOGRAPHY : DATA SETS ELABORATION

Known microsatellite loci are documented into public genetic information databases (for instance the NCBI[1]) what allows researchers to localize them and to carry out analyses. After amplification, a DNA sequencer and an associated software provide the size of the microsatellite markers at the different loci. The latter are stored in the columns of a raw data set containing a list of all animals (stored in rows) of all breeds whose DNA has been analyzed (figure 4.12).

	A	B	C	D	E	F	G	H
1	ID	BREED	BM1329	BM8125	DYMS1	HUJ616	OARAE129	OARCP34
2	OAALBAR07	ALBAR	164 166	120 120	175 183	126 162	146 148	120 126
3	OAALBAR08	ALBAR	162 168	120 122	175 183	122 156	136 148	126 128
4	OAALBAR09	ALBAR	162 180	120 120	185 195	126 162	136 148	120 122
5	OAALBAR10	ALBAR	164 166	120 124	187 197	126 140	136 136	120 122
6	OAALBAR11	ALBAR	164 166	120 122	185 197	114 122	146 146	122 126
7	OAALBAR12	ALBAR	162 164	122 122	181 181	126 156	148 148	120 126
8	OAALBAR13	ALBAR	164 168	122 122	181 197	122 162	146 146	118 120
9	OAALBAR14	ALBAR	162 180	118 120	175 195	122 124	148 148	120 122
10	OAALBAR15	ALBAR	164 166	118 120	175 197	122 126	136 148	126 130

Fig 4.12. A raw sheep microsatellite data set with the ID of the animal, the abbreviation of the breed and then one locus per column. A column is containing two numbers which are the lengths of the microsatellites in base pairs for both alleles.

For each locus, the length of the microsatellites (expressed in number of base pairs) is displayed for both alleles. An animal is homozygote when both numbers are equals.

---

1. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/ (09.11.2005)

This raw information is then re-encoded in order to transform it into frequencies. In fact, both alleles of a given locus are distributed among two columns. All existing microsatellite lengths of all loci are tested for all animals. A «1» is coded in the column when a microsatellite length is existing for a given animal, and a «0» if not (figure 4.13). This permits then to obtain presence frequencies of the loci to calculate diversity indexes[1] (see the genetic diversity section), and provides a qualitative information about those loci to be tested versus geoenvironmental variables (see chapter 7).

Geographic coordinates are added simply by joining the genetic table with the sampling table on the basis of the animal ID.

Fig 4.13. Selected loci with presence or absence of a microsatellite length by allele in a sheep georeferenced table.

For AFLP, the elaboration of the data set is rapidly done as this technology is producing binary information, a (dominant) marker being present or absent at a given locus (figure 4.14).

Fig 4.14. A table with georeferenced AFLP markers in sheep, with the ID of the animal, the corresponding farm where it was sampled, the breed, geographic coordinates, the species, the country of origin, and then primer pairs combinations (markers) with a 1 when present and a 0 when absent for each animal.

This is how the - henceforth geographic - genetic information looks like before being processed by GIS tools, as well in exploratory spatial analysis (chapter 5), as in cartography (chapter 6) and also in the case of the detection of natural selection signatures[2] (chapter 7).

Before going to analysis, it is necessary to provide a few explanations about population genetics concepts, and more specifically about the different genetic diversity variables or indices which are calculated on the basis of raw genetic information, and which will be used in the forthcoming two chapters.

---

1. Various specialized statistical genetics software do process this type of data set automatically. See http://www.biology.lsu.edu/general/links.html (09.11.2005)
2. «Signature» here means a distinctive mark or characteristic.

POPULATION GENETICS AND GENETIC DIVERSITY

Our interest will mainly focus on genetic diversity because it reveals variations at the level of individual genomes, and divulges a mechanism for populations to adapt to their ever-changing environment. The more variation, the better the chance that at least some of the individuals will have an allelic variant that is suited for a new environment, and will produce offspring with the variant that will in turn reproduce and continue the population into subsequent generations. Having recourse to the «gene» concept, we can consider that the gene reservoir of a population is a complete set of unique alleles that would be found when investigating the genetic material of each of its members. A large gene pool indicates a large genetic diversity, which is associated with a robust population able to survive intense selection. Meanwhile, low genetic diversity can cause reduced fitness and increased chances of extinction.

Several genetic diversity variables have been calculated on the basis of microsatellites and AFLP data described here above. Those ones are providing a way to assess a level of diversity and thus determine a vulnerability status in a genetic resources conservation perspective, or a general indicator of the genetic health of a population.

Heterozygosity

An important notion is the one of heterozygosity which is a measure of the genetic variation in a population, with respect to the fraction of individuals in that population which are heterozygote for one or several given loci.

The *observed heterozygosity* of a population is measured by determining the proportion of loci that are heterozygote and the number of individuals that are heterozygote for each particular locus. For a single locus with two alleles, H_o is the number of heterozygotes at this locus divided by the total number of surveyed individuals. Over a series of several loci, H_O is the sum of H_o heterozygotes calculated for each locus divided by the number of considered loci.

Expected heterozygosity (H_E) is the probability that two alleles drawn at random are different alleles. H_E differs from the H_O as it is a prediction based on the known allele frequency from a sample of individuals, if the population mates at random. Deviation of H_O from H_E is used as an indicator in population dynamics in accordance with Hardy-Weinberg principle (see Hardy-Weinberg law on page 53).

The way to calculate it for a single locus is :

$$H_E = 1 - \sum_{i=1}^{k} p_i^2$$

Where p_i is the frequency of the ith of k alleles.

To get the expected heterozygosity over several loci, a double summation is required :

$$H_E = 1 - \frac{1}{m} \sum_{q=1}^{m} \sum_{i=1}^{k} p_i^2$$

(EQ 4.2)

Where the first summation stands for the qth of m loci. The average over the m loci is made via the 1/m term. The second summation is equal to EQ 4.1.

Hardy-Weinberg law is an important theory in investigating genetic varieties in an idealized[1] large population stipulating that alleles frequencies remain constant from generation to generation, if mating is at random and if there is no selection, neither migration, nor mutation (equilibrium situation). In the case of a single locus with two alleles *A* and *a* with allele frequencies of *p* and *q* respectively, the Hardy-Weinberg principle predicts that the genotypic frequencies for the *AA* homozygote to be p^2, the *Aa* heterozygote to be *2pq* and the *aa* homozygote to be q^2.

F-statistics

F-statistics are measures of genetic structure developed in the 1920s by Sewall Wright (University of Chicago), one of the primary founders of population genetics, related to statistical analysis of variance.

Within a subpopulation, the F (without subscript) is the ratio of [the difference between expected and observed heterozygosity] to [the expected heterozygosity].

$$F = \frac{H_E - H_O}{H_E}$$

(EQ 4.3)

F_{ST} is the proportion of the total genetic variance contained in a S subpopulation relatively to the T total genetic variance of the whole population. Values can range from 0 to 1, and a high F_{ST} implies an important differentiation among populations.

F_{IS}, used for the microsatellite data set in our case, is an *inbreeding coefficient* which assesses global variation in I individuals, relative to the variation in their S subpopulation. If the F_{IS} is positive then the set of subpopulations, as a whole, is inbred (deficiency of heterozygotes). If the F_{IS} is negative, then the set of subpopulations, as a whole, is outbred (has an excess of heterozygotes).

---

1. The idealized population is infinite (so as to eliminate *genetic drift*), sexually reproducing and diploid.

Jaccard similarity index

The Jaccard index (1908) is a measure of similarity between individual ***genotypes*** which is applied to AFLP in our case.

$$S_{ij} = \frac{a}{(a + b + c)}$$

(EQ 4.4)

Where *a* is the number of genotypes present in both *i* and *j*, *b* is the number of genotypes present in *i* but not in *j*, *c* the number of genotypes present in *j* but not in *i*.

Other variables

Finally, and briefly, the last three genetic variables utilized in this research are 1) the frequency of the recessive genotype which is the frequency at which a homozygote recessive genotype occurs when a particular locus comprises two alleles for the recessive trait (*aa* for instance), 2) the mean number of alleles and 3) the number of polymorphic loci which are two clearly explicit denominations.

It is therefore possible to plunge into data analysis...

EXPLORING THE SPATIAL DIMENSION OF MOLECULAR DATA

Chapter outline

Considering the vast quantity of information collected within the Econogene project, exploratory data analysis methods are likely to help to extract useful and so far unknown information from large spatially explicit genetic data sets. A specific category of GIS tools may facilitate investigations to understand the geographic distribution of genetic diversity among sheep and goat breeds as well as its variation according to environmental parameters.

EXTRACTING INFORMATION FROM LARGE GENETIC DATABASES

Genetic research projects generate enormous quantities of data. For instance, in order to provide an idea of the size, Genbank which is the US National Institutes of Health (NIH) molecular database[1], is composed of an annotated collection of all publicly available DNA sequences. In february 2004, it contained more than 44 billions of base pairs and approximately 40 millions of sequences[2].

Comparatively, Econogene teams have collected biological samples from over 3'400 animals distributed among 104 breeds, for which 5 types of molecular analyses were carried out, generating hundreds of markers to be investigated and compared with more than 100 environmental variables likely to make any interesting relationship emerge. In this research context, only one solid *a priori* work hypothesis arose about genetic diversity through the whole study area, being that animals reared at the location of centres of domestication or next to them are expected to contain more genetic diversity and to be very distinct from the ones living at the ends of migration routes (Cavalli-Sforza *et al.*, 1994; Hanotte and Jianlin, 2005). But the amount of available data is of course likely to promise much more interesting discoveries, and it was not worth deploying such noteworthy means just to confirm one postulate. So how to extract information from this important amount of data ? In the category of database like Genbank, computational **genomics** is essential. It is a field leaning on supercomputer technologies and dedicated to the development of large-scale genome research applications, notably built on processes of database mining (Houle *et al.*, 2000). More modest being the Econogene data set, an alternative way was found to investigate this nevertheless considerable amount of data by taking advantage of the available geographic location of samples in order to *explore* the variation of molecular variables through the whole

---

1. http://www.psc.edu/general/software/packages/genbank/genbank.html (10.11.2005)
2. http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html (10.11.2005)

project area, this in comparison with geoenvironmental data and making use of specialized GIS softwares[1].

EXPLORATORY DATA ANALYSIS

The Exploratory Data Analysis (EDA) field was first defined in the John Tukey Exploratory Data Analysis book (Tukey, 1977). In 1962, Tukey issued a call for the recognition of data analysis as a branch of statistics and distinct from mathematical statistics in a paper entitled «The Future of Data Analysis» (quoted in Friendly and Denis, 2005). He conceived a wide variety of new graphic displays under the designation of «Exploratory Data Analysis». This approach employs a variety of mostly graphical techniques in order to maximize insight into a data set, that is to uncover underlying structures, to extract important variables, to detect outliers and anomalies, etc. Instead of looking for a known model and checking if data is conform, EDA proposes a more direct approach of allowing data itself to reveal its underlying structure, stimulated by spontaneous successive rough hypotheses outlines produced by researchers. EDA is mainly resorting to graphical techniques for the reason that its main role is to «open-mindedly» explore data; visualization of graphics provides matchless power to do so, making it possible to discover structural hidden aspects, and to gain some new insight into the data. Scientific visualization was first used as an informal way to analyze information (Unwin cited in [Tobon, 2001]) until it was recognized as a scientific method by the end of the 1980s (MacEachren, 1994).

GEOGRAPHIC VISUALIZATION (GVIS)

On the basis of EDA, a complementary approach arose to exploit the spatial dimension of data, when available. Exploratory Spatial Data Analysis (ESDA) tools included additional methods elaborated to take into account the specificities of geographic information (MacEachren, 1994; MacEachren, 1995; Banos, 2001; Haining, 2003). Indeed, cartography gradually had to deal with an increasing number of data sources which were becoming larger and larger, and developments in GIS made it possible «to rejoin data storage with display» (MacEachren *et al.*, 2001), transforming traditional maps into real interfaces able to support «knowledge construction activities» (MacEachren *et al.*, 2001), while keeping their representation function. So emerged a «modern cartography» (MacEachren *et al.*, 2001) likely to face the changes occurring in geographic information management and analysis. *Geovisualization* (GVIS) is an approach stemmed from these developments and offering *dynamic* and *interactive* access to geodata, fitted to facilitate search for unknowns, information exploration and finally knowledge construction in the absence of pre-determined hypotheses.

In practical terms, GVIS tools are providing interactivity in the sense that they allow users to choose and visualize different variables to assess their simulta-

---

1. «Wandering» among large data sets according to Banos (2001).

neous variation, together with a constant access to the spatial location of the considered objects, in order to facilitate *visual thinking*. An interactive and dynamic link is established between the geographic representation of analyzed objects and the genetic information they contain. Compared with thematic cartography, GVIS softwares are more powerful to investigate and visualize data, as it is possible to find a value, to see the corresponding location on a map, and to get the values of all other variables describing the breed or the environment at this location. But representation features are not as extensive as those of cartography and this makes GVIS essential tools to be used *upstream* - that is before fixing an interesting situation with cartography (chapter 6) - during the first steps of the scientific reasoning.

COMBINED ANALYSIS OF GENETIC AND ECOLOGICAL DATA OF GOAT AND SHEEP BREEDS

In the following sections, exploratory analysis will be demonstrated by exploiting four different GVIS functions on molecular and geoenvironmental data[1] characterizing breeds. A *breed* is represented by 33 to 47 individuals. As for farms, genetic data sets are most often composed of three and sometimes up to seven or eight individuals, they were used solely for informal exploratory investigations; they could not be considered as statistically representative samples (Wonnacott and Wonnacott, 1995). During discussions punctuating the Econogene project, it has been argued that the multiplicity of molecular markers was likely to compensate the lack of individuals forming farm samples. Against this argument one can first argue that the present knowledge about the location of the different kinds of markers within the genome is not sufficient enough to pretend they can be complementary. Then each type of marker is considered as a separated variable what doesn't improve statistical representativeness; actually, different ways of measuring diversity don't increase the number of animals in a farm. Sampling was especially designed for analyses at the breed level, and information at the farm level (individuals) was only used when all individuals were globally taken into account to be compared to geoenvironmental data (chapter 7).

As preamble to spatial analysis, some preliminary remarks are necessary. First, it was only possible to make use of 41 goat and 50 sheep breeds among the respectively 47 and 57 originally sampled[2] (their geographic location is displayed in figures figure 5.1 and figure 5.2). Then it is essential to keep in mind that the number of considered breeds - our statistical individuals - is not that important and this should temper the different observations (Wonnacott and Wonnacott, 1995). Finally, let us mention that data have been standardized : with the exception of latitude, longitude and variables representing frequencies, each value has been divided by the maximum of the distribution it is belonging to, in order to make the variables comparable.

---

1. A codebook of the different molecular and geoenvironmental variables used in this chapter is proposed in appendix 4.
2. Six goat breeds and seven sheep breeds were not taken into account mainly because of missing data.

Chapter 5

EXPLORING THE SPATIAL DIMENSION OF MOLECULAR DATA
Combined analysis of genetic and ecological data of goat and sheep breeds

Fig 5.1. Spatial distribution of goat breeds through the Econogene study area. 1. Brava 2. Verata 3. Payoya 4. Florida 5. Malaguena 6. Cabra del Guadarrama 7. Pyrenean 8. Rove 9. French Alpine 10. Valdostana 11. St.Gallen Booted 12. Swiss Alpine 13. Valais Black Neck 14. Grisons Striped 15. Peacock Goat 16. German Alpine 17. Sarda 18. Corsican 19. Orobica 20. Camosciata delle Alpi 21. Bionda dell'Amadello 22. Thuringian Forest Goat 23. Pinzgauer 24. Tauernschecken 25. Girgentana 26. Grigia Molisana 27. Argentata dell'Etna 28. Polish Fawn Colored Goat 29. Dukati 30. Hungarian Native 31. Muzhake 32. Hasi 33. Capore 34. Carpathian 35. Skopelos 36. Greek Goat 37. Angora 38. Baladi 39. Hair 40. Abaza 41. Gurcu.
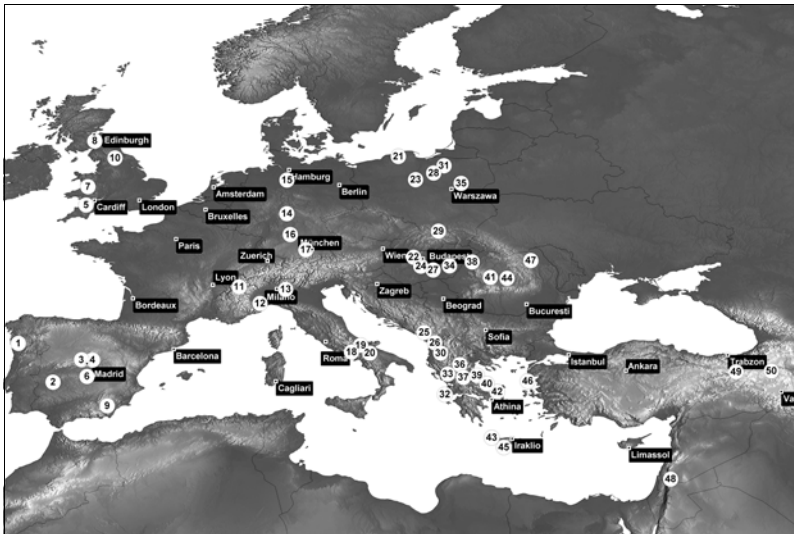
Fig 5.2. Spatial distribution of sheep breeds through the Econogene study area. 1. Churra Braganzana 2. Spanish Merino 3. Rubia del Molar 4. Colmenarena 5. Exmoor Horn 6. Manchega 7. Welsh Mountain 8. Scottish Blackface 9. Segurena 10. Swaledale 11. Thones et Marthod 12. delle Langhe 13. Bergamasca 14. Rhoensheep 15. German Grey Heath 16. German Merino 17. White/Brown Mountain 18. Laticauda 19. Gentile di Puglia 20. Altamurana 21. Pomeranian 22. Cikta 23. Polish Merino 24. Hungarian Tsigaia 25. Shkodrane 26. Bardhoka 27. Hungarian Merino 28. Polish Heat 29. Polish Mountain 30. Ruda 31. Kameniec 32. Keffaleneas 33. Orino 34. Magyar Racka 35. Zelazna 36. Kalarritiko 37. Karagouniko 38. Transylvanian Merino 39. Pilioritiko 40. Skopelos 41. Turcana 42. Kymi 43. Sfakia 44. Romanian Tsigaia 45. Anogeiano 46. Lesvos 47. Black Karakul 48. Ossimi 49. Akkaraman 50. Morkaraman.

It is a rather difficult exercise to write about exploratory analysis and to describe its usefulness through applied examples as it is in essence highly interactive and dynamic. Hopefully interactivity will stand out through the different examples analyzed in the following pages[1]. Different exploratory methods have been selected to illustrate in a didactic way how GVIS is likely to make us gain insight into relationships between genetic data and geoenvironmental profiles with regard to their geographic distribution. This will help characterizing breeds according to a genetic point of view *while* taking into account their geographic distribution *and while* considering their relationship with the natural environment.

Sheep and goat breeds data sets won't be fully analyzed, but an approach is proposed to show a way among many others to extract information, the point rather being to demonstrate the usefulness of spatial exploratory analysis applied to molecular data[2]. Anyway, there was a development to invent as ESDA is precisely free of any strict procedure. Thus, successive ESDA features will show :

- how to detect best correlations in order to identify promising associations to be investigated with additional exploratory tools, and which one could be exploited in a conservation perspective;

- how variables so brought to the fore can be used to produce interactive thematic maps to visually compare spatial patterns and to identify possible regional effects;

- how breed classifications can be built on the basis of genetic and geoenvironmental data simultaneously;

- how to interpret breed classes with the help of additional investigation tools;

- and how to assess genetic data congruence by analysing molecular markers signatures.

The thread of the developments presented hereafter is based on the observations carried out during the exploratory phase. Aside from the general hypothesis formulated by Cavalli-Sforza *et al.* (1994) that early farmers (Neolithic) came to Western Europe from the Middle East and that theses migrations would have generated circular gradients of gene frequencies around the region of origin (genetic diversity decrease; see also Pringle, 1998; Zeder and Hesse, 2000; Luikart et al., 2001; MacHugh and Bradley, 2001), there is no other underlying work hypothesis to be considered.

Correlations

Exploratory spatial analysis was carried out with the help of an easy to use and efficient software called CommonGIS[3] and developed by the CommonGIS international consortium in the context of an ESPRIT european research project. Some operations were also realized with Geovista[4], a Java GVIS software developed by the Department of Geography of the Pennsylvania State University. In addition to

---

1. The representation of maps and graphs is «mono-display» in this chapter. It is important to keep in mind that when making use of spatial exploratory analysis, all chosen functions are simultaneously displayed allowing the user to directly compare respective positions or ranking of studied individuals. Each point of each scattergram, each bar section of each histogram is clickable and allows the immediate identification of an individual on a map.

2. The number of possible exploring ways is very important and reporting for all of them is impossible in this context.

3. http://commongis.jrc.it/, http://www.commongis.com (10.11.2005)

4. http://www.geovista.psu.edu/index.jsp (10.11.2005)

standard GVIS features, this application allows to develop complex applications for data exploration and knowledge construction, and is consequently a bit more complex to utilize.

As a starting point, correlations were calculated between all genetic and geoenvironmental variables of the data set (see names and description of the variables in appendix 4). A specific tool displays all correlations to quickly and visually detect higher ones. Then it is possible to set a threshold under which coefficients will not appear in the result table (see appendix 5), and to spatially restrict on the map the zones in which individuals have to be taken into account to process the calculation, and to compare those partial results with the whole study area. These operations are realized «on the fly» and each modification is immediately processed to show up-to-date results.

Focusing on relationships between molecular and geoenvironmental data and considering the whole area, after having visualized all correlations, a threshold was arbitrary set to 0.44 to keep only the most significant ones which are detailed in table 5.1 for goats and table 5.2 for sheep. First of all, it is to observe that very few noteworthy correlations exist for both sheep and goat data sets. In fact, it is only in goats that two relatively significant correlations can be identified with a positive (line g4 in table 5.1) and a negative one (g7). The others can be regarded as low, and that is the reason why following analyses will be carried out and illustrated on goat breeds only.

	Molecular variable	Geoenvironmental variable	Correlation coefficient
g1	Microsat. Obs. heterozygosity	Longitude	0.5
g2	Microsat. mean number of alleles	Number of wet days	- 0.46
g3	Y chrom. haplogroup C frequency	Latitude	- 0.54
g4	Y chrom. haplogroup C frequency	Diurnal temperature range	**0.73**
g5	Y chrom. haplogroup C frequency	Number of days with ground frost	-0.47
g6	Y chrom. haplogroup C frequency	Temperature	0.29 (0.49*)
g7	Y chrom. haplogroup C frequency	Relative humidity	**- 0.59**
g8	Y chrom. haplogroup C frequency	Maximum possible sunshine	0.56
g9	Y chrom. haplogroup B frequency	Precipitations	0.54

Table 5.1. Goat data : table of correlation coefficients > 0.44. See appendix 4 for a complete description of the different variables. 9 relationships are presented here on a total of 204 correlations calculated between molecular and geoenvironmental variables in goats. On line G6, the correlation coefficient of 0.29 is obtained when calculated on all breeds. When the turkish Abaza and Turku (breeds with the highest altitude in the data set) are removed, the coefficient of correlation is 0.49.

	Molecular variable	Geoenvironmental variable	Correlation coefficient
s1	mtDNA haplogroup A frequency	Longitude	- 0.45
s2	mtDNA haplogroup B frequency	Longitude	0.45
s3	Y chrom- haplogroup E frequency	Latitude	0.44
s4	Y chrom- haplogroup E frequency	Number of wet days	0.46
s5	Y chrom- haplogroup E frequency	Relative humidity	0.44

Table 5.2. Sheep data : table of correlation coefficients > 0.44. See appendix 4 for a complete description of the different variables. 5 relationships are presented here on a total of 312 correlations calculated between molecular and geoenvironmental variables in sheep.

ESDA tools can facilitate the computation of statistics by subclasses. Notably, there is a function designed to dynamically restrict the number and the location of breeds to be analyzed. It was exploited to roughly assess correlations coefficients on the basis of various geographic criteria. For instance, when selecting Southern goat breeds only, *mtDNA* variables show correlations higher than 0.44 with environmental variables which is not the case when selecting only Northern ones. Or in the same way, the «response» of mtDNA variables is also much more «sensitive» to environmental data when selecting only goat breeds with a high microsatellite observed heterozygosity (> 0.65 with a range of 0.44 - 0.71). Finally, if we consider the five sheep breeds of Northern Poland only (Pomeranian, Polish Merino, Polish Heat, Kameniec and Zelazna all located in low lands), a number of genetic variables show a high correlation with environmental data (notably the microsatellite expected heterozygosity with precipitations = 0.88), and the coefficients decrease in an important way just by adding one breed, located in the polish Southern mountains. This interactive feature is typical of GVIS applications and is clearly illustrating the process in which «the tremendous human capacities, in terms of visualization, of intuition, of reasoning by analogy and of spontaneous hypotheses generation, are thus fully exploited in the context of a playing and realistic human-to-computer relationship, making the best use of the qualities of both parts» (Banos, 2001).
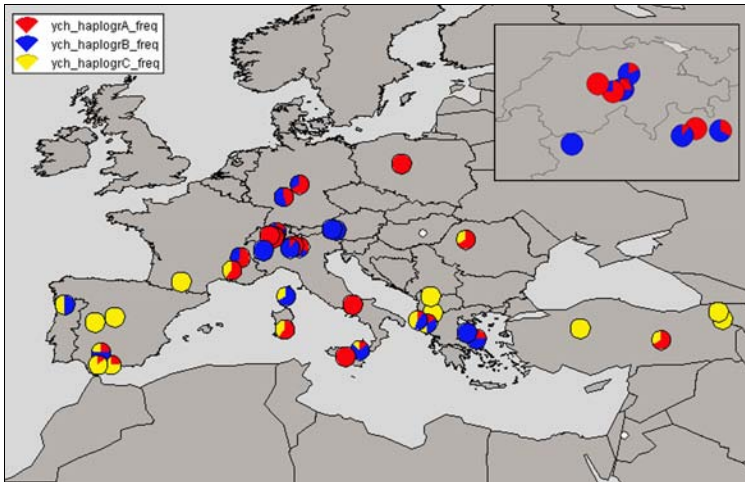
Fig 5.3. Spatial distribution of Y chromosome *haplogroup*s in goat breeds. White points are representing breeds with missing data. This figure is revealing specificities of GVIS in comparison with cartography (see figure 6.5 on page 87) : the view is temporary and it is therefore not necessary to take care over the design of the map. On this figure, breeds are displayed according to their Econogene location (see explanations on page 39). On the contrary, swiss breeds are displayed according to their historical origin on figure 6.5 on page 87 in order to improve readability where pies are very close to one another.

We will go further into the analysis of Y chromosome[1] haplogroup[2] C in goats which seemingly is globally (considering all breeds) «sensitive» to environmental

---

1. This means information collected in males only, that is one third of the animals.
2. Large groups of *haplotype*s (see chapter 4). A haplotype is a set of closely linked genetic markers present on one chromosome and which tend to be inherited together.

characteristics - compared to other molecular markers at least - and which provides information allowing to surround regions where it is likely to be found : rather south (see correlations for relationships g3, g6, g8 in table 5.1), probably in mountains (g4, g7), but not too high (g5). This profile is to be relativized by the modest values of the observed correlation coefficients, and in addition it is necessary to have recourse to the map of the spatial distribution of Y chromosome haplogroups in goat breeds (figure 5.3) to complete the «Identikit». Actually it is possible to visually notice a general Southern favorite location of the C haplogroup, confirming g3.

A way to check this Y chromosome haplogroup C potentially preferential ecological profile is to consult geoenvironmental data describing the rearing areas of the breeds. Figure 5.4 shows two goat breeds groups, the black one for which the Y chromosome haplogroup C is present, and the grey one for which it is not, with the mean of the different geoenvironmental variables. Although it was not a variable involved into highlighted correlations, the mean altitude of breeds containing the Y chromosome haplogroup C provides a first general indication confirming the *a priori* profile, as it is 776 meters high in comparison with the other group for which the mean altitude is 594 meters. And going on with this control, for breeds containing the Y chromosome haplogroup C, one can observe that for each variable involved in a positive correlation (g4, g6, g8), values are higher (green arrows) than the ones of the other group, and for each variable involved in a negative one, their values are lower (red arrows). The initial interpretation is apparently confirmed.
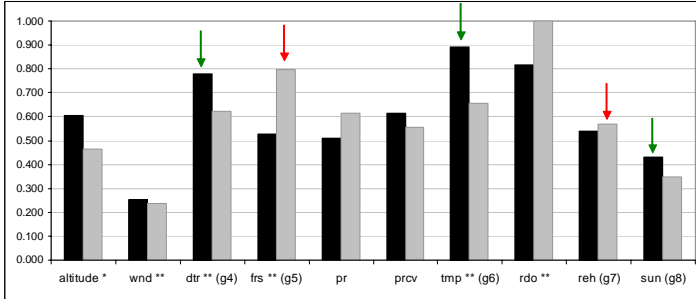
Fig 5.4. Ecology of Y chromosome haplogroups. Black bars are representing the group of breeds containing Y chromosome haplogroup C and grey bars the group of breeds without it. Values on the Y-axis are standardized means of the geoenvironmental variables displayed on the X-axis. The green arrows point out variables for which the group with haplogroup C shows higher values what confirms a positive correlation; the red arrows confirm a negative correlation. Arrows have been placed only for the variables appearing in table 5.1. Last element to confirm the a priori interpretation of the ecology of the Y chromosome haplogroup C is the altitude whose mean is higher for the breeds containing the haplogroup C. (*) real value x 1E-10; (**) real value x 10.

It is interesting to pick out that goat breeds with the Y chromosome haplogroup C are often homogeneous, showing no diversity (haplogroup C only, see figure 5.3 on page 61), or just a low one. This general homogeneity is conform to observations of weak phylogeographic structure in domestic goat revealed by the analysis of a specific *mtDNA* segment led on 400 goats originating from various countries of Europe, Asia, Africa and Middle/Near East (Luikart *et al.*, 2001). Econogene mtDNA data by the way show the same characteristic (figure 5.5 on page 63).

In the case of data presented on figure 5.3 on page 61, this observation is particularly amazing for turkish breeds supposed to be genetically more diverse according to hypotheses about livestock origins and migrations during the Neolithic expansion (Loftus *et al.*, 1999). Bruford *et al.* (2003) are explaining it by opposing cattle - which is showing a progressive loss a diversity from South-East to North-West with mtDNA data - to goat breeds, assuming that the latter are much more «portable» and that they have been more moved from regions to others, in «extensive intercontinental transportation» according to Luikart *et al.* (2001) who also suppose on that basis the importance of goats in historical human migrations and commerce. Those considerations about breed origins and prehistorical preoccupations are important to identify specific rare lineages. Indeed, uniqueness of a breed is one of the most obvious reasons for conservation, provided that it shows either a very high or a very low diversity (this is conservation in a narrow sense, meaning «preservation» of a breed which is rare). But it is also important to conserve different breeds that most likely possess different alleles and gene combinations. This to ensure that a range of traits that might be important for adaptation, for production or also scientific purposes are conserved (this is conservation in a broad sense, referring to operations in the management of animal genetic resources).
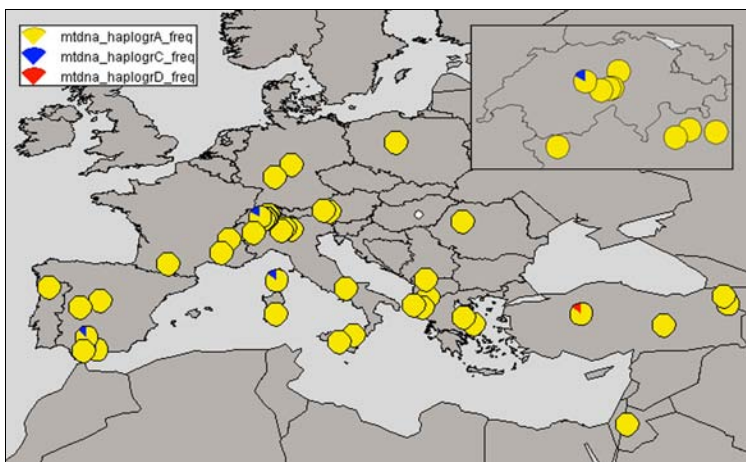
Fig 5.5. Spatial distribution of mtDNA haplogroups (maternal origin) in goat breeds to be compared with Y chromosome haplogroups (figure 5.3 on page 61). This map highlights the weak phylogeographic diversity in domestic goats. The white point shows missing data.

We will now assess how these genetic specificities (uniqueness/diversity) are distributed among the breeds and how the latter could be classified on the basis of these criteria in order to facilitate the undertaking of any conservation action.

Cluster analysis

Exploratory analysis makes us face the diversity of data, and it may happen that despondency arises because encouraging hypotheses emerging during one manipulation are immediately swept away by the next one. Ordering the information is a way to reduce an apparently «chaotic diversity into understandable, man-

Chapter 5

EXPLORING THE SPATIAL DIMENSION OF MOLECULAR DATA
Combined analysis of genetic and ecological data of goat and sheep breeds

ageable arrangements before scientific explanations are possible» (Mayr and Bock, 2002). Classifications, defined as a subset of ordering systems by Mayr and Bock (2002), is a way to arrange «a diversity of entities into sets of classes based on similarities possessed by the included individual entities». In parallel with specific goals which are developed farther, several generic ones are existing for ordering an apparent diversity in information. They were here adapted to breeds :

- to recognize similar classes and find delimitations between groups because «a greater number of propositions can be made» (Mill [1843] quoted by Mayr and Bock [2002]) about those groups;
- to identify a possible unknown breed;
- to predict characters of any additional breed located near one of the constituted groups;
- to serve as point of reference in comparative studies.

Means are provided within spatial data exploration tools to precisely sort out information through its different dimensions (in our case 19 molecular and 10 geoenvironmental variables, see appendix 4) in order to produce a more apprehensible reality and to reach the different goals expressed here above. Clustering consists of partitioning a data set into subsets (clusters), so that the data in each subset share some common trait - often similarity or proximity - for some defined distance measure. Data clustering algorithms can be hierarchical or partitional : with hierarchical algorithms (much appreciated in the life sciences), successive clusters are found using previously established groupings, whereas partitional algorithms determine all clusters in one calculation. Among its various functions, CommonGIS offers the possibility to make use of a partitional clustering algorithm and thus to end up at a bio-environmental[1] classification on the basis of a purely statistical grouping of individuals and of their interpretation.

K-means algorithm

The K-means algorithm (MacQueen, 1967) was used within the CommonGIS environment thanks to the integration of the WEKA public domain data mining software[2]. In comparison with other partitional clustering algorithms (Quality Threshold Clustering, Fuzzy c-means clustering, spectral clustering[3]), K-means is fast, doesn't require any specific preparation of the different data sets and is particularly easy to use. Its main weakness consists of the fact that it has to be told the number of clusters (k) to be found. Initially, it is necessary to define k a priori temporary centres (one for each cluster) which are located at random in the multidimensional scatter of points.

All points belonging to the different data sets are associated with their nearest centre and this constitutes an early grouping together. Then each one of the k centres is calculated as the centroid of the points it «owns» and a new association is established with the nearest points of the data sets, and so on. The k centroids change their location step by step until they don't move any more.

---

1. The term «environment» is combined with «biological» to specify that the classification is made according to both molecular and geoenvironmental data. About the concept of «biological classification» which is not single, please refer to Mayr and Bock (2002). By resorting to it, we simply mean that studied breeds can be arranged according to criteria based on molecular data.
2. Waikato Environment for Knowledge Analysis, Ian Witten and Eibe Frank (2005) «Data Mining : Practical machine learning tools and techniques», 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
See http://www.cs.waikato.ac.nz/ml/weka/ (10.11.2005)
3. Romesburg, H.C. (2004) Cluster Analysis for Researchers, Lulu Press Incorporated, Morrisville.
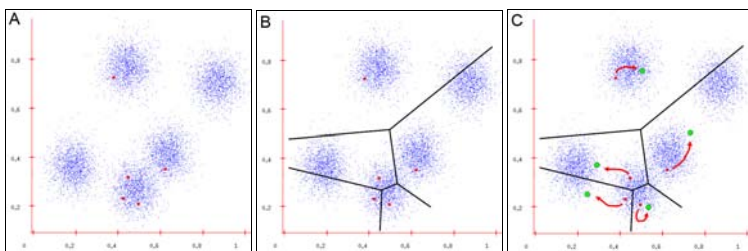
Fig 5.6. K-means algorithm. Examined individuals, the blue points, are defined by two variables. This is a 2D example while the operation on goat breeds is carried out with 29 variables. The user arbitrary chose to classify this information in 5 clusters. A) The K (5) clusters centres are randomly located. B) Each data point is attributed to its closest centre. C) Each initial centre finds the centroid of the points it owns and moves to this location. Then a new iteration occurs, taking into account the new centroid positions, and this until the cluster membership of data points remains stable. Source : Copyright © 2001, 2004, Andrew W. Moore, http://www.cs.cmu.edu/~awm/tutorials (10.11.2005)

The algorithm aims at minimizing an objective squared error function.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

(EQ 5.1)

Where k = number of clusters and n = number of individuals

and

$$\left\| x_i^{(j)} - c_j \right\|^2$$

is a measure of the distance between a data point $x_i^{(j)}$ and the cluster centre c_j.

Classification of goat breeds

Several configurations with different number of classes were tested in order to establish a classification whose number of groups best fitted both emergent molecular profiles and an apparent geographic structure (see appendix 6). This exploration phase led to the choice of a 5 goat breeds classes clustering, the most relevant to show how simultaneously investigated genetic and geoenvironmental variables might shape a resulting classification (figure 5.7 on page 67). The main objective of breed classification based on genetic criteria is to identify uniqueness and to manage biodiversity in order to avoid further loss of genepools (Mukesh *et al.*, 2004). But instead of focusing on genetic aspects only, the ecological dimension has been added to observe how it can improve the discrimination between classes, and to which extent this dimension can be linked to genetics to better grasp possible evidences of environmental pressure shaping breeds traits. In other words, this is a way to identify given characteristics of genetic qualities compared to ecological conditions found at a set of locations where a breed is raised. This is approaching a potential habitat notion whose use and delimitation could facilitate conservation actions. Reservations may be put forward as studied species are domestic and thus eminently transportable, and of course less liable to show signs of environmental pressure. But one can bank on the fact that autochthonous

breeds didn't significantly move out of their original rearing region in comparison with cosmopolite ones.

It is amusing in passing to pick up a common element between the present grouping of breeds and an aspect stressed by Mayr and Bock (2002) in their deep clarification about ordering systems. They argue that classifications are ordering systems, but that not all ordering systems are classifications like the Hennigian[1] system of cladification which consists of the ordering of branches of phylogenetic trees. The latter is based on one single criterion what «does not lead to classes of entities possessing similar phenotypic attributes», unlike Darwinian classification (or evolutionary classification) which is using *two* criteria leading to the recognition of classes of similar entities (phenotypic similarity *and* genealogy). In our case, we strengthen a grouping based on genetic similarities - which are possibly holding the common descent information[2] - by ecological ones which are providing information on breeds environment and thus potentially participating in shaping **phenotypes**. So that it corresponds to Mayr and Bock's arguing.

A five classes configuration

The present breed classification(see figure 5.7 on page 67) is made up of two main groups (colors have been arbitrarily chosen) :

- Red class with 12 breeds, that is 29.3% of the total number of breeds;

- Yellow class (11 breeds, 26.8%);

and of smaller ones :

- Purple class (8 breeds, 19.5%);

- Green class (6 breeds, 14.6%);

- Blue class (4 breeds, 9.8%).

This grouping configuration was not chosen at first attempt, but established by using results of the first exploration step (correlations) and consulting different graphs constructed on the basis of temporary-experimental clustering information. As the interactive exploratory process cannot be presented, the reviewing of scattergrams, histograms and parallel coordinates plots based on the configuration illustrated in figure 5.7 on page 67 is used to show indications which are supporting this choice - or more precisely this *adjustment* - as a relevant classification.

---

1. Willi Hennig (1913 - 1976), german biologist known as the founder of phylogenetic cladistics which is an approach that classifies organisms according to the order in time at which branches arise along a phylogenetic tree, without considering the degree of morphological divergence. This reveals an important conceptual difference in the use of different grouping criteria : «order of branching» vs. «similarity and difference», or clades vs. taxa (Grant, 2003).
2. «The sorting of species into similarity classes is simultaneously also a process of phylogenetic sorting, because usually a class of similar species consists of the descendants of a common ancestor» (Mayr and Bock, 2002).
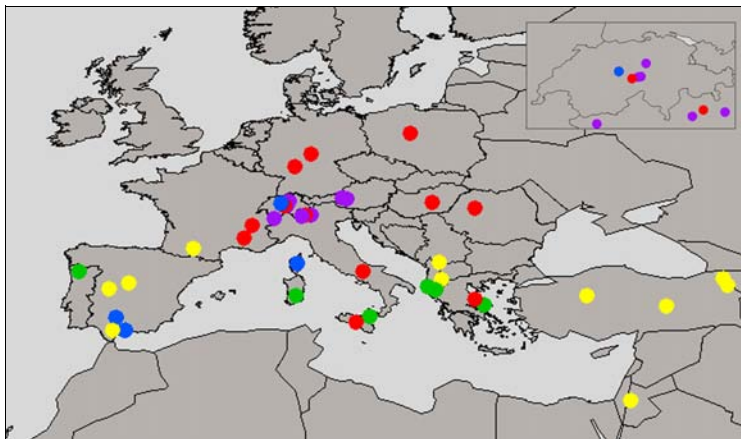
Fig 5.7. Map showing goat breeds clustering in 5 classes (see figure 5.1 on page 58 for breed names). **Red class** : Camosciata, Carpathian, French Alpine, German Alpine, Girgentana, Grigia Molisana, Hungarian Native, Polish Fawn Colored, Rove, Skopelos, Swiss Alpine, Thuringian | **Yellow class** : Abaza, Angora, Baladi, Cabra del Guadarrama, Capore, Gurcu, Hair, Hasi, Payoya, Pyrenean, Verata | **Purple class** : Bionda dell'Adamello, Grisons Striped, Orobica, Peacock, Pinzgauer, Tauernschecken, Valais Black Neck, Valdostana | **Green class** : Argentata dell'Etna, Brava, Dukati, Greek Goat, Muzhake, Sarda | **Blue class** : Corsican, Florida, Malaguena, St Gallen Booted.

Main relationships detected between genetic and environmental data

Among the higher positive and negative correlations early detected in analysis and displayed in figure 5.8 on page 68, the relationship between the frequency of Y chromosome haplogroup C and both mean diurnal temperature range (*dtr*) and relative humidity (*reh*) is strongly discriminating the main red and yellow groups, and the purple one. Indeed, considering diurnal temperature range, this haplogroup is totally absent from purple breeds and from many red breeds (excepted Carpathian and Rove) while it is present for each yellow breed.

Two additional interesting scattergrams are displayed in figure 5.8 on page 68 :

- The coefficient of correlation observed between microsatellite heterozygosity and longitude (G1), even though not very high, is possibly showing the supposed South-Eastern origins (Fertile Crescent) of livestock breeds (Pringle, 1998; Zeder and Hesse, 2000; Luikart *et al.*, 2001), despite breeds goats transportability as mentioned above, and despite the absence of Eastern Turkey breeds information (missing data). Indeed, the trend of an increase in diversity towards Eastern regions is clear in spite of the outlier behaviour of Pyrenean (in yellow, with the lowest microsatellite heterozygosity) and of two blue spanish breeds, Florida and Malaguena. An hypothesis could be that outliers breeds whose genetic diversity *grosso modo* doesn't follow the trend precisely are more «portable», and thus would be breeds which have been relocated for commercial purpose. It is to note that there is no breed homogeneity according to the microsatellite heterozygosity, all of them presenting a large range of values.

- The negative correlation observed between the mean number of alleles in microsatellites and the number of wet days (G2) ensues from the same phenomena, that is an increase in genetic diversity following a decrease in the number of wet days. In the context of the present sampling, this last point can be translated as moving towards South Eastern regions. Considering this environmental parameter, we can observe that yellow breeds are much more compactly grouped than the other ones.
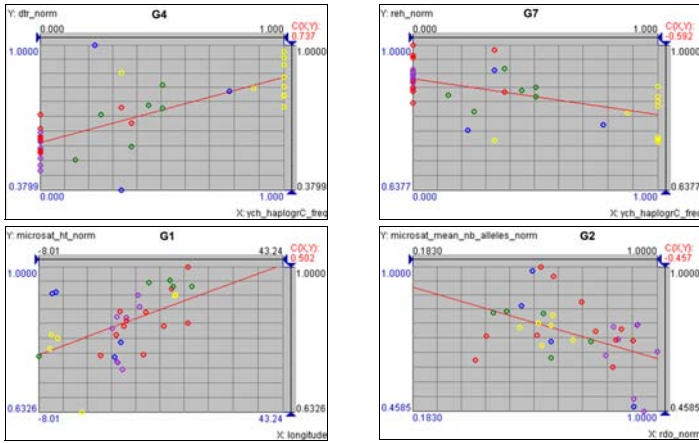
Fig 5.8. Scattergrams of variables with the 2 higher positive or negative coefficients of correlation presented in table 5.1 on page 60 (G4 and G7), and of relationships involving microsatellite variables (G1 and G2). Colors of the points representing breeds are corresponding to the different classes revealed in table 5.7 on page 67.

The fact of containing haplogroup C doesn't make Carpathian and Rove males behave differently compared with the rest of the red breed. Relative humidity has the same discriminating effect, inducing a negative correlation in comparison with diurnal temperature range. Both are globally composing dry versus humid environments to which at least males with Y chromosome haplogroup C are «responding».

Genetic composition of the classes

Another way to study the composition of the classes and to propose interpretations for the resulting classification is to analyze the distribution of any variable through the use of histograms[1]. Figure 5.9 on page 69 is displaying histograms of molecular variables and farther figure 5.14 on page 75 will show histograms of geoenvironmental variables. According to the systematics we adopted since the beginning of spatial data exploration, variables which were shown to be significantly involved in genetics-environment interaction processes in goats are displayed first.

This way of representing the structure of groups shows again that Y chromosome haplogroup C (histogram A in figure 5.9) is well discriminating the yellow and the red class. However, this histogram permits to identify the yellow Hair breed (like in figure 5.8 on page 68, *dtr* and *reh*) in which this haplogroup is present in only one third of the animals, what is much different from the other breeds of the same class, and from the same region especially. Y chromosome haplogroup B (B) characterizes the purple class for which it is always present at a rate of more than 60%. Microsatellite observed heterozygosity (C) apparently doesn't provide obvious information, showing a low value for the Pyrenean (yellow, on the left), and indi-

---

1. Variables used for histogram analysis have not necessarily been used to constitute the classes.

rectly gives credit to goat breeds low genetic structuring mentioned above and reported by Luikart *et al.* (2001).
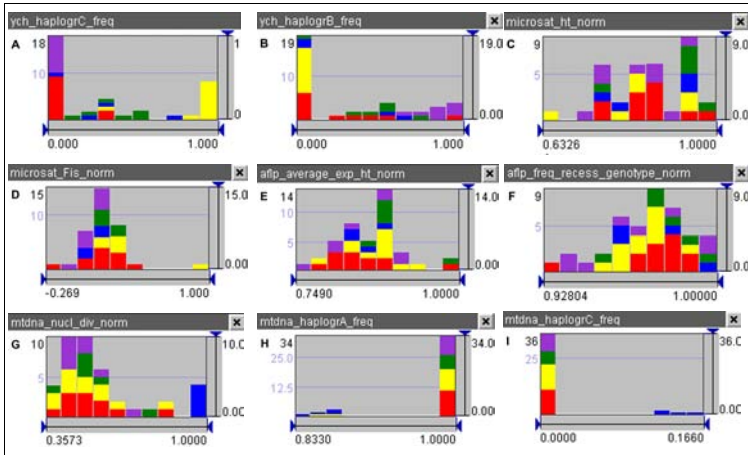
Fig 5.9. Histograms of the molecular variables. The X-axis is representing the standardised values of the variable and the Y-axis the frequency of breeds within quantiles 10. The color of the bars is corresponding to one of the classes presented on figure 5.7 on page 67. The numeral on the top left of the graph stands for the number of breeds in the higher bar. Please refer to appendix 4 to read the complete name of molecular variables.

The microsatellite F_{IS} (D) indicates a global homogeneity between breeds whose majority is almost at equilibrium. This makes the Pyrenean clearly emerge as an inbred breed (F_{IS} = 0.25, isolated yellow on the right) in opposition with Hungarian Native (red) and Pinzgauer (purple) which are slightly outbred (respectively -0.02 and -0.07).

AFLP data (E and F) show a compact red group with the notable exception of Carpathian which presents more diversity than the others (average expected heterozygosity, on the right). Looking at the last row of figure 5.9, mtDNA data (mtDNA is inherited from the mother) especially give information about the blue group, and about the green one to a lesser extent. At a first glance on all three mtDNA histograms (G, H, and I), it seems obvious that this is the marker making the identity of blue breeds. About *nucleotides* diversity, St.Gallen Booted, Malaguena, Florida and Corsican - the four breeds constituting the blue group - manifest the four higher rates respectively. The haplogroup A (which is not correlated with any environmental parameter) is almost absent from those breeds in opposition with all others for which it is always present. It is very different about haplogroup C which is present only in blue breeds and not at all for the other groups. This haplogroup C is somewhat (0.34) correlated with the coefficient of variation of monthly precipitations and with wind speed (0.24) but no need to say it is negligible. Finally, the identity of the green group can apparently hardly be assessed with the help of molecular variables only. One characteristic of this class is a high microsatellite observed heterozygosity with the exception of the Brava which is located far in the West, making this observation compatible with the general theory of livestock's Middle-Eastern origin.

Chapter 5

EXPLORING THE SPATIAL DIMENSION OF MOLECULAR DATA
Combined analysis of genetic and ecological data of goat and sheep breeds

In general, the genetic structure of breeds classes is not very neat although divulging rare clear indications (mtDNA nucleotides diversity). One can wonder if the fact of constituting clusters on the simultaneous basis of genetic and geoenvironmental data do not tend to dilute molecular «informativity». In fact, when running simple K-means clustering algorithm on molecular data only, the constitution of the groups differs very few from the «ecogenetic» one. This could be an indication about the fact that the ecological influence is already integrated within genetic information which is effectively expressing more or less - and globally - the «pressures» the environment puts on organisms. I cannot put other arguments forward to support this hypothesis for the moment and refer to developments made in the first part of chapter 7 about natural selection to better tackle the way organisms may be moulded by their environment. Let us rather focus on the class changes caused by this strictly genetic classification : it appears that this is precisely about the not well defined green class that most changes do occur (figure 5.10). By «inheriting» the «former» yellow Baladi and Hair, as well as the red Carpathian, and by «losing» the Sarda, it becomes clear that this group is made up of breeds presenting a really low mtDNA nucleotides diversity, in parallel with a compact behavior in AFLP data which is illustrated by high values of average expected heterozygosity, thus a high number of polymorphic loci, and finally low values for the Jaccard similarity index which is coherent with the diversity trend of this group.
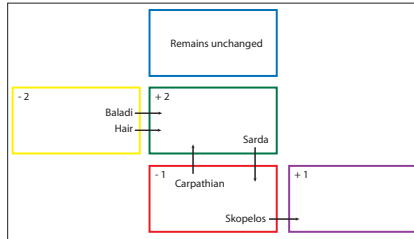
Fig 5.10. Changes occurring in the classification when the clustering is based on molecular variables only.

Apart from specifying useful elements regarding the interpretation of classification, the checking carried out on genetic data only is typical of a spatial data exploration approach. It illustrates rather well one of these innumerable «*va-et-vient*» (see chapter 3 on geographic modelling) to compare situations based on different data, while remaining reasonably time consuming.

Exploiting Principal Component Analysis results

The genetic interpretation of groups can also be refined by using the factorial scores of a principal component analysis (PCA) carried out on genetic distances[1]. A PCA was run on goat breeds AFLP data, using Reynolds (1983) distances. This analysis showed a first component accounting for 50% of the total variance individualizing the italian Orobica and the austrian Tauernschecken which have probably been «artificially» introduced in the Alps a long time ago from a remote area

---

1. A genetic distance is a measure of the genetic similarity between any pair of breeds, based on the allele frequencies for instance, on phenotypic traits, or DNA sequences.

(Negrini *et al.*, 2004). The second component (19.2% of the variance) is revealing a south-east to north-west pattern of genetic variation, conform to the direction of agriculture expansion.

This is interesting, but in the context of exploratory analysis, the important is rather the way the results of a PCA can be used : the factorial scores of each breed on the respective principal components are used as a genetic similarity variable to be projected through the different classes determined by the cluster analysis (figure 5.11).
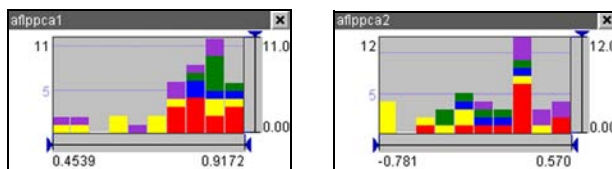
Fig 5.11. Histograms of factorial scores of goat breeds on the first two principal components (AFLP, Reynolds distances).

The histogram of the first component illustrates the individualization of the Orobica and the Tauernschecken (purple, on the left), here grouped together with Turkish Gurcu and Abaza breeds (in yellow). The green class is homogeneously defined on this component which is best represented by four different breeds (the reds German Alpine, Grigia Molisana, Carpathian; the yellow Cabra del Guadarrama; the blue Malaguena; the green Argentata). The figure concerning the second component presents a perfect dissociation of Turkish breeds (on the left with a negative correlation) and shows a global and gradual geographic longitudinal shifting towards West when reaching higher correlations on the right of the histogram.

Before examining the ecological interpretation of the classification, it is worth broaching an additional exploratory technique in order to assess molecular markers congruence and their behavior within and among goat breeds categories.

Parallel Coordinates Plots (PCP)

The parallel coordinates method is a multidimensional visualization technique which was originally proposed and implemented by Alfred Inselberg (1985). Since plotting more than 3 orthogonal axis is impossible, parallel coordinates schemes plot all the axes parallel to each other in a plane on which the geometric structure of the different variables is projected.

Each breed in the data set is represented by a line segment which intersects horizontal axes, each being scaled to a different variable. The value of each variable for a given breed is plotted along each axis relative to the minimum and the maximum of the distribution. The points are then connected using line segments. The result is a breed «signature» across n dimensions (figure 5.12 on page 72).

With the help of PCP, the intention is to globally compare the molecular signature of the five classes of breeds in order to complete the rough descriptive work previously made on the basis of histograms. The 5 goat classes and associated molecular signatures are displayed in figure 5.13 on page 73.
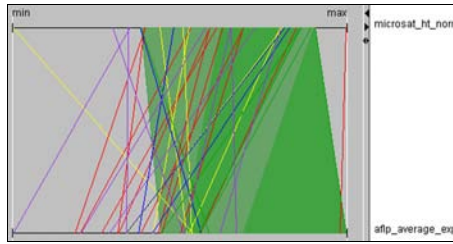
Fig 5.12. PCP didactic graph on which two variables are represented by black horizontal lines. Variables are standardized and all values are included between a minimum of 0 and a maximum of 1. Each line is a breed whose color is corresponding to its class. The initial and the terminal point of a colored line are the values on the respective variables. It is possible to show the range of the different classes and to dynamically switch them on and off in order to make comparisons. In this case, the green surface represents the range of the green class : it shows that the AFLP genetic diversity of this group is slightly higher than the microsatellites one.

These signatures permit to evaluate the behavior of the statistical distributions thanks to the display of quantiles (darker lines within the colored signature) while black horizontal intervals are showing the range for each variable among all breeds.

At an elementary visualization level, the shape of the signatures is carrying information. The more narrow it is, the more homogeneous and discriminant the different genetic variables. The purple group shows a general reduced covariance in comparison with the other classes. On the other hand, red and yellow groups present a rather large variance on many variables. Further main visual signals are a large variance in all Y chromosome variables, this being particularly obvious in green and blue groups.

To assess the way the 15 selected molecular variables are discriminating breeds, they have been arranged according to their variance in an increasing order in each class, from top to bottom (see figure 5.13 on page 73). This ranking therefore makes it possible to assess which are the more discriminants variables in each class.

Pair of classes	Kendall's W
Green - Red	0.904
Green - Blue	0.868
Red - Purple	0.866
Blue - Red	0.855
Green - Purple	0.807
Blue - Purple	0.779
Red - Yellow	0.732
Green - Yellow	0.691
Blue - Yellow	0.684
Purple - Yellow	0.577

Table 5.3. Discriminating power of molecular variables : coefficient of concordance[1] between breed classes (Kendall's W) in a decreasing order. The global Kendall's coefficient over the 5 classes is 0.642.

The corresponding Kendall's W, the coefficient of concordance between the respective ranking among the 5 groups, is 0.642 and thus reveals a medium level

---

1. The coefficient W ranges from 0 to 1, with 1 indicating complete inter-rater agreement, and 0 indicating complete disagreement among raters.
    Lecoutre, J.-P. and Tassi, P. (1987) Statistique non paramétrique et robustesse. Economica, Paris.

of correspondence between the discriminant power of molecular variables in the different classes.

The results of partial Kendall's tests are displayed in table 5.3 : molecular variables have almost the same effect in green and red groups. On the contrary, the lowest W between the purple and the yellow classes is affecting the value of the global coefficient of concordance between groups.
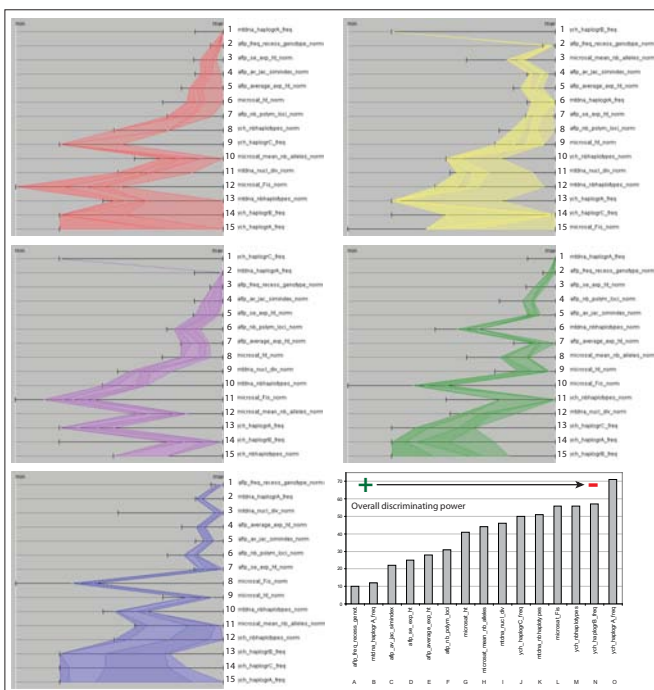
Fig 5.13. PCP molecular signatures of breed classes. The colored lines within a group stand for quartiles. The horizontal black lines show the range of the concerned variable for all breeds. Molecular variables are displayed in an increasing order according to their variance from top to bottom (excepted for the yellow and the purple groups in which the first variable is absent). A smaller variance has a higher discriminant power. MtDNA haplogroups C and D were ignored because the first is present only within blue breeds, and the second within Gurcu and Angora (Yellow class).

A) AFLP frequency of recessive genotype / B) mtDNA frequency of haplogroup A / C) AFLP Jaccard similarity index; D) AFLP standard error of average expected heterozygosity / E) AFLP average expected heterozygosity / F) AFLP number of polymorphic loci / G) microsatellite observed heterozygosity / H) microsatellite mean number of alleles I) mtDNA nucleotides diversity / J) Y chromosome frequency of haplogroup C / K) mtDNA nb of haplotypes L) microsatellite FIS / M) Y chromosome number of haplotypes / N) Y chromosome frequency of haplogroup B O) Y chromosome frequency of haplogroup A.

On the basis of the ranking carried out on the five PCPs, an overall score of molecular variables (sum of their ranks among the 5 breed groups) is showing an indication of the overall discriminating power of molecular variables (see histogram in figure 5.13). It clearly turns out that AFLP variables have the highest discriminating power in most of classes, and that the two first of them are similarity indices. On the contrary, Y chromosome variables show a large variance in all classes,

confirming the previous visual observation, although somewhat reduced in red and yellow groups.

PCP introduces the concept of a general genetic signature which is especially useful in order to quickly and visually compare groups or populations between them. Its main interest resides in the ability to provide a simultaneous synthetic view of every dimension of a molecular data set, allowing to check for a possible convergent behavior of genetic variables. Nevertheless it is rather delicate to use as one would tend to concurrently compare all the groups or individuals constituting a data set. Therefore it remains necessary to make choices to keep the information readable by displaying groups separately like in figure 5.13, or to test different successive combinations implying a few observations instead of producing indistinct heaps of colors.

One may think that this kind of tool is not more efficient than traditional box plots for example. But it is important not to lose sight of the fact that the PCPs shown on figure 5.13 are always displayed together with a map, with other graphs produced during the exploratory process, and that they all are dynamically linked : a single breed can constantly be identified among others as well as geographically localized.

Environmental composition of classes

To complete the interpretation, we will now examine the comportment of ecological variables to determine how they affect the composition of classes .

On figure 5.14, the first histogram clearly shows that the yellow group is mainly constituted of breeds located where diurnal temperature range is high, especially in comparison with the red and the purple classes. But though the *dtr* variable is distinctly structuring breed groups, its range within the yellow and the red classes is really wide, showing that this parameter allows to make a distinction between them, but that it is probably not determining as habitat condition (see map on figure 5.15 on page 75). Abaza and Gurcu turkish mountain breeds, together with Hair, Angora and the jordanian Baladi to a lesser extent, are constituting an Eastern subgroup of the yellow class which is the most influencing the *dtr* histogram. These yellow breeds are definitely the ones of sun and dryness, what is confirmed by histograms of frost (*frs*) (high altitude yellow breeds vs. low altitude but Southern ones) and of temperatures where the identification of the Eastern mountainous group is obvious.

The relative humidity histogram indicates a clear and compact grouping together of red and purple classes towards high *reh* values. The purple group is strictly present in the Alps in which winters are particularly moist in comparison with summers, the same observation being valid for red breeds located north of the Alps. Having a look on temperatures (*tmp*) confirms the existence of this purple alpine class from which Alpine breeds are absent (!), the French, German and Swiss Alpine being members of the red class. In fact, in opposition with the locations of purple and a part of yellow breeds living in altitude (500 to 1200m and 1300 to 2200m respectively), the Alpine are raised at a rather low altitude (250m for the German and 800m for the others).
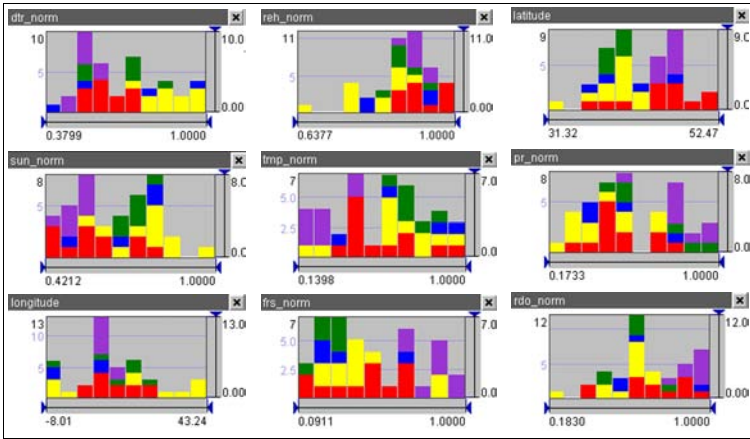
Fig 5.14. Histograms of the geoenvironmental variables. The X-axis is representing the normalised values of the variable and the Y-axis the frequency within quantiles 10. The color of the bars is corresponding to the one of classes presented on figure 5.7 on page 67.
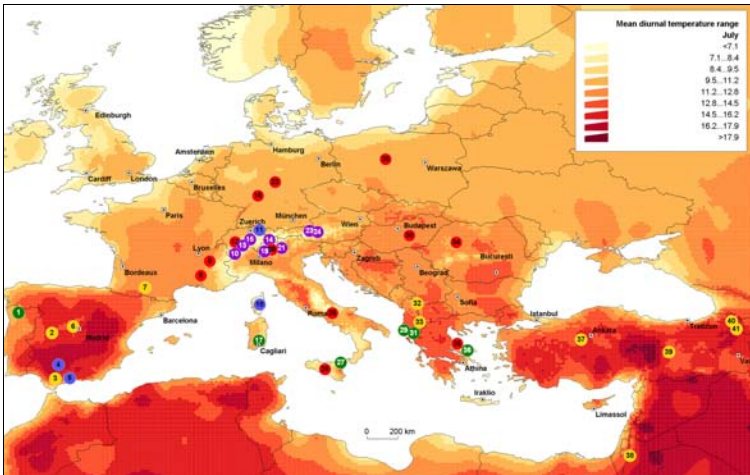
Fig 5.15. Map of mean diurnal temperature range in july (in order to accentuate contrasts). Temperature ranges are indicated in degrees C°. Please refer to figure 5.1 on page 58 for goat breed names.

About the geographic distribution of classes, longitude and latitude histograms are interesting insofar as they show something like a profile view of those two spatial dimensions. They probably help the reader in materializing in one dimension what he approximately gets from a two-dimensional map. For instance, given the broad longitudinal spread of the red class, the information provided by the 2D map (figure 5.15) is rather fuzzy in order to compare this group with the green, the blue or even the yellow one. It is not the case on the histogram on which the

structure of groups clearly appears with two extreme yellow breeds groups (Western and Eastern), a red central, a purple concentrated in a few kilometers, a widely distributed green and finally a rather Western blue class.

Precipitation (*pr*) values provide the same information like relative humidity especially because of the already described «Alps effect», which is here to be better defined as a «relief effect» on the one hand (the Pyrenean is also affected by a relatively important quantity of precipitations), and a «sea effect» (Brava and Muzhake) on the other hand. The number of wet days (*rdo*) above all individualize the «yellow» Baladi and the two most Southern red Girgentana and Skopelos. breeds.

With regard to this geoenvironmental interpretation attempt, we are confronted to two types of goat breeds. On the one side the ones which are definable on the basis of obvious *geographic* criteria, and on the other side breeds which are not. Yellow, green and purple groups at least show an obvious link with geography : latitude for green and yellow, and a compact localization in the Alps for the purple. The green class is not present beyond the Bosphorus, but chances exist that this tangible limitation in latitude nevertheless has an influence on specific genetic characteristics (total absence of Y chromosome haplogroup B for instance). As for the second type, the blue and the red classes are spread across vast areas, extended in longitude and in latitude. There is *a priori* no spatial evidence for distinguishing them.

To simultaneously take into account environmental and geographical observations previously made, a synthetic map produced by the European Environment Agency (EEA[1]) presents the biogeographical regions of Europe (Roekaerts, 2002; see figure 5.16 on page 77), derived from original Natural Vegetation Map units (scale = 1:3'000'000). The evidences this document shows are that the purple class is definitively Alpine and that the Carpathian (#34) is also raised in this kind of environment. Moreover, the green breeds are mediterranean. Turkish goats which seemed to constitute a yellow subgroup are all included in an Anatolian region where numerous steppes are gradually converted into arable lands thanks to intense irrigation (Roekaerts, 2002). Otherwise the yellow group is distributed amongst Alpine, Mediterranean and Atlantic (Pyrenean breed, #7) zones, blue breeds are not mediterranean because of the St.Gallen Booted (#11), and the red ones are spread among Mediterranean, Alpine, Continental and Pannonian zones which are mainly hungarian and whose characteristics are a mixing of eutrophication of large lakes due to the intensification of agriculture and heavy metals pollution due to mining industry (Roekaerts, 2002).

In fact, it turns out that the initial observations based only on information provided by a few environmental variables were correct, and above all that unfortunately the approach adopted by the EEA doesn't bring any new additional element. But it has the non negligible advantage to provide a complete representation of the biogeographical areas we are talking about and to situate breeds with this reference before the eyes.

---

1. http://www.eea.eu.int/ (11.11.2005)
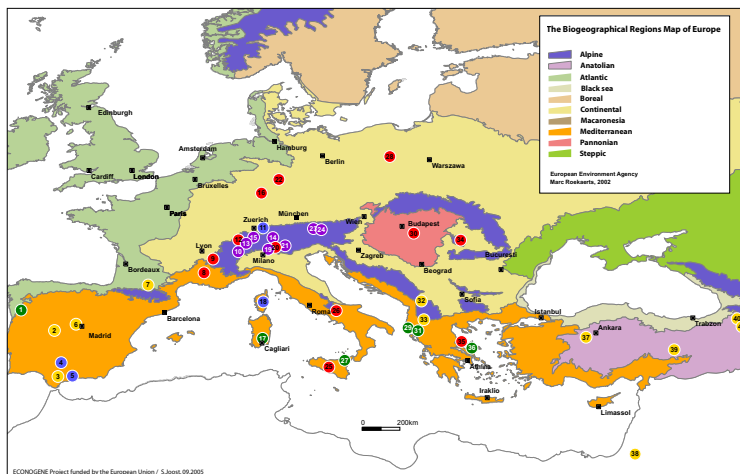
Fig 5.16. Map of the biogeographical regions of Europe with the location of Econogene goat breeds. Please refer to figure 5.1 on page 58 for goat breed names. Source : European Environment Agency, Roekaerts (2002).

Goat breeds classification digest

To summarize the information gathered with the help of different exploratory analysis tools about the description and the interpretation of the different goat breeds groups, here is a synthetic reminder.

- The yellow class is mainly related to sun and dryness. This is the only group present beyond the Bosphorus. A notable element is the total absence of Y chromosome haplogroup B in opposition with all other classes. A general high variance on most of genetic variables has to be noted;

- The red class is widely spread in longitude and latitude, centred on the Alps. Genetic specificities are : a) the presence of mtDNA haplogroup A for all animals of the group, as it is the case for the green and the purple one, and b) high microsatellites **and** AFLP diversity values (as it is also the case for the green class);

- The purple class is grouping together mountain breeds raised in the Alps (strong geographic cohesion). This is the only group for which Y chromosome haplogroup C is absent. AFLP similarity indices are high in opposition with the microsatellite F_{IS} which is low.

- The green class is the more difficult to characterize. It is localized in Southern areas, on islands or in zones strongly influenced by the sea. This class has a high genetic diversity both in microsatellites and AFLP (like the red one), but with a comparatively low number of microsatellite mean number of alleles and AFLP number of polymorphic loci.

- The blue class particularity is about mtDNA. The C haplogroup is present in this category, in opposition with all other breeds for which it is absent. Moreover, the *nucleotides* diversity is very high. Spatially indefinable.

Fig 5.17. Pictures of goat breeds. From the left to the right, top to bottom : 1. Verata, yellow class 2. Cabra del Guadarrama, yellow class 3. Camosciata, red class 4. Valdostana, purple class 5. Tauernschecken, purple class 6. Argentata dell'Etna, green class 7. Florida, blue class 8. Malaguena, blue class.

SUMMARY

This flying over a substantial case of exploratory spatial analysis applied to genetic data above all reveals that investigation possibilities are numerous, considering the fact that only major features have been illustrated in this chapter. ESDA and GVIS methods are easy to use, intuitive and flexible; they really allow to check any complementary hypothesis for a reasonable time cost. The spatial dimension can be constantly accessed and acts as an efficient support to bring investigated variables into coherence.

Molecular data very well lend themselves to this exercise and their comparison with geoenvironmental parameters indicated that deepened investigations on the relationships between genetic and environmental data are relevant. At least strong similarities between combined ecogenetic and strictly molecular clustering configurations let it suppose.

CARTOGRAPHIC REPRESENTATION OF GEOREFERENCED GENETIC DATA

6

Chapter outline

ESDA efficiently contributes in detecting relevant information out of large spatial molecular databases. Once identified and merely visualized, significant patterns in analyzed statistical individuals have to be staged in order to refine spatial analysis thanks to geo-graphic tricks, and to be communicated later on. Thematic cartography is resorting to the semiology of graphics to assume this role.

FROM GEO-GRAPHICS TO CARTOGRAPHY

In a similar way like Karl Popper articulated his thesis of a Science having to proceed by deduction, arguing that it was in the interplay between the tentative theories (hypotheses) and error elimination (refutation) that scientific knowledge was progressing towards complexity (Popper, 2002), the use of ESDA entails the generation of many hypotheses of which many are rejected, but among which some are supposed to improve knowledge. The latter, as long as they are not proven to be false (to keep on referring to Popper!) are concretized through raw spatial representation (see figure 5.7 on page 67 for instance). Then it is worth fixing those relevant observations - ephemeral until then - by representing them on maps to record, communicate or possibly analyze[1] the information they are carrying, to cite the three main functions of thematic cartography (Bertin, 1983).

At this point, a brief parenthesis is necessary. «Cartography», «map» and «mapping» are terms which have to be used with caution as they all are employed both in the geographic information field (the maps whose elaboration will be described in this chapter) and in genetics. With regard to a standard geographic map, a genetic map permits to locate a gene or a DNA sequence in a specific region of a chromosome in relation to known genes or DNA sequences (figure 6.1). But the lexical analogy is not restricted to those rather well known appellations as genes «atlas» are also existing, among which «Genatlas»[2] developed by the René Descartes University in Paris is an example[3]. End of the parenthesis.

In the previous chapter - with the rare exception of a few necessary maps whose role was to localize breeds (for instance figure 5.1 on page 58) - non elaborated spatial representations were used in order to allow a simple recognition of territo-

---

1. Most of the time it consists of interpretation.
2. http://www.dsi.univ-paris5.fr/genatlas/ (14.11.2005)
3. While picking up GIScience and genetics ambiguities of lexical or semantic distinctions, let us also mention the anecdotal but real Genome Institute of Singapore (GIS !). http://www.gis.a-star.edu.sg/internet/site/ (14.11.2005)

ries thanks to their shape. The well known general shape of Europe and its coun-
tries are widely assimilated in mental representations of the readers, and the fact
of displaying no additional information makes the attention mainly focus on sta-
tistical information transmitted through colored objects.

Fig 6.1. Genetic map. The on-line Genbank map viewer[1] is an interactive application to map genomes regions. Here
the *Homo sapiens* chromosome 1 with the illustration of the zoom function from the left to the right. Source :
http://www.ncbi.nlm.nih.gov (14.11.2005)

During exploratory analysis, no particular care was dedicated to the representa-
tion of geographical objects which are displayed on *geo-graphs* rather than on
maps. These geo-graphs are mere supports for data representation, other graphic
displays like the many invented by John Tukey (1977). We are confronted to the
subtle notion of geo-graphic intermediary analyzed by Rappo (1994); he was
alluding GIScience tools («computer geography»), intermediaries providing the
possibility to be situated between a real space (a region, a territory) and a graphic
one (a world of models or representations). It is possible to go one step further
within the graphic space only - and keeping the intermediary concept - to con-
sider that ESDA representations precisely constitute geo-graphic intermediaries
between an unknown reality represented by statistics which are explored and sit-
uations selected among them suspected to reveal pieces of knowledge. These ones
have to be exploited and will overtake this intermediary condition by a) being
represented with care[2] and b) being spatially contextualized. The first case alludes
the staging of geographic information and related representation rules, and the
second one aims at making the most of relevant statistical observations by reveal-
ing a geographical context composed of chosen objects (rivers, mountains) pre-
cisely because they are supposed to interact - or to be put in touch - with statistical
observations. This interaction between data and a model of geographical reality is
likely to produce information and knowledge in turn. Both points constitute
improved ways of showing spatial information I'm naming thematic cartography
and which will be described in this chapter.

---

1. http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi (14.11.2005)
2. «Care» signifies that the semiology of graphics exposed in the following pages will be applied, and also that
aesthetic aspects have to be taken into account. The latter are developed by Edward Tufte (1983) and won't be
treated in detail in this chapter. Tufte notably argues that simplicity of design together with complex data are two
key-elements of good design. Moreover, he advocates visually attractive graphics and proposes a list of guides to
enhance the quality of statistical information display (Tufte, 1983, p.177).

THEMATIC CARTOGRAPHY
· ·

A consensual and synthetic definition of thematic cartography - or thematic map design - is the art of communicating information regarding geographical objects (points, lines and areas) by using various shading techniques, and possibly having recourse to a physical geographic context. By shading a spatial object based upon the attribute values or range of attributes, spatial relationships and patterns can be communicated (MacEachren, 1995; Brunet, 1987).

Here are a few historical landmarks to show how thematic cartography developed, to realize how it is positioned with regard to both graphical design and geographic information management and analysis. Before the 17th century[1], the earliest signs of maps or visualization supports consisted in tables showing the positions of stars, or in navigation and exploration maps. Later, the 17th century was mainly dedicated to physical measurement of time and distances, and the context of territorial expansion entailed map making. Moreover, theoretical developments like theory of errors, probability and demographic statistics, which will turn out to play a role within thematic cartography and visual thinking, were emerging. During the 18th century, new graphic forms were proposed (isolines and contours) and thematic mapping of physical quantities was initiated to produce maps on the base of geologic, economic, and medical data. It was the beginning of statistical theory and empirical data began to be systematically collected. Technical innovations like lithography facilitated the reproduction and the use of images containing data. But it was during the first half of the 19th century that statistical graphics and thematic cartography made decisive progresses. In 1826, Charles Dupin published in France the first choropleth[2] map entitled «Carte figurative de l'instruction populaire de la France», what constituted the beginning of modern data graphics. Many original forms of symbolism were introduced in cartography, and in parallel a lot of the modern forms of data display like bar and pie charts, histograms, etc., were invented. Toward 1850, official national statistical offices were established in Europe, and statistical theory initiated by Gauss and Laplace made it possible to exploit large bodies of data. The period from the early 1900s to the mid 1960s is described as a time of dormancy by Friendly and Denis (2005). It is a time for application and popularization, rather than for innovation. Then data visualization began to rise from lethargy, launched and stimulated by significant developments. First in 1962, when John Tukey issued his call for the recognition of data analysis (see chapter 5, page 56) and launched EDA. Then in 1967, when Jacques Bertin published «Sémiologie Graphique», a huge work dedicated to the organization of the visual and perceptual elements of graphics according to the features and relations in data (Bertin, 1983). Those two major contributions constituted the back-cloth of the beginning of computer data processing. Computer science research at Bell Laboratories (software tools, C language, UNIX, etc.) together with developments in data analysis - notably pioneer GIS developments by Roger Tomlinson in Canada in the early 1960s - and in display

---

1. This review of thematic cartography progressive advances was made possible thanks to a considerable research work carried out by Michael Friendly and Daniel J. Denis of the Statistical Consulting Office at the York University in Toronto (Friendly & Denis, 2005).
2. A map with areas colored or shaded such that the darkness or lightness of an area symbol is proportional to the density of the mapped phenomena or is symbolic of the class.

or printing devices drastically improved thematic cartography features. This led to an explosive growth in visualization methods and techniques, in parallel with emerging theories based on perception and cognitive aspects of graphic elements.

We will now appreciate the way these different elements have indirectly contributed in serving the cartographic representation of molecular data.

MAP DESIGN

Like EDA often resorts to the visualization of graphics to efficiently investigate important data sets, geographic maps are exploiting human cognition features which are recognized to be essentially sensitive to spatial processes (Wood, 1994). Thus maps are particularly well suited to stimulate creative thinking by generating mental imagery in the analyst's brain, depending on their personal culture (Eco, 1985) and educational background.

However, while several sophisticated developments have led to the elaboration of advanced interactive geographic visualization tools (see chapter 5), one could believe that the old-fashioned cartography has become totally out of interest. Wood (1994) has listed and well argued numerous points to make us still consider static maps as valuable tools, main aspects being that cartography allows - as early mentioned in this chapter - to fix, communicate and analyze relevant views. Communication aspects will be dealt within the next section with regard to this genetic data context, and this will make guidelines for molecular data cartography implicitly emerge.

Applying cartographic rules

Because maps are synoptic[1], careful design must be used to ensure that the information is conveyed effectively. Indeed, any map is a signal and is emitted by a designer who is working according to a point of view. This signal has two faces or two sides (see figure 6.2 on page 83) as it is composed on the one hand by a sign that *shows* information, and on the other hand by a concept, that is a *meaning* to be interpreted (Saussure, 1995; Hussy, 1998). This is important because it demonstrates that if a signal is approximately built it will lead to a number of different interpretations and not reach its objective at all.

To transmit his message and to build his signal, the designer has recourse to signs (shapes, colors, etc.) constituting *a language* which allows to communicate information, and possibly to produce knowledge. This language is *the graphics* which is a system of signs allowing to transcribe ordering differences or proportionability relationships between qualitative (**haplogroup** membership for instance) or quantitative data (heterozygosity).

Moreover, geographic information and accompanying elements (for instance a scale, an author name, a framework, etc.) are staged in order to facilitate the signal

---

1. In the sense of «pertaining to, or affording an overall view».

efficiency (Bertin, 1983). This means that whatever spatial analysis is computed, the creation of a map has to follow a publishing process in order to improve communication through cartographic content. Each cartographic document is defined and structured by usual rules and constraints that lead to the correct construction of a map.

SIGNAL or MAP

Sign	To show
Concept	To interprete
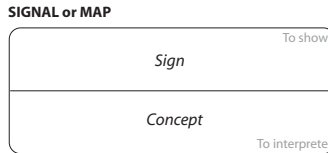
Fig 6.2. The two sides of a map. A simplified representation of any signal meaning unity. Sources : de Saussure (1912), Hussy (1998).

There are two sets of cartographic rules when elaborating a map : on the one hand, the semiology of graphics, and on the other, constraints concerning the «dressing» of a map, what we have called «accompanying elements» in the above paragraph (Ertz *et al.*, 2001). The first set is about methods, rules and tools developed to efficiently represent information on maps; it will be detailed in the section dedicated to the semiology of graphics applied to genetic data. The second set of rules is concerning all elements which have to be displayed together with an edited cartographic document. The role of the map dressing can be compared to the role of the metadata in the field of databases. The reader of a map should have the possibility to read a title, a legend, to identify the author and the data sources, to recognize a scale and the orientation, to know when the document was produced, and to read in a commentary a synthetic description and analysis of what the author wants to show.

The main objective of cartographic rules is to allow an optimal transmission of the information - or of the message - the author wants to convey. A well conceived cartographic representation permits to clearly understand the objective of the map, to quickly grasp its contents and to understand it. The application of both semiology of graphics and map dressing helps composing well presented and organized maps and above all avoids interpretation errors occurring.

Semiology of graphics applied to the representation of spatial genetic data

The graphic language makes use of elementary elements (like words) allowing to construct the map, that is to visually transcribe information consisting of qualitative or quantitative data, and to express existing relationships between them. These elements are *visual variables*, each of them offering specific visual differentiation possibilities and providing determined perceptual properties. Efficiency and relevance of a graphical representation are necessarily depending on the adequate choice of visual variable between *position*, *size*, *value*, *resolution* or *grain*, *color*, *orientation* and *shape* (figure 6.3 on page 84).

Position, *size* and *value* are «variables of the image». They are constructing change-able visibility zones and make forms appear. Among them *size* and *value* are creating a visual ordering or hierarchy. In opposition, *shape*, *orientation*, *color* and *grain* are «separation variables» which are used to build homogeneous visibility zones, «flat», without relief whose only goal is to separate elements.
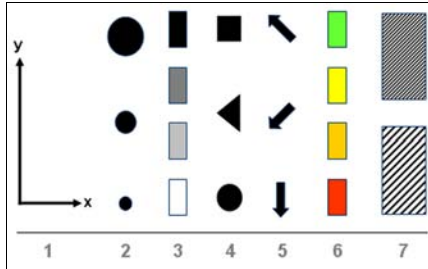
Fig 6.3. Visual variables, the words of the graphic language. 1. Position 2. Size 3. Value 4. Shape 5. Orientation 6. Color 7. Resolution or grain. Sources : Bertin, 1983; MacEachren, 1995; Rappo, 1996.

In order to create efficient maps with the help of visual variables, there are several important *readability* notions to know.

Visual variables and perceptive thresholds

According to the different visual variables, various corresponding *perception thresholds* are existing. The most important is the length of a variable which is the number of sensible levels a visual variable is able to support; in other words the more possibilities a visual variable has to vary, the longer it is. Shape is the longest visual variable, position is the second one. On maps showing quantitative data, an essential element to examine is the value mainly because its variation is dominating any other interfering visual variable : see for instance figure 6.7 on page 92 in which the variation in value of two distinct hues is perceived before the different sizes of the circles. The length of value is depending on the number of classes which have been defined to show a phenomenon on a map (for instance, in figure 6.3, the length of the value variable is 4). Here is a synthesis of some valuable characteristics to know related to the length of value :

- Value is related to the *luminosity* notion in theory of colors;
- the number of perceptive thresholds is depending on the value of the background of the map and on the size of graphic elements. Thus the ideal number of thresholds is a maximum of 6;
- the variation in value is *ordered* and it is essential to use it this way. Otherwise a semantic problem is engendered;
- the white color shouldn't be used unless it indicates missing data.

In this semiology of graphics application to genetic data, color is the second visual variable on which our attention will be focused. It is by far the most difficult to use because of its complexity and the numerous aspects color is indirectly influencing.

A few important points about color are :

- Color has three dimensions and can be defined by a) *luminosity* (also named *brightness* or *lightness*) b) hue (wavelength) c) saturation (pureness);
- color is highly selective: it means that it enables the reader to immediately isolate the correspondences belonging to the same category of a given variable;
- distinctions between colors are more easily perceived in reds and purples, and less perceptible in yellows and greens;
- color variation is not adapted to communicate ordered information : a unique hue is sufficient in this case. Nevertheless, two colors can be used when a distribution is centred on zero (a hue for negative values and another one for positive values), or when a specific differentiation has to be emphasized (see explanations on page 89 and figure 6.7 on page 92);
- when a series of different colors is used (greens for positive and reds for negative data for instance), brightness is perceived first and hue is perceived afterwards[1];
- the efficiency of color is decreasing in proportion with the surface of the object on which it is represented;
- main advantages of color are that a) it benefits from a strong psychological attractiveness and b) it is easily memorizable. Moreover color can be associated with cultural meanings like green for «yes» (or positive) and red for «no» (or negative) according to the traffic light system convention.
- main problems with colors are a) possible perception anomalies (color blindness) of readers and b) unfortunately persisting high paper-diffusion costs.

The notion of efficiency is practical to conclude on visual variables. A considered map is an *image*, that is a relevant visual element which can be perceived within a minimum moment of vision. The shorter a perception time is required to transmit an information, the more efficient a map. The efficiency of an image is also related to the number of information components it is showing. In fact, it is not possible to build a unique map with more than 3 variables. Then is it necessary to build several separated maps to insure a minimum number of simultaneous perception moments and an optimal efficiency.

Qualifying the talk about color and perception

Mastering color and the way given hues can be used or not in thematic cartography is a very difficult matter. It is therefore useful to propose general common sense rules, but it is definitively not adapted to define advanced guidelines going too much into details. First because color is very personal in the sense that the perception process is related to a physiological functioning which may vary among people. I already mentioned perception anomalies like color blindness for instance which is a major one, but given the complexity of the ocular system, minor variations are inevitably existing about the way different wavelengths are perceived by photoreceptor cells in different humans. For instance, some colors with wavelengths situated at the limits of the visible (either around 440 nanometers for violet or 740 nm for reds, see figure figure 6.4 on page 86) are likely not to be perceived at all by a given number of people, or to be perceived differently from what was expected by the author of a map (MacEachren, 1995).

These are physical arguments to which subjective ones have to be added. Indeed, culture is much influencing the way colors are perceived, distinguished or named,

---

1. See Brewer (1999) for detailed guidelines about color use for data representation.

and also the meaning which is attributed to them. This is important about color discrimination which is essential in (thematic) cartography.
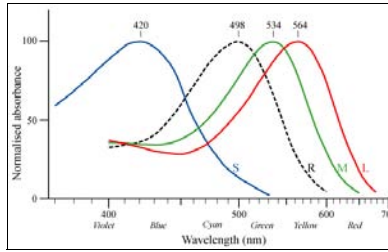
Fig 6.4. Spectral absorption curves of the short (S), medium (M) and long (L) wavelength pigments in human cone and rod (R) cells. Source : Bowmaker and Dartnall, 1980.

The discriminatory power of color is «truly astronomical» (Luria *et al.*, 1986) and the Optical Society of America even classifies a range of 7.5 to 10 million of colors which can be discriminated (Eco, 1985), but as soon as the number of colors being simultaneously compared - or used together - is high, the number of colors which can be discriminated is rapidly dropping (MacEachren, 1995). In a test, Luria *et al.* (1986) are showing that there is a 98% of correct discrimination among 10 colors which falls to 72% when 17 colors are presented. Umberto Eco (1985) is quoting the Farnsworth-Munsell test, which is including 100 hues, to demonstrate on the one hand that the discrimination rate is unsatisfactory for 68% of the population (without color defectives!) which is making between 20 and 100 errors in rear-ranging hues on a continuous gradation scale. But he also uses this test to highlight the fact that the majority of subjects do not have linguistic means with which to categorize those hues. Maerz and Paul (quoted by Eco, 1985) wrote a «Dictionary of color» containing more than 3'000 color names but of which only 8 are commonly used in current english. This is also a number of 3'000 different colors which are recognized and named by New Zealand Maori according to David Katz (quoted by Eco, 1985). Where culture is joining those statistical observations is that ranges of hues for a given culture can be considered a single relevant unit for another one. For instance «our» red and orange are considered a unique color by Hindus, and «our» blue is segmented in «goluboj» and «sinij» by Russians (Eco, 1985). However and despite those apparent discrepancies, it was shown possible to elaborate (impose) international conventions overlapping cultural specificities, to regulate the traffic for instance. This model is often applied to cartography when positive values are represented in green (indicating a free way) and negative ones in red (inciting to stop), as a metaphor of the traffic lights functioning. This method was applied for one of the data categories we have to represent on maps (see figure 6.6 on page 88).

These elements do highlight the facts that make it impossible to offer a strict and rational method to apply when mapping genetic data (and most of other data types). In spite of this remark, there is no doubt that *color* remains an efficient - and convenient - visual variable to use judiciously in conjunction first with the *value* and with the *size* of symbols representing geographic objects.
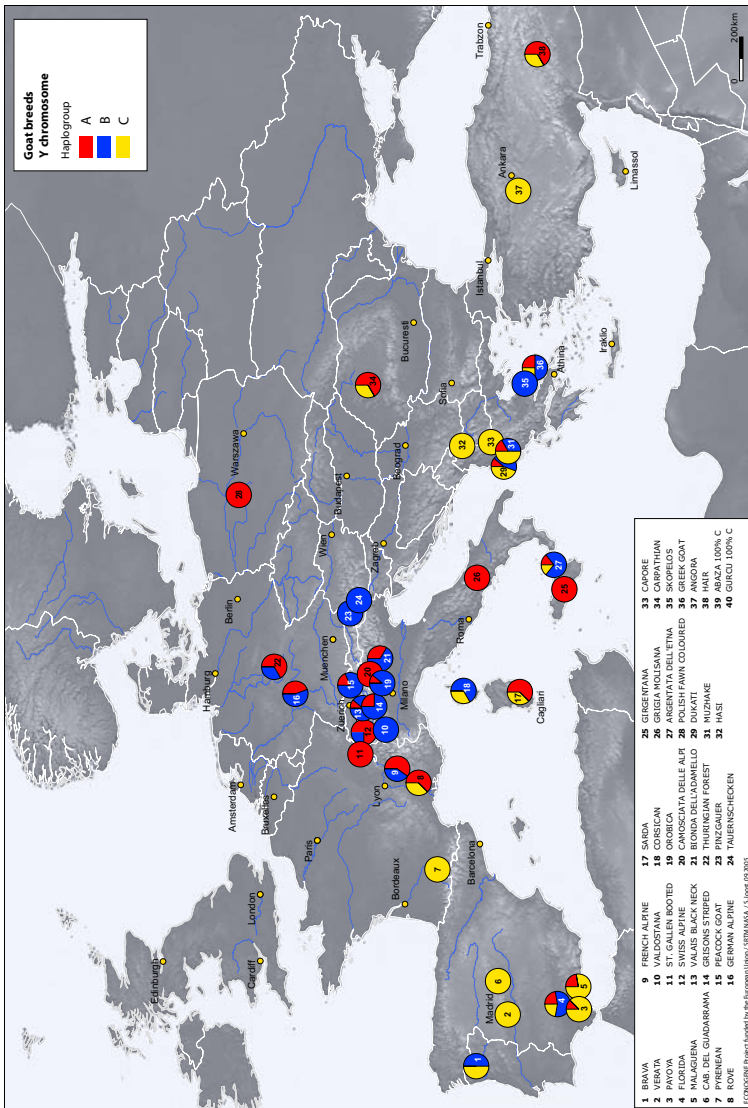
Fig 6.5. Map of Y chromosome haplogroups in goat breeds. This is a cartographic version of figure 5.3 on page 61 which focused on genetic data only, without considering the geographic context (see section dedicated to geographic context on page 90). Those haplogroups are named A3 (A in red), A4 (B in blue) and B (C in yellow) by Lenstra (2005) and shown on a NeighborNet graph of Nei standard genetic distances (see appendix 7). Most of turkish breeds have haplogroup C, but one also has A : this suggests that both were introduced in Europe, but have now variable frequencies in different breeds (Lenstra, 2005). In central and Northern Europe, A and B are predominant, confirming a common origin as suggested by the genetic distances. Haplogroups A and B are more common in Italy than in other Mediterranean breed what may indicate an exchange of paternal lineages across the Italian peninsula (Lenstra, 2005). Also refer to explanations provided on page 58 about information provided by mtDNA and Y chromosome.

87
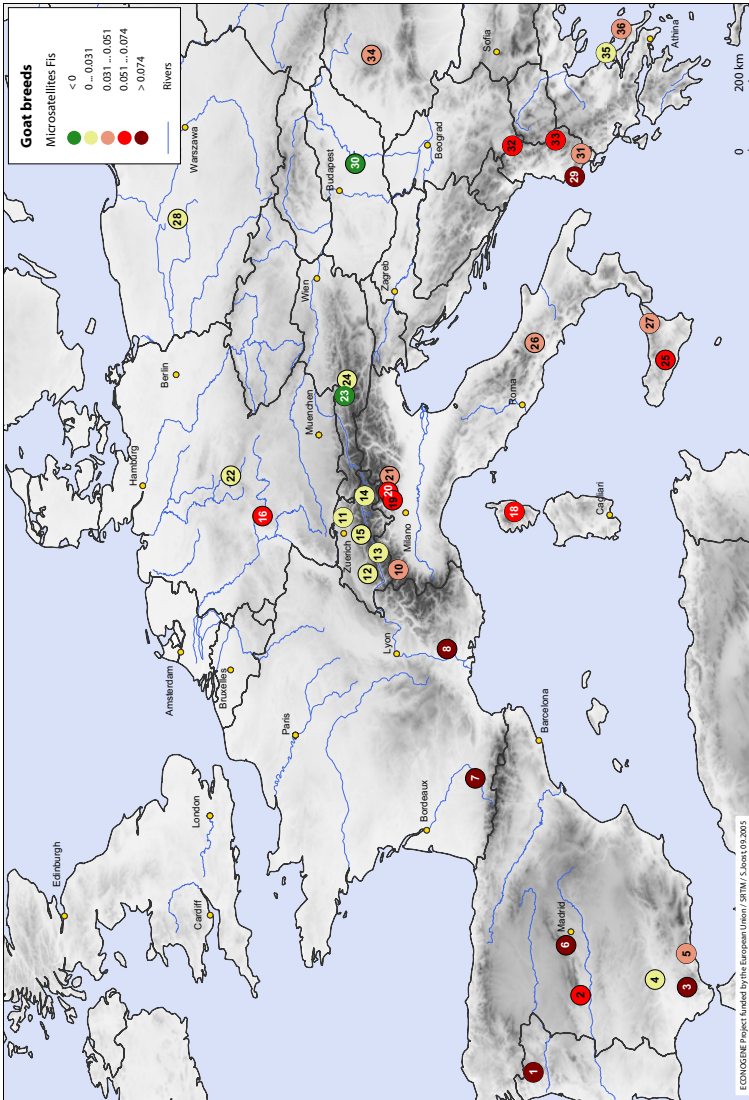
Fig 6.6. Map of inbreeding coefficient in goat breeds. A positive FIS means that the sampled breed is inbred (deficiency of heterozygotes) and a negative FIS means that the breed is outbred (has an excess of heterozygotes). See F-Statistics on page 47. This figure only shows the Western part of the Econogene study area as microsatellite FIS was not available for the other breeds. Moreover, data were not available also for the French Alpine, and the Sarda. The topographical background is different from the other maps presented in this chapter : valleys are represented in light and mountains in dark greys. This makes valleys better appear, and so is it for the hydrographic network. But topography in low and flat areas is less discernible. Breeds can be identified thanks to figure 5.1 on page 58. Pinzgauer and Hungarian Native are the only outbred breeds. The equilibrium in swiss goats is clearly apparent, possibly being the result of a small dynamic market favouring exchanges.

REPRESENTING ECONOGENE GENETIC DATA

In the context of the Econogene project, we are confronted to several kinds of molecular variables : a) quantitative variables composed of positive values which are showing a diversity indication, b) quantitative variables composed of negative and positive values which are representing either the intensity of a relationship (correlation coefficients on principal components) or inbreeding/outbreeding coefficients, and c) categorical variables which are showing a membership to particular **haplogroups**. For each of them specific color schemes were chosen.

For the first category, the representation of genetic data was made in order to distinguish at best *diversity* values variation. Reaching this goal is a subtle balance between the choice of the most adapted discretization method and the most adapted number of classes on one side, and the most adapted choice of colors and hue on the other side, to fit crucial known criteria about visual perception (MacEachren, 1995). The chosen discretization method was «natural breaks» for all maps, completed with a few manual adjustment notably in the case of proportional circles. Natural breaks[1] has the main advantage to conserve spatial relationships what is a determining criteria in our case (homogeneous classes are generated) (Faucher, 2002). The main color scheme was chosen in order to show as much variation as possible within a considered molecular variable. According to cartographic rules mentioned above, a variable which is expressing a progression from a minimum positive to a maximum positive value should be represented with a single hue in conjunction with a gradation in brightness (clear grey to black for instance). Nevertheless I had recourse to a color scheme composed of three hues : a gradation of a cold hue from dark to light blue to represent low values, an intermediate and neutral hue to represent mean values, and a hot hue from light to dark red to represent high diversity value (see for instance figure 6.9 on page 94). Playing with cold and hot colors has the advantage to extend the perception range by adding a hue and so to refine the appraisal of the spatial distribution of the variable variation.

Data of the second category present quantitative variables with negative and positive values. For correlation coefficients on principal components, negative values are represented in cold colors (blue) and positive ones in hot colors (red). In-between, an intermediary neutral hue (light yellow) is used to separate negative and positive values (see figure 6.11 on page 98). It also happens that factorial scores are positives only, and in this case the same color scheme was applied. The only tinge with the previous paragraph is that cold colors can represent negative or positive values according to the data to represent.

Always in this second category, negative values of the F_{IS} inbreeding coefficient mean that populations are outbred. Greens have been chosen to represent them as the diversity aspect is positive in our context. The values around zero represent populations at equilibrium or near it, and light greens to light neutral yellow have been chosen. Finally, inbred populations are represented in red to emphasize the negative aspects of inbreeding (see figure 6.6 on page 88). This green/red opposition was chosen because of the meaning attributed to those colors by an international convention (see previous section).

---

1. Also called «observed thresholds» or «Jenks method». For further details please read Smith, R.M. (1986) Comparing traditional methods for selecting class interval on choropleth maps, The Professional Geographer, 38:1, p.62.

The last type of data is categorical information. **Haplogroups** have been represented according to standard pies used in population genetics. The choice of colors aimed at reaching a maximum discriminating power according to the number of categories (see figure 6.5 on page 87).

Shape and size of the symbols

The circle is the only symbol chosen to represent objects on the maps presented in this chapter. It is the «natural» extension of a geographic object defined by a pair of geographic coordinates representing a farm or a breed.

For all presented maps, a particular attention was paid to the choice of an adapted size of the symbols in order to ensure a maximum readability of the maps and their precise identification. The main issue is to avoid any symbols overlap in area were a high density of breeds populations is observed.

Proportional symbols were used in bivariate maps. Figure 6.9 on page 94 simultaneously represents microsatellite heterozygosity and mean number of alleles to assess the way the number of alleles is influencing genetic diversity. To calculate a correlation coefficient is a way to get an answer, but the map is able to provide a sufficient indication while showing the geographic context. In this case, the cartographic representation is globally confirming the intuition of the more alleles the higher the diversity, with exceptions like the Kymi or the Romanian Tsigaia breeds whose mean number of alleles is higher than expected. Figure 6.12 on page 99 proposes another bivariate map : microsatellite heterozygosity is injected in proportional symbols in order to provide an interpretation key to factorial scores expressed by hue. And indeed, it is possible to observe that breeds with high diversity globally have a high correlation coefficient on the first component of the analysis.

Apart from the semiology of graphics, let us now comprehend how spatial analysis and graphic edition are also adequate to supply additional means in order to improve the depiction of molecular data on maps.

Providing a geographic context

Most of spatial representations realized so far in geographic genetics (for instance Epperson, 1993, p.191; Petit *et al.*, 2001, p.306) have recourse to a neutral geographic space, strongly suggesting that only the general location and an approximative distance between symbols is important. Of course, this depends on the goal the map is expected to reach, as one may only wants to focus on a simple function like to compare **haplotypes** and showing the relationship between diversity and geographic location (Bruford *et al.*, 2003, p.907). But often are cartographic representations unintentionally limited in my opinion.

In addition to their role of spatial index[1] and of communication tool, geographic representations of genetic data are produced because spatial processes are supposed to be explanatory. Then it seems consistent to use available contextual spatial objects at best, and not being satisfied with a mere longitude/latitude descrip-

---

1. A geographic key to access information.

tion of analyzed individuals. Nowadays, as representation technologies henceforth do allow it, giving a concrete expression of landscape (showing its main and explicit forms) is invaluable to improve cartographic representation of genetic data, with the constant concern of keeping a high level of readability. This permits to anchor a phenomenon in the landscape and helps analysts to understand a spatial distribution of data and to produce new working hypotheses. Contextual objects may be on the one hand relief, forests, rivers, etc., that is to say natural landscape components, and on the other hand anthropic objects like roads, railways, etc. Depending on the working scale, both are likely to play the role of barrier and supply explanatory elements when analyzing their respective position with that of the animals. These points of reference are helpful to locate and inlay observed objects into the geographic space. This has the considerable advantage to reduce the analyst's first intuitive intellectual effort mobilized to locate an object before being able to initiate the visual thinking process to make research hypothesis emerge (Wood, 1994; MacEachren *et al.*, 2001). To illustrate this point, figure 6.7 on page 92 and figure 6.8 on page 93 are displaying the same genetic information. Only the geographic context is removed in figure 6.8 excepted countries frontiers which are constituting minimal landmarks. We can consider two reading levels, that is firstly the efficiency with which the genetic information is communicated to the reader (where is which value ? does it constitute patterns ?), and secondly the way this information can be analyzed and interpreted. Figure 6.8 is sufficient and more effective - because of the neutral background - for the first reading level, but is weak and even unable to provide clues for analyze and interpretation, unless resorting to previously acquired knowledge of any kind related to approximate location(s) where remarkable values of analyzed objects are observed. On the contrary, and despite its relative «deficient efficiency» in transmitting genetic information, a map like in figure 6.7 offers much more immediate analytical perspectives. These ones are limited but sufficient to situate the examined information (breeds) in its topographical context. With the exception of figure 6.6 on page 88 (read comment), relief is represented with grey shades, from dark for low areas to light for higher altitudes. This combination makes mountains chains (the Alps, the Pyrenees, the Carpathians) and also high regions perfectly stand out (Turkey, north west of Madrid), despite the fact that elements are represented on a small cartographic scale[1]. Moreover, the choice of these rather neutral grey shades permits to preserve an acceptable efficiency in showing the main information. A colored topographical background would be disastrous for readability. In addition to a global elevation indication, topography may also highlight isolation situations of certain breeds, in conjunction with rivers. But the more convincing argument in favour of the geographic contextual map is probably that it is the only way to know where is which studied individual, and this as accurately as possible thanks to the combination of different kinds of complementary informative geographic elements (administrative boundaries, hydrography, topography, localities, etc.), with the constant concern of keeping the whole distinctly readable.

---

1. Scale issues are not tackled in this research as mentioned on page 39. Data representation is also implied in scale issues and the usefulness of GIS also lays in features allowing to access to more detailed information in order to provide more accurate representations.
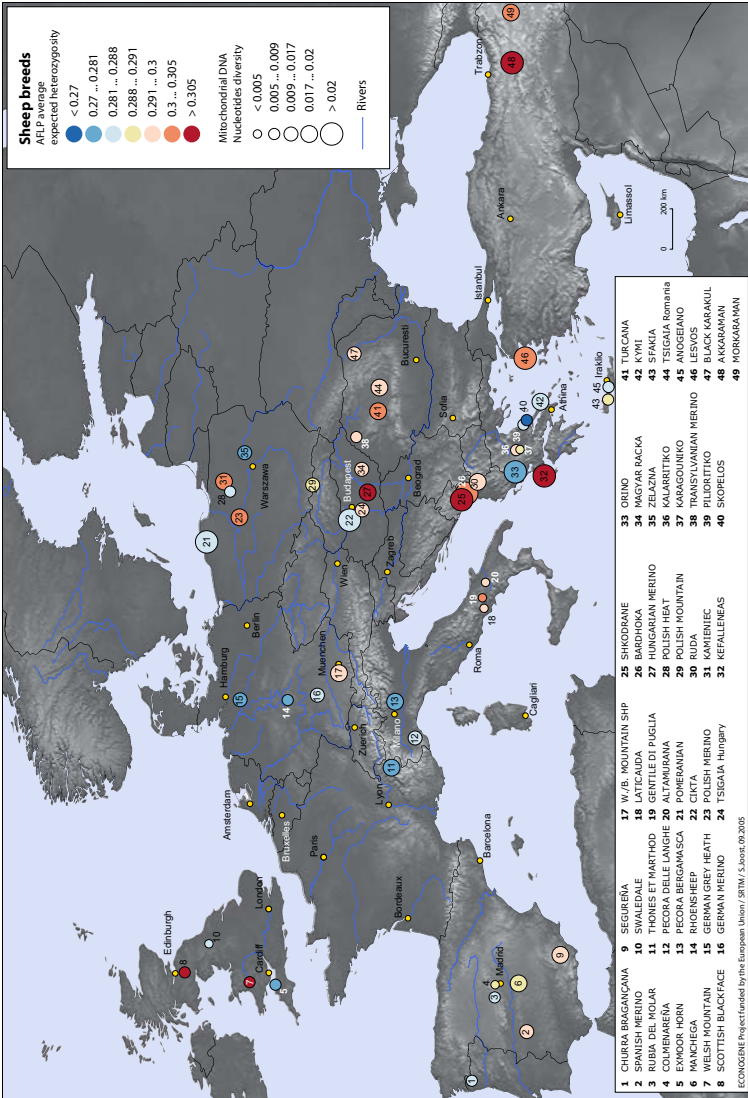
Fig 6.7. Map of AFLP average expected heterozygosity and MtDNA *nucleotides* diversity in sheep breeds with a geographic context. A neutral grey shades permits to preserve an acceptable readability and efficiency in showing the main information. Cities and countries limits provide a localization information likely to favour the emergence of indirect information stored in readers memory. Topography is providing direct ecological information likely to make hypotheses to be formulated about one breed and with regard to data shown by its neighbourhood.
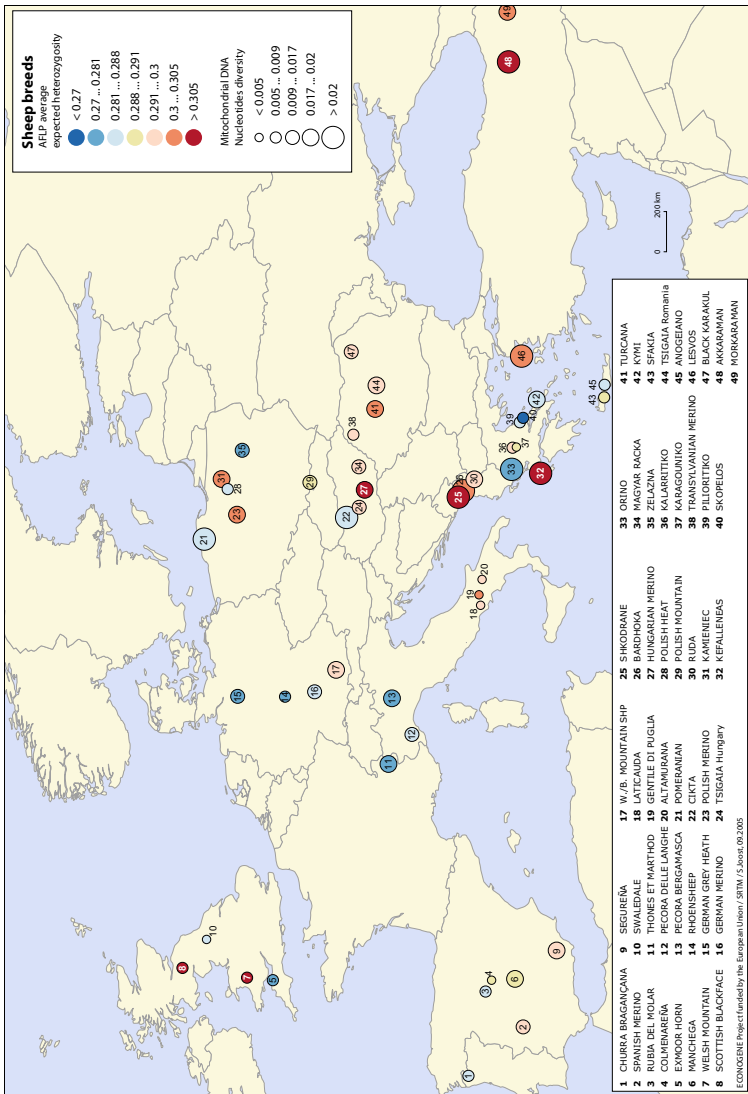
Fig 6.8. Map of AFLP average expected heterozygosity and MtDNA nucleotides diversity in sheep breeds without geographic context. Only countries limits are visible. This «flat» representation possesses the undeniable advantage to very efficiently show the spatial distribution of analysed data and to make possible spatial patterns immediately appear. But it is simultaneously very poor to help generating hypotheses in relationship with simple geographic features.
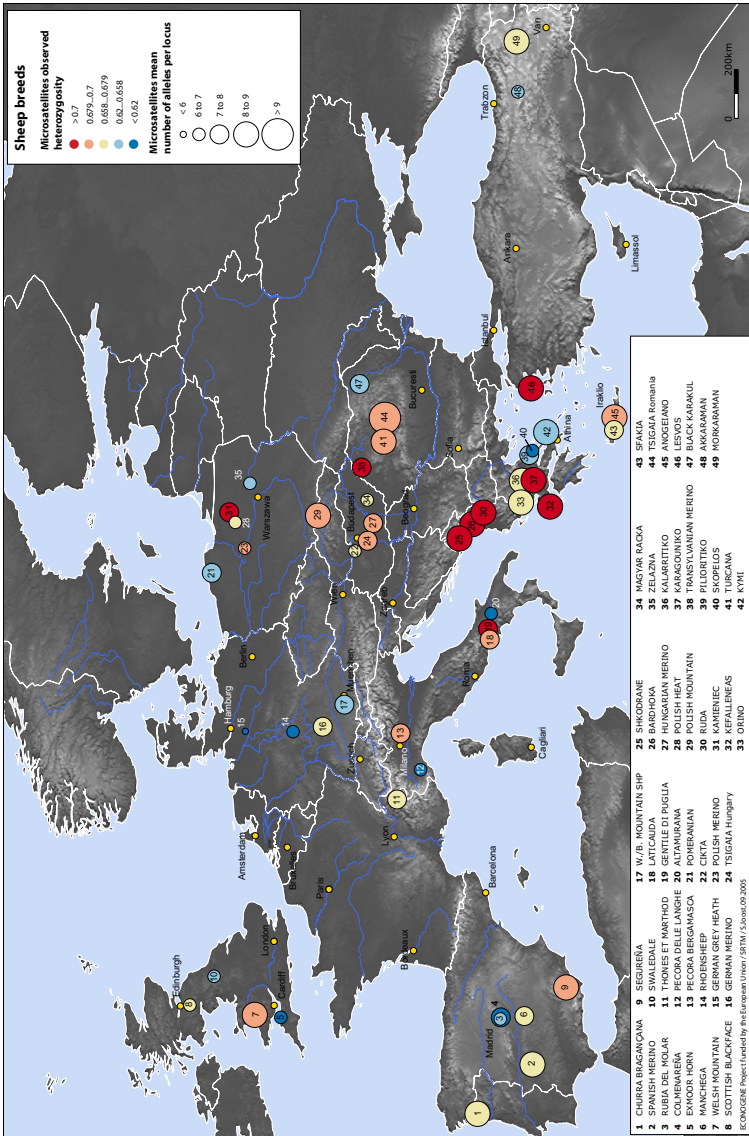
Fig 6.9. Map of microsatellite observed heterozygosity in sheep breeds. Genetic diversity is represented by hue variation and the microsatellite mean number of alleles per *locus* is represented by mean of variation of the perimeter of the proportional symbols. Graphical tricks are used to represent at best centroids which are close to each other : those ones can be superimposed (3 and 4 near Madrid, or 39 and 40 near Athens) while preserving the readability. This map is noticeably showing the amazing behaviour of the Skopelos breed (40) with regard to higher diversity values expected in Eastern regions. This apparent vulnerability is confirmed by a FAO animal production and health paper (Brooke and Ryder, 1979) but contradicted («not endangered breed») by the up-to-date School of Veterinary Medicine (Hannover) database (http://www.tiho-hannover.de) (7.12.2005).

Regionalization[1]

In «The history and Geography of Human Genes», Cavalli-Sforza, Menozzi and Piazza (Cavalli-Sforza *et al.*, 1994) resorted to regionalization to elaborate most of the maps of what the Time newspaper once called «the first genetic atlas of the world». More exactly, they had recourse to smoothing, that is to fit a surface which is close to sampling points without forcing it to go through them. This choice was made to avoid the problem of possible false outliers interfering in the spatial distribution of gene frequencies trend surfaces (typically breed 27 on figure 6.12 on page 99). The goal was to make thematic mapping efficient to visualize general trends in genetic frequency patterns, without displaying «parasites» (which can be real outliers, too).

To make thematic mapping efficient to visualize genetic data, regionalization can be used as an *artefact* to enforce the visual impact of the spatial distribution of diversity measures, and to create a continuity that facilitates its rendering. Brodlie (1994) mentions this approach as a possible sequence of processes, or «visualization pipeline» : a) interpolate scattered data on to a grid, b) generate contour map from grid, c) render contour map. This technique was used in several maps and turns out to be really effective to emphasize gradients in the variation of genetic diversity, provided that a warning is explicitly formulated about the sense to attribute to interpolation (see figure 6.11 on page 98 and figure 6.12 on page 99). Indeed, an important point about maps having recourse to interpolation is that the staged observed phenomena are analyzed through a given sampling made according to chosen criteria (see chapter 4, page 36). In front of a map, the reader is obviously not confronted to reality; but it is spontaneously difficult to straight-forward integrate this necessary distance or reservation as a filter when elements are concretely represented on a map. This is also true for maps showing symbols only, but the absence of continuous information makes it easier to realize.

Regionalization is closely related to geostatistics which first applied theories of stochastic processes and statistical inferences to geographic phenomena. Geostatistics were initially used in geosciences (geology and particularly petroleum research). Its functioning is based on the concept of «scales of spatial variation». Spatially independent data show the same variability regardless of the location of data points. However, spatial data in most cases are not spatially independent and data values which are close to each other are revealing less variability than data values which are farther away from each other. This is a principle of autocorrelation (variography) which is usually exploited to predict values in places where sampling was not carried out. The maps presented here (figures 6.11 and 6.12 or 6.13 and 6.14) have not recourse to interpolation in order to predict genetic diversity values, but to emphasize visual patterns, to make it easier to visualize a possible scenario of spatial continuous information. Real and trustable information is only contained within breeds centroids or breeds farms. Moreover, by creating a layer of continuous information, this representation technique allows the superimposition of another variable to be compared as it is illustrated by figures 6.13 and 6.14 where a variable showing sustainability of farms activity has been overlaid to the interpolated genetic diversity.

---

1. Regionalization is here used in the sense of a tendency to form regions on the basis of similar information contained by neighboring points.

The different interpolation examples presented in this chapter where obtained by using the normal kriging method. This is a gaussian process regression technique used in geostatistics named after its inventor, D.G. Krige[1]. Kriging was then formalized - and first named - by Georges Matheron[2] in the early 1960s. This method is often used to point out regionalization despite the fact this is only one of many existing ways of interpolating information. Other methods are Natural Neighbors, inverse Distance Weighting (IDW), Triangular Irregular Network (TIN), etc.

To limit false interpretation or abusive representation in spite of the above reservations, masks have been created to hide areas where no breeds were sampled. This makes neutral areas appear to cut the continuous interpolated information. This constraint is mainly due to France where the sampling was carried out in the Alps, in the Southern Alps, and in the Pyrenees only.

Cartography of Principal Component Analysis results

Representations of the genetic relationships among populations may be obtained by using multivariate procedures. Among multidimensional analysis methods, principal component analysis (PCA) offers a simple and powerful mode to analyze sets of populations genes frequency data. This technique condenses the information from several alleles and loci into a few synthetic variables. Moreover, it was observed that the first splits in tree procedures (when inferring phylogenies and constructing evolutionary trees) habitually correspond to the separation of populations generated by the first components of multivariate procedures (Cavalli-Sforza *et al.*, 1994; Moazami-Goudarzia and Laloë, 2002).

Data processed by PCA are genetic distances. A genetic distance is a way of measuring the amount of evolutionary divergence in two separated populations of a species by counting the number of allelic substitutions per *locus* that have cropped up in each population. There are several ways of calculating genetic distances, among which the one proposed by Reynolds, Weir and Cockerham (1983) which is used in both PCA results presented in figures 6.11 and 6.12. It assumes that all differences between populations arise from *genetic drift* only. This distance measure is based on the coancestry coefficient θ[3]. Three estimators of the distance $D = -\ln(1 - \theta)$ are constructed for multiallelic, multilocus data. In such a drift situation, from which *mutation* is excluded, this weighted estimator appears to be an appropriate measure of distance (Reynolds *et al.*, 1983).

Microsatellite data were produced by Christina Peter[4] who also ran the PCA analysis. This was made on the basis of data out of thirty-one microsatellite markers covering 22 chromosomes, including ten markers recommended by the FAO[5], characterizing the genetic variability of 57 european (most of which Mediterranean) and middle Eastern sheep breeds from 15 countries. Those results were used in order to produce the map shown in figure 6.11, page 98. Factorial scores

1. Krige, D.G. (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand, J. of Chem., Metal. and Mining Soc. of South Africa, Vol. 52, No. 6, pp. 119-139.

2. Matheron, G. (1962) Traité de géostatistique appliquée. Tome 1, Editions Technip, Paris.

3. The coancestry coefficient is the probability fAB that two homologous genes, one from individual A and the other from individual B, are identical by decent or in other words descend from the same ancestral gene. The complementary probability, 1-fAB, is the probability that these two genes come from unrelated ancestors.

4. Department of Animal Breeding and Genetics, Justus-Liebig-University of Giessen, Germany.

5. http://dad.fao.org/en/refer/library/guidelin/marker/pdf (16.11.2005)

on the second component are punctually represented and extrapolated according to the considerations proposed in the previous section. This component was chosen because it illustrates how cartography may be precious to detect distinct behavior of given breeds. The map shows a clear separation of the German Grey Heath breed which is a short-tailed breed thought to have descended directly from the european Mouflon. It is an autochthonous breed of the Lüneburg Heath in Northern Germany. Early reports about this breed have emphasized the excellent adaptation to feeding upon the sparse heather (Peter *et al.* and references therein, submitted).

As for the first component (here not shown) which is explaining 69% of the variability, it revealed a clear distinction between Western European breeds on the one hand and south Eastern European and middle Eastern breeds on the other hand. Migration during the Neolithic demic population expansion and subsequent adaptation to the environment could have caused this structure (Peter, 2005).

The cartography of PCA factorial contributions was also applied to AFLP markers produced and processed by Riccardo Negrini[1] (figure 6.12 on page 99). The first principal component accounts for the 58% of the total variance. It clearly separates German Grey Heath and the British Swaledale from the other breeds, and plots the latter separating a Western-Central European group from Eastern breeds (red circle in figure 6.10), with a few exceptions (Negrini, 2004). This is clearly apparent on both plot of two first principal components (figure 6.10) and map of the first factor in figure 6.12 on page 99.

Fig 6.10. Sheep breeds : plot of the two first principal components of a PCA analysis based on Reynolds distances between sheep breeds populations, using 93 AFLP polymorphisms. Source : Riccardo Negrini, Piacenza (Negrini, 2004).

It confirms observations made on the basis of microsatellite markers for the German Grey Heath as an original breed. As for the Swaledale[2], it is a breed whose origin almost certainly emerged from the genetic group of horned sheep from which also came the Blackface. Anyway, both breeds appear in the same cluster within a microsatellite STRUCTURE analysis (phylogenetic trees, Pritchard *et al.*, 2000) assuming K = 4 classes (together with the british Exmoor Horn and Scottish Blackface) (Peter *et al.*, submitted).

---

1. Istituto di Zootecnica, Facoltà di Agraria, Università Cattolica del Sacro Cuore, Piacenza, Italy.
2. http://www.swaledale-sheep.com (30.11.2005)

Fig 6.11. Sheep breeds - PCA on microsatellite data : cartography of Principal Component Analysis factorial scores. The PCA was run on Reynolds genetic distances based on microsatellites. The interpolated and represented value is the factorial score of each breed on the second component (6.2% of the total variance). This component was chosen because it is revealing a gradient from South East to North West, supporting the hypothesis mentioned several times about the migration of early farmers. It also shows a clear separation of the German Grey Heath (GGH, #15) from all other breeds. This german breed has kept a lot of its nativeness, as crossbreeding with other breeds in former times failed due to strong adaptation to heathland environmental conditions. [As it is not specified in the legend, the grid is showing continuous values of factorial scores ranging from the minimum to the maximum indicated for breed centroids].
Interpolation is not carried out in order to predict values of genetic diversity, but to emphasize visual patterns.
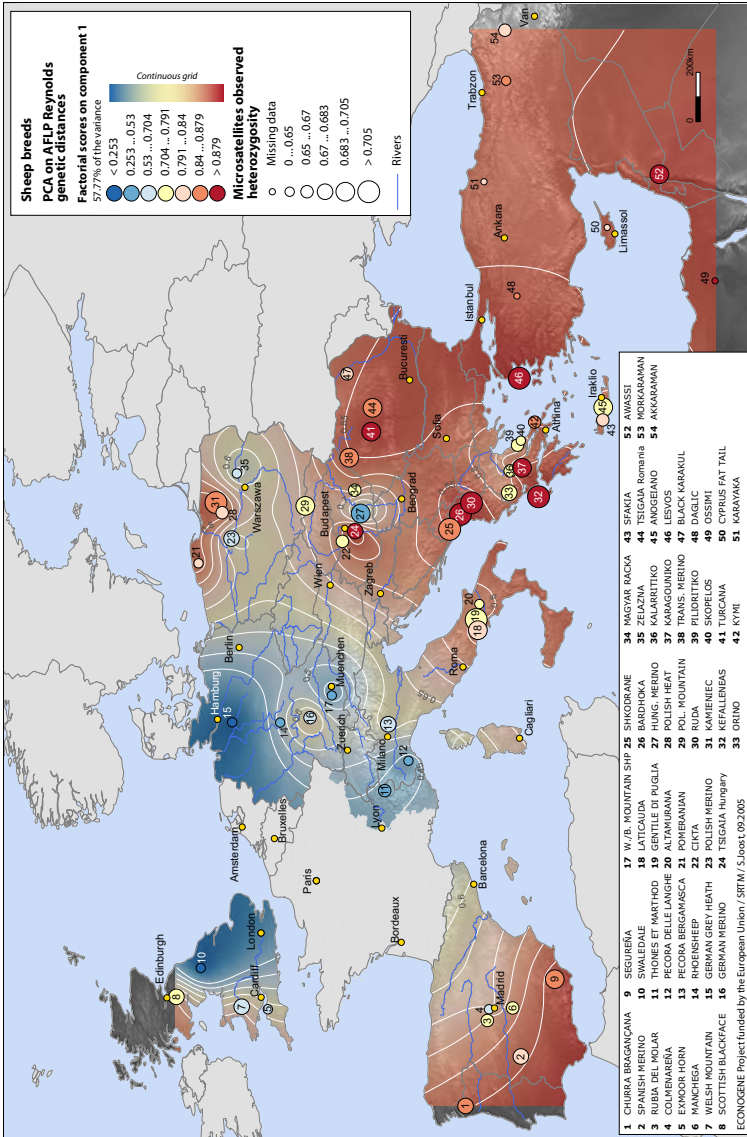
Fig 6.12. Sheep breeds - PCA on AFLP data : cartography of Principal Component Analysis factorial scores. The PCA was run on Reynolds genetic distances based on AFLPs. The interpolated and represented value is the factorial score of each breed on the first component (57.8% of the total variance). The diameter of the symbol is proportional to the microsatellite observed heterozygosity. German Grey Heath and the British Swaledale are leading a Western central european group separating far Western from Eastern breeds. Hungarian Merino seems to behave as an outlier on the Eastern part. [As it is not specified in the legend, the grid is showing continuous factorial scores values ranging from the minimum to the maximum indicated for breed centroids].
Interpolation is not carried out in order to predict values of genetic diversity, but to emphasize visual patterns.

99

Moreover, figure 6.12 on page 99 is another example likely to show the direction of the agriculture expansion and migration towards north Eastern Europe (see description of the hypothesis on page 23) expressed by sheep genetic diversity information. More than a simple progressive loss of genetic diversity towards north east, PCA results are consolidated as they are based on genetic distances and thus show patterns of breeds being genetically nearby. On both figures 6.9 and 6.12, the behavior of the Welsh Mountain and the Scottish Blackface breeds have to be noted as their diversity is amazingly high with regard to the famous agriculture expansion hypothesis, what makes them rather look like most of Eastern breeds (see *OAGBWMT* and *OAGBSBF* in red on figure 6.10 on page 97).

The combination of PCA results cartography and of interpolation as a support to visualization appears to be efficient in order to grasp trends in statistical individuals spatial behavior. It above all allows to raise the impact of contrasts between behaviors - by creating regions - what discrete information (centroids) is not able to show. This may however be deceptive and special attention has to be paid to sampling density and to the adopted representation scale with regard to the proposed interpretation. In the case of figures 6.12 and 6.10, sampling density is low but the explanations are rather general and matched up with different information sources.

Superimposition of information layers

Data interpolation is also used in figure 6.13 on page 102 and in figure 6.14 on page 103 to show two informations of different nature on the same map : over genetic diversity represented by microsatellite heterozygosity are represented farms revealing a probability of sustainable activity. This index was calculated by Marco Bertaglia[1] on the basis of information collected with questionnaires distributed in the farms where Econogene breeds were sampled, and according to general statistics provided by Eurostat[2]. It is an ordered logit model which is taking into account as well demographic, as socio-economic and breeding strategies variables concerning the farms and the surrounding region (Bertaglia, 2004, p.93). The index has been computed for Econogene countries belonging to the EU-15 only as statistics for non-EU countries was not available. Both maps represented on pages 102 and 103 are well illustrating the difficulty to represent information on the farm level as they often are regrouped together in heaps. There is no need in this case as a general representation of the phenomenon is wanted, but when needed an zoomed extraction of a specific area is the only way of clearly showing a situation.

Interpolation of genetic diversity was made according to farm data and this is what explains the high fragmentation of patterns in Switzerland. The swiss farms are not appearing because the index of sustainable activity was not calculated for them[3]. For goats, the south east to north west gradient of genetic diversity emphasizing the direction of agriculture expansion is apparent, despite several differences. On a general reading level, goat breeds show an Eastern high genetic diver-

---

1. Centre for Environmental Policy, Faculty of Life Sciences, Imperial College London.
2. http://epp.eurostat.cec.eu.int (29.11.2005)
3. The index is calculated with data provided by Eurostat and available for countries of the European Union.

sity, an alpine mixing of high and low diversity values, and a Western low genetic diversity mainly due to the Pyrenean and part of Rove animals. On a more detailed level of reading, the main observation is the high diversity revealed by Southern Spain breeds, what makes a Northern Africa migration route scenario emerge. A complementary sampling of Egyptian, Libyan and Maghrebian goat breeds would of course be necessary to check this hypothesis.

About goats farms probability of sustainable activity, trends are mixed up all over Europe, but it is to note the homogeneity of good values for the Rove in the south of France. In a conservation perspective, such a map also allows to identify the debatable case of farms raising the Pyrenean breed which are showing rather good probabilities for going on with their activities though the breed exhibits a really low genetic diversity.

Genetic diversity in sheep is much more patchy. Of course microsatellite data for turkish breeds are missing and prevents us from assessing their values on the Eastern part, but this is the same for goats. In fact, sheep genetic diversity is very heterogeneous in all regions of Europe, for all breeds.

As for the index of sustainable activity in sheep farms, it is very expressive. The first observation is about Great Britain where this index is very low, excepted for one Welsh Mountain farm. This is probably an early effect of the epidemic of foot and mouth disease which took place in the UK in 2001. This case is in opposition with the one mentioned about the Pyrenean goat breed because we are confronted to a low probability for the farms to carry on with their activities while the genetic diversity of the breeds is high (particularly Scottish Black Face and Welsh Mountain). Out of Great Britain, there are 3 poles where a high probability of sustainable activity is observed : first in France (Haute-Savoie and Savoie) around the Thones et Marthod rearing, then in Hungary and Romania where the index is high for all farms, and finally in Crete.

SUMMARY

Commensurate with the elementariness the maps which are accompanying spatial genetics papers are shaped, the role of the cartographic representation of georeferenced molecular data seems of secondary importance in population genetics, although it is really appropriate to communicate information or discoveries between research teams and to the general public.

The minimalist standard way of representing spatial genetic data is considerably improved by applying thematic cartography rules, widely relying on semiology of graphics. These guidelines take advantage of observations related to the functioning of human visual perception. Their use leads to *the production of efficient maps*, whose main purpose is to communicate information as quickly and as precisely as possible. On that basis, the cartographic representation of molecular data takes even more magnitude when adequately complemented by the display of an adapted *geographic context* and with the contribution of techniques of regionalization which must be used with caution.
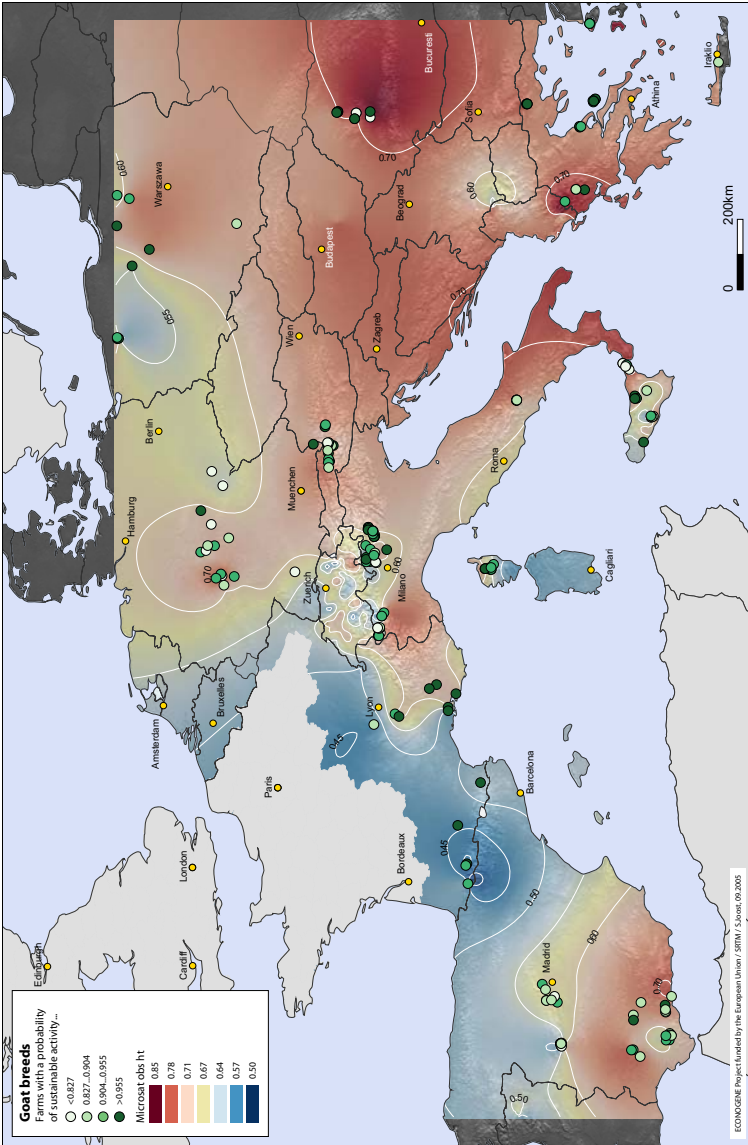
Fig 6.13. Microsatellite heterozygosity in goat breeds together with the probability of sustainable activity of farms. Interpolation of genetic diversity was made according to farm data and this is what explains the high fragmentation of patterns in Switzerland. In a conservation perspective, the homogeneity of good sustainability values for the Rove in the South of France together with a rather high genetic diversity could serve as an example; in opposition, the map also allows to identify the questionable case of farms raising the Pyrenean breed which are showing diverse sustainability indices while the breed exhibits an homogeneous low genetic diversity.
Interpolation is not carried out in order to predict values of genetic diversity, but to emphasize visual patterns.
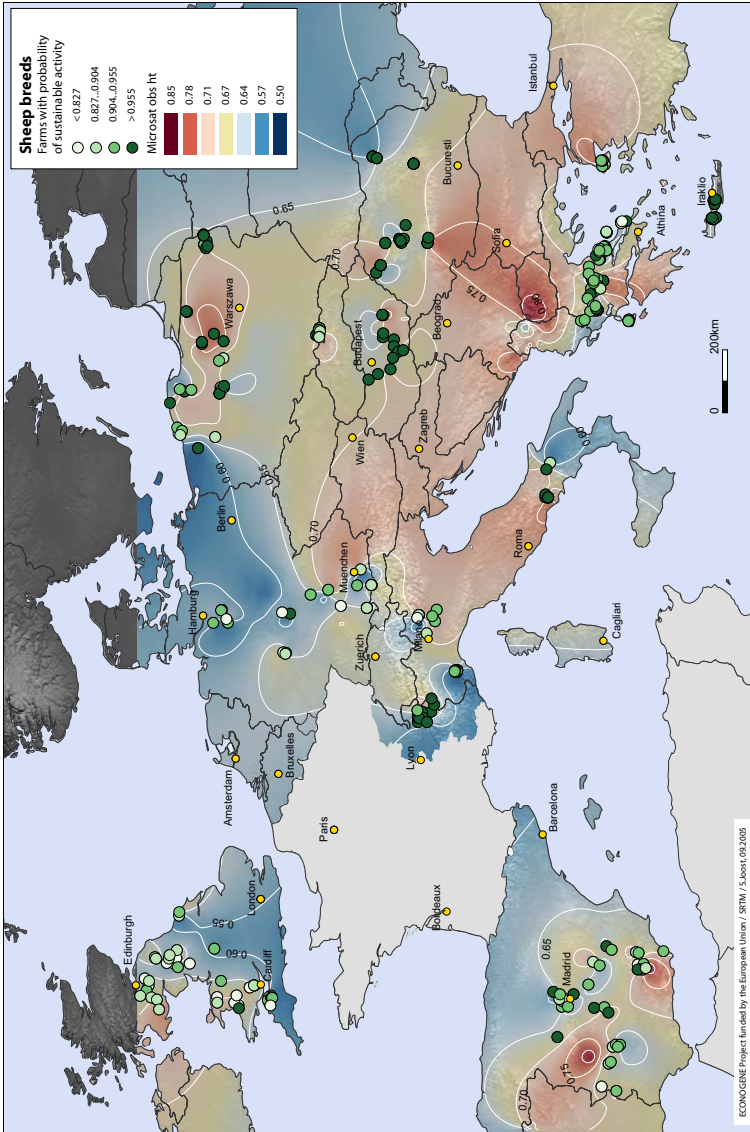
Fig 6.14. Microsatellite heterozygosity in sheep breeds together with farms probability of sustainable activity. This map mainly emphasizes the possible consequences of the epidemic of foot and mouth disease which took place in the UK in 2001 : the sustainability of activities of farmers rearing sheep breeds is really low in Great Britain. Otherwise, three poles are observed where the activity of farms apparently is solidly established : in France (Haute-Savoie and Savoie) around the Thones et Marthod rearing, in Hungary and Romania where the index is high for all farms, and in Crete.

Interpolation is not carried out in order to predict values of genetic diversity, but to emphasize visual patterns.

Spatial Analysis to Detect Signatures of Natural Selection

7

Chapter outline

In the previous chapters, GIScience tools application to molecular data were related to visualization and general analysis of breeds to study their respective spatial behavior. Now we will go back over Darwin's preoccupations quoted on page 11 to exploit GIS features in order to establish whether there are detectable relationships between environmental characteristics and specific regions in the genome. In other words, to detect natural selection signatures within studied genomes.

INTRODUCTION

In this chapter, I propose a Spatial Analysis Method (SAM) to detect, identify and measure the hypothetical sensibility of microsatellite alleles and AFLP markers to the environmental stimulus. This is undertaken from the GIScience angle, with the help of its own tools and of statistical methods.

Michael Goodchild (1996) distinguishes six concepts of spatial analysis, from the simplest function that is to organize data (the map as a spatial index), to more complex ones like spatial dependence or spatial heterogeneity. In order to connect genetic information with geoenvironmental data, that is information characterizing animals organisms with properties of their surroundings, we refer to the spatial coincidence concept (Goodchild, 1996). This consists in associating information levels and to compare them thanks to their common geographical coordinates. In the present case, the spatial coincidence is determined by the inclusion of farms coordinates - carrying molecular information - within cells of grids within which climatic or topographic information are available.
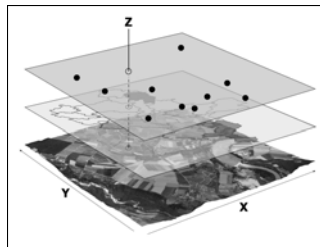
Fig 7.1. Spatial coincidence. (X,Y) coordinates differentiate individuals or populations (see chapter 2) but they may also constitute a common information to share in order to compare and analyse a Z coordinate (genetics, climate, etc.). Source : A. Pointet, modified.

The method was developed independently of any population genetics model, and of any approach developed by population geneticists in order to detect abnormal loci behavior within the genome, potentially involved in adaptation processes. Its usefulness is related to the study of the shaping of molecular variation by natural selection processes, which is a domain likely to improve our understanding of the genetic mechanisms of evolution and to «speed up the discovery of genes that are important for health and human medicine» (Luikart *et al.*, 2003). And regarding applications, ***conservation biology*** researchers are increasingly interested in this kind of approach, as loci confirmed to be under natural selection[1] may be used :

- to help endangered species, and breeds within species, to survive present and future environmental changes by choosing source populations for the translocation of individuals to supplement and rescue declining populations (Luikart *et al.*, 2003);
- to include as criteria also information on the presence and frequency of valuable alleles for adaptive traits to prioritize populations for conservation;
- to identify the provenance of endangered species in the context of illegal trafficking (Luikart *et al.*, 2003).

Apart from conservation genetics, we can also mention that this type of approach is useful for the detection of fraudulent food products. Luikart *et al.* (2003) mention the case of France where differences in the frequencies of ***SNP*** alleles (high F_{ST}) were used to detect the fraudulent use of cheap Holstein milk for producing speciality cheeses exclusively made of expensive mountain-breed milk[2].

The proposed method is not a mere application of GIS and statistical tools, but its design and application is relying on a particular approach of environmental modelling based on the specificities of natural sciences. This thought process is lauding a modelling simplicity whose utilization's grounds and advantages are described as an introduction to the case studies analyses.

But on a general level, the matter discussed in this last chapter tackles the domain of Evolution, and the proposed method is likely to show if GIScience has a role to play in the task of understanding how life is working. The following section is providing some landmarks about the present evolution debate, to position this research, and to highlight some possible stakes it may touch.

EVOLUTION AND NATURAL SELECTION

The analysis of the relationships between any component of the natural environment and animals organisms at the genome level constitutes the core material of this research. It addresses an essential issue at the junction of many sciences, among which biology and GIScience. The assessment of environmental pressure and corresponding genetic adaptation is included in a current universal fundamental debate about living systems and evolution. Discussion is still open about the mechanism of adaptation, but it is generally accepted that living organisms

---

1. Called «adaptive loci».
2. The technique was developed by Maudet and Taberlet who sequenced thousands of ***nucleotides*** from many coding genes to find high FST single nucleotide polymorphisms to differentiate the black Holstein (common cattle breed) from the red French Alpine which is a rare one (Maudet, C. & Taberlet, P. (2002) Holstein's milk detection in cheeses inferred from melanocortin receptor 1 (MC1R) gene polymorphism. J. Dairy Sci. 85, 707–715).

are in constant interaction with the environment (Rose, 2003). They absorb parts of their environment like oxygen or food materials, and at the same time, they excrete waste and therefore continuously modify it. Steven Rose (2001) even talks about «Envirome»[1] to propose an environmental unit corresponding to the Genome, both (genomes and enviromes) being «abstractions from this continuous dialectic». Darwin's observations mentioned on page 11 about the gradation in the size of the beaks of the chaffinch birds living on different islands in the Galapagos constitute an explicit example of the influence of geoenvironmental specificities on organisms.

Darwin's theory of evolution (Darwin, 1985) set up the first fundamental land-marks looking into the relationship between species and environment. It is a theory about adaptation to changing environments that above all provided a simple[2] universal mechanism for evolution, but on the basis of which numerous different interpretations were made. We have to remember only two that are important in our context. The first one is universal darwinism which claims that all life in the universe is governed by the Darwinian rules of variation, inheritance and selection. This means that evolution will occur in a population if ways to introduce variation, consistent selection process, and mechanisms for preserving or propagating the selected variants exist. In fact, universal darwinism presents a generic formulation of evolution that is applicable to any domain. The second one is ultra-darwinism, for which the micro-evolutionary mechanisms of organismal selection can be extrapolated to explain all phenomena in life's history.

Presently, we can distinguish three main evolution movements in the modern biological thought within which environmental pressure and genetic adaptation take place. On the one hand a gene-centred, ultra-darwinist and deterministic concept of evolution, led by Richard Dawkins who wrote his reference book «The Selfish Gene» in 1976. Darwin's view was that all biological evolution could be explained by natural selection acting on organisms. For him, all higher level order in nature could be explained by natural selection acting on organisms pursuing their own self-interest. Ultra-Darwinians still embrace this view. This vision really seeks explanations for every aspect of the living condition in terms of evolutionary imperative. Living organisms are considered to be survival machines pro-grammed for the preservation of the selfish molecules which are genes (Dawkins, 1989). To really understand all implications of this conception at a philosophical level, it is radically opposed to the freedom idea defended by Sartre in 1946 in «Existentialism and Humanism» for which a human being is what he makes of himself, and getting what he is willing : there is no determinism. On the contrary, we find in ultra-darwinism the roots of biological determinism and sociobiology, known now as evolutionary psychology. This vision of evolution was used to characterize much of modern biological thought till the end of the 1990s and still remains one of its major trends.

The second current is the naturalist approach defended by Niles Eldredge and Stephen Jay Gould, materialized and initiated in 1972 by their common theory of «Punctuated Equilibria». The naturalist view suggests that selection is simulta-

---

1. Contraction of «Environment» and «Genome».
2. «How stupid not to have thought of that», T.H.Huxley, 19th century biologist.

neously acting on several levels in nature, and that the nature's units of selection include genes, organisms, and species. Eldredge and Gould welcome Dawkins' idea of evolution in which selection operates on selfish genes, considering this as a contribution to unravelling evolution's hierarchy. But they dismiss his reductionist claim that all of evolution can be extrapolated from the selection of genes only (Stoelhorst, 2002). In 1989, Gould added the important «contingency» notion to the naturalist standpoint. In the course of history, contingencies are unpredictable sequences of antecedent states, and not a combination of determinism and randomness, chanciness, or accident (Gould, 1989).

This was relatively recently completed by Steven Rose who proposes in «Lifelines, Life beyond the gene» (1997) a global vision of living systems based on interactions between cells, organisms and ecosystems. Rose recognizes the power and role of genes without subscribing to genetic determinism. Human cannot be considered as empty organisms reducible to nothing but DNA replication machines. While recognizing the importance of genes and natural selection, this vision emphasizes the trajectories of living organisms through time and space. Reductionism moves science forward when research is made in laboratories, but it has to be complemented by an integrated study of complex interactions that occur within and between cells, organisms, and ecosystems. Organisms moves along lifelines, unique developmental and behavioral tracks from conception to death.

Fig 7.2. Contemplating snowman evolution or the morality of throwing one's precursor at someone. © Bill Watterson, Warner Books, 1998.

Finally, as third main evolution movement, there exist a major trend supported in particular by Christian de Duve and Simon Conway Morris and which is the theory of constrained evolution. In opposition with Gould's claim that if the tape of life were rerun it is very unlikely that anything resembling humans would emerge (Gould, 1989), the idea is that evolution has been constrained to follow certain paths leading more or less inevitably to the development of intelligence (Conway Morris, 2003). De Duve, although much less polemist than Conway Morris, Gould or Dawkins, is coming over the evolutionary convergence theory, notably using examples of parallel independent evolution of animals like big cats, anteaters and herbivores in North America, South America, Africa and Oceania which ended up at the same particularities and functioning (de Duve, 2005a). Moreover, he is proposing a cultural evolution characterizing the history of human civilization and funded on the transmission of cultural traits and on a cumulative process allowing the conservation of earlier acquisitions (de Duve, 2005b)[1]. This means that in

---

1. On the same theme, see also Cavalli-Sforza, L. (2004) L'evoluzione della cultura. Proposte concrete per studi futuri. Codice, Torino.

parallel with the natural selection mechanism, another cultural kind of selection would exist for humans. De Duve's vision of evolution also considers two directions for evolution. One *horizontal* concerning small genetic changes which are not affecting the general functioning of an organism, responsible for a diversity of forms. With horizontal evolution, de Duve is rejoining Gould and the contingency predominant notion as chance is the rule in this direction. But evolution not only created variations of forms on the basis of a standard life configuration, it also created new models which implied much more scarce but important genetic changes, altering organisms in a considerable manner. This is *vertical* evolution which is involving a major increase in complexity (de Duve, 2005b) and which can be brought closer together with John Maynard Smith major transitions (Smith, 1999), those «key events» in evolution like the origin of life itself, the first eukaryotic cells, reproduction by sexual means, the appearance of multicellular plants and animals, the emergence of cooperation, etc.

The distinction between those currents is useful to position the present research. The adopted approach may seem deterministic and gene-centred considering the fact that the interaction and its significance are assessed through specific loci determined by precise molecular analyses in comparison with a selection of climatic and topographic variables characterizing the animals' environment. But I stress the fact that it has to be understood as a simplification of reality that aims at identifying links between parts of the genome and environmental characters in order to better understand the genome functioning, and in turn in order to better understand the way an organism is adapted or adapts to a specific surrounding environment; it is not to be perceived as a reductionist gene's-eye viewpoint.

ENVIRONMENTAL MODELLING IN NATURAL SCIENCES

This research lies within a natural sciences context. It means that we are involved in the study of the physical world and its phenomena, and that we attempt to explain the working of some aspects of this highly non uniform world via natural processes. Studying and analyzing natural processes have specific implications we will stress by making a quick comparison between experimental and natural sciences approaches.

Paradigms in experimental sciences tend to be more universalistic and less local than the ones of natural sciences (Rose, 1997). They attempt to understand elementary processes controlled by a few parameters. The problems these disciplines are dealing with are very difficult, they have a high abstraction level, and moreover they can hardly be cut off from their context. On the contrary, complexity in natural sciences comes from the important number of intervening variables that are interacting simultaneously, from the fact that parameters are not fixed and properties non-linear. Moreover, as most of properties of studied systems are not quantifiable, natural processes are not capable of being captured in mathematical formula, and attempts to put numbers on them may even produce mystification (Rose, 1997). Obviously, it is not possible to separate an aspect among others and to measure it leaving aside all interactions between them. The main implication is that the global functioning of a system cannot be described by the sum of its ele-

mentary processes. This is a basic principle in the systems theory first proposed by the biologist Ludwig von Bertalanffy in 1968[1] as a reaction against reductionism, and attempting to revive the unity of science. Jean-Louis Lemoigne in «La modélisation des systèmes complexes»[2], goes in the same direction and shows that the solving of complex problems is made possible when people are respecting the complexity of reality rather than by having excessively recourse to reductionism[3].

Uncertainty

Experimental sciences simulate conditions in laboratories that aim at recreating a situation as simple as possible, where a minimum of parameters researchers want to study are considered. In natural sciences the situation is totally different as it is not possible to reproduce a natural situation without creating a different phenomenon. What is done instead is to measure combinations of parameters in order to find out which ones are the most significant, the most relevant. But even the elaboration of very accurate and sophisticated models won't prevent researchers from setting up approximations. Those advanced models may even produce vague information as every time a parameter is added, it comes with its associated uncertainty. It is often argued that it is nevertheless useful to know the respective part of variance explained by several parameters (Randin, 2005, oral communication). I understand the argument, but this is to be qualified by the fact that the total variance (reality) will never be known. In any case, those models will only catch a part of the variance, and will provide *indications* on how the studied phenomenon behaves. Analyzing natural processes induces the production of indicative information rather than physical values. The corollary being that all results we can obtain in natural sciences fall within the scope of a rather significant *uncertainty*.

Uncertainty follows from specificities of quantitative analysis in natural sciences, and constitutes a research domain in itself. Error, inaccuracy, imprecision, vagueness, ambiguity, etc., are tracked and studied by many people in many different fields and so it is for geographical information. Uncertainty is measured and its causes are identified to improve techniques to reduce it, to provide guidelines to prevent from generating it. In a paper about the limits of geographic knowledge, Helen Couclelis (2003) argues that error in (geo)information is inevitable and not only because of human limitations. Instead of focusing on data quality, she first proposes to shift the interest to the quality of knowledge this information allows to produce. Reasoning within the context of the limits of knowledge in general, she demonstrates with different examples that there are a lot of things about information we cannot know and which are not the result of imperfect information, nor depending on empirical facts or human limitations. Instances of intrinsic limits to knowledge are well known especially in experimental sciences. For instance,

---

1. See Bertalanffy, L.v. (1969) General system theory: Foundations, development, applications, The Penguin Press, London, 1971. See also de Rosnay, J. (1979) The Macroscope : a new world scientific system, Harper & Row, New York, and Schwarz, R. (1992) A Generic Model for the Emergence and Evolution of Natural Systems toward Complexity and Autonomy, in the Proceedings of the 36th Annual Meeting of the International Society for the Systems Sciences, Denver, Vol. II, p.766.
2. Lemoigne, J.-L. (1990) La modélisation des systèmes complexes, Dunod.
3. In french «Amplification réductrice».

in its incompleteness theorem, Gödel demonstrated in 1931 that within any given subject of mathematics, there is always a statement that cannot be proven either true or false using the rules and axioms of that mathematical subject itself. This statement can only be proved to be true or false within an expanded deductive system, but by doing so, a larger system is created with its own unprovable statements, etc. This means that every logical system of any complexity is incomplete by definition. Physics provides another example as it is known since Newton and then Poincaré that there is no analytic solution to the gravitational equations describing the dynamics of n bodies, where n > 2.

Couclelis considerations about knowledge and spatial information can be applied to the information produced in natural sciences analysis. She says that because of the existence of this intrinsic uncertainty, bad data and human fallibility shouldn't take all the blame. I argue that the probable existence of intrinsic uncertainty in natural sciences information - even if it cannot be quantified and even if it is likely to be very small - relativizes the care we take and the energy we spend to identify and determine errors. Caloz (2005) shows in a reflection paper dedicated to uncertainty propagation applied to spatial analysis that it is often impossible to quantify this uncertainty, and that one should rather have recourse to a qualitative reliability level of the measures carried out. It doesn't mean that specialists should stop their efforts in trying to reduce uncertainty as it is important to yield improvements in the quality of information. But it is one reason more to definitively accept the fact that measures and analysis in natural sciences are producing relative and indicative information that has to be completed by expert advises to make acceptable predictions. Methods and techniques allow the production of indicative information to come closer to reality, but they don't make it possible to rebuild or reconstitute reality. «Any representation of reality we develop can be only partial. There is no finality, sometimes no single best representation. There is only deeper understanding, more revealing and enveloping representations» (Woese, 2004). In these conditions, as it is not possible to exactly describe the functioning of phenomena in a deterministic way even by - or especially by - outbidding with the number of parameters, researchers have to correctly *interpret* the results produced by models. In this perspective, they have to improve as much as they can their personal knowledge of the studied matter, and then to commit this knowledge and the one of specialized people to provide partly subjective and perceptual but expert interpretations (Beven, 2001; 2002).

Equifinality

At this point, and keeping on relativizing many-parameters models, it is necessary to quote the very elegant and relevant concept proposed by Keith Beven called *equifinality* (Beven and Binley, 1992; Beven, 2001). Beyond considerations previously made about cumulated uncertainty, Beven is rejecting the usual habit that consists in considering one single optimum parameters set only. «A priori any model structure and parameter set that predicts a required variable in an application is a potentially useful simulator. However, it is often very difficult to accept that a particular (...) parameter set (...) is dominant in fitting available observations.» (Beven, 1998). Beven rightly proposes not to be interested only in one model fitting at best observations, but to assess the relative performance of all

possible models in terms of likelihood measures. The Generalized Likelihood Uncertainty Estimation (GLUE) is based on this principle and the analysis focusses on parameter sets rather than on the behavior of individual parameters and their interaction.

In concrete terms, Monte-Carlo simulations are applied, and for each run parameter values are chosen randomly by uniform sampling across their ranges. A degree of belief or likelihood value[1] is attributed to all processed parameter sets, making it possible to compare them; it represents a form of Bayesian averaging[2] over all behavioral models. This highlights the fact that the usual optimum parameter set we consider in multivariate models is one solution among many others which is often not explaining much more variance than the following ones in terms of goodness of fit. And it clearly shows that different combinations of parameters can lead to almost similar effects with a rather similar goodness of fit (see table 7.1).

Ranked Parameter Sets					
Rank	M	LnTo	SRmax	SRinit	Eff
1	0.034	5.228	0.026	0.007	0.858
2	0.036	9.64	0.025	0.001	0.856
3	0.032	4.735	0.011	0.064	0.85
4	0.033	5.218	0.007	0.534	0.849
5	0.031	5.646	0.01	0.518	0.846
6	0.031	8.727	0.012	0.349	0.844
7	0.033	9.372	0.007	0.178	0.844
8	0.032	8.223	0.008	0.214	0.844
9	0.032	8.609	0.008	0.11	0.843
10	0.031	7.858	0.013	0.6	0.843

Table 7.1. Illustrating equifinality. Columns «M», «LnTo», «SRmax» and «SRinit» are four variables whose combinations are constituting different models. These models are ranked according to their degree of belief which is displayed in the right column. Source : GLUE software, Beven (1992).

Realising this aspect of modelling is important; then the advantage of such an approach is that the possibility exists for an expert to chose the most appropriate model on the basis of any subjective criteria constituting his knowledge. The systematic recourse to the optimum parameter set may be seen as a characteristic of *pragmatic realism* (recomposing reality), the predominant philosophy underlying environmental modelling and to which Beven (2002) proposes equifinality as an alternative.

The difference between both philosophies can be appreciated by referring to a Carl Woese paper (2004) in which he discusses the notion of reductionism which is confused according to him. He interestingly distinguishes a fundamentalist from an empirical reductionism. The fundamentalist - compared to the reductionism of the 19th century classical physics - is metaphysical in essence and states that living systems can be completely understood in terms of the properties of their constituent parts. One must admit that it shows some analogy with pragmatic realism. On the other hand, empirical reductionism is methodological and is a mode of analysis asserting that the dissection of an entity into its constituent

1. This value allows to assess the performance of the parameter set. This can be the sum of squared errors.
2. In the Bayesian view, using a single model to make predictions ignores the uncertainty left by finite data as to which is the «correct» model; thus all possible models in the model space under consideration should be used when making predictions, with each model weighted by its probability of being the «correct» model. This posterior probability is the product of the model's prior probability, which reflects our **domain knowledge** (or assumptions) before collecting data, and the likelihood, which is the probability of the data given the model. This method of making predictions is called Bayesian model averaging (Domingos, 2000).

parts will allow a better understanding of its functioning. This is much like equifinality or the systems theory whose goal is to assess the different combinations of some factors of a whole in order to better understand the way it is working.

Equifinality is illustrated thanks to a simple example in table 7.1 and this let us suppose what could be the number of conceivable combinations when applying a complex model ! It relativizes in a convincing way the constant and «blind» use of the - fundamentally - reductionist optimum parameter set. On this basis, one should wonder if it wouldn't be worth choosing an opposite way than making models more sophisticated, and propose very simple models which are taking into account the different observations made here above.

SIMPLICITY IN MODELLING

A possible remedy to apply in order to avoid too much complexity arising because either of complicated models or in sophisticated error measurement techniques, is to employ simple models to study phenomenon in order to move as near as possible to reality and to produce a few pieces of certainty, rather than a lot of fragments of vagueness. Pierre-Gilles de Gennes, Physics Nobel prize in 1991, but also involved in chemistry and biology researches[1], well exemplifies the concept. In an interview, he explains the virtues of simplicity when he and his team were studying cell adhesion in the Curie Institute. They resorted to models with a minimum of adjustable parameters in order to bring out a very simple phenomena and to produce a minimal and acceptable error. He was surprised that other teams had recourse to complex chemical reactions implying heavy and time consuming verification processes (Leach and de Gennes, 2005).

This reasoning shows some resemblance with the parsimony principle, also called «Ockham's Razor». Applied to modelling, it states that if two models in some way adequately model a given set of data, the one that is described by a fewer number of parameters will have better predictive ability given new data (Jaynes, 2003). This form of «methodological reductionism» or «law of economy» is attributed to an english logician and franciscan friar of the 14th century, William of Ockham (1287–1347)[2].

Simplicity is also a preoccupation for science philosophers. In 1927, Hermann Weyl (quoted by Popper, 2004) recognized the role of simplicity and asserted that «The problem of simplicity is of central importance for the epistemology of the

---

1. «I'm a physico-chemisto-biologist!», (Leach and de Gennes, 2005).

2. «Pluralitas non est ponenda sine neccesitate» or «Given two equally predictive theories, choose the simpler». The parsimony principle is often expressed as «Entities should not be multiplied beyond necessity». This can be interpreted in two different ways. The first is a preference for the simplest model that adequately fits the data. The second is a preference for the simplest subset of any model which fits the data. The difference is that, beyond equifinality properties, it is possible for two different models to explain the data equally well, but have no relation to one another. One problem with the original formulation of the principle is that it only applies to models with the same explanatory power. A more general form can be derived from Bayesian model comparison and Bayes factors, which can be used to compare models that don't fit the data equally well (Jaynes, 2003), on which is based equifinality. These methods can sometimes optimally balance the complexity and power of a model [Sober, E. (1981) The Principle of Parsimony. British Journal for the Philosophy of Science 32:145–156]. It is to note that Francis Crick warned about the parsimony principle applied to biology : «While Ockham's razor is a useful tool in the physical sciences, it can be a very dangerous implement in biology. It is thus very rash to use simplicity and elegance as a guide in biological research» [Crick, F. (1988) What Mad Pursuit. New York, Basic Books].

natural sciences»[1]. In general, most philosophers believe that simpler theories are better[2] because of the syntactic simplicity, of the elegance, of the conciseness of the theories[3].

But in «The logic of scientific discovery» Popper (2004) is rejecting those aesthetic and pragmatic aspects arguing that they have little interest from the point of view of the theory of knowledge. He analyzes simplicity in probabilistic terms and affirms simpler hypotheses are desirable, not because they are more likely to be true, but because they are easier to eliminate if false. This is the quality - to be easily eliminated - which will be amply exploited in the modelling applications applied to natural selection signatures detection.

These philosophical considerations about environmental modelling constitute the foundations of the reasoning the approach presented hereafter is relying on. Even if «The philosophical subtleties are not really necessary to the practicing environmental modeler», Beven said (2002) when describing identified flaws of pragmatic realism, that analysis constitutes an excellent opportunity to forthwith reinforce a theory or a particular approach of environmental modelling with a concrete application.

---

1. Weyl, H. (2000) Philosophie der Mathematik und Naturwissenschaft, Scientia Nova, Oldenburg, 7th edition.
2. Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/simplicity/ (17.11.2005)
3. «(...) any scientist who has succeeded in representing a series of observations by means of very simple formula (...) is immediately convinced that he has discovered a law». Schlick, Naturwissenschaften 19, 1931, quoted by Popper (2004).

DETECTING SIGNATURES OF NATURAL SELECTION

In accordance with the ideas about simple modelling described above, it was decided to resort to univariate logistic regression. This is a variant of standard linear regression, applicable when the dependent variable is a dichotomy such as the presence or absence of given alleles at specific loci characterized by quantitative climatic and topographic information[1]. This is likely to provide a measure of the association level between microsatellite or AFLP alleles and the environmental parameters favouring or not their presence.

Logistic regression

Logistic regression has become, in many fields, the standard method of data analysis concerned with the description of the relationship between a response variable following a binomial distribution and one or more explanatory variables. It is used to model the probability of occurrence of a binary or dichotomous outcome. The principle of logistic growth[2] is used nowadays to model as well animal or plant habitat, population growth, as economical modelling[3] (Cramer, 2002).

Significance of coefficients and modus operandi

In the context of the proposed method, logistic regression is used in a particular way. The idea is to assess the significance of the models constituted by all possible *locus ↔ environmental variable* pairs, and to highlight the markers involved in the most significant models as «possibly under natural selection» or «potentially implied in adaptation processes». To get a probability of presence of examined loci or alleles given specific values of investigated environmental variables **is not** the main goal of the method applied in our context.

Estimation of the parameters

The significance of models in logistic regression is based on the comparison between the values predicted by the model and the observed values (Hosmer and Lemeshow, 2000). It is thus necessary to calculate the parameters of an estimated function that best fits observed values.

According to the properties of logistic regression explained in appendix 13, as a logit is a natural logarithm of odds, and that odds are a function of p, the probability of a molecular marker to exist for a given value of the independent variable is

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

(EQ 7.1)

In this case, the natural logarithm of odds (or *logit*) is assumed to be linearly related to x, the independent variable. Thus it could in theory be possible to run an ordinary regression with *logits* as dependant variable to calculate the parame-

---

1. See explanations about environmental variables in chapter 4, page 38. A list of the variables is displayed in appendix 4.
2. See appendix 13.
3. Daniel McFadden earned the Nobel price in Economics in 2000 jointly with James Heckman thanks to works linking the logit model to discrete choice theory from mathematical psychology with the aim to provide a theoretical underpinning for the use of the logit.

ters of the function. But in reality, we don't have logits but only presences and absences of a marker (1 and 0). In such a model, there is no mathematical solution to produce the equivalent of least squares estimates applied in linear regression to determine the parameters. It is necessary to apply a loss function[1] called «maximum likelihood». This method yields values for the unknown parameters which maximize the probability of obtaining the observed set of data (Hosmer and Lemeshow, 2000; Morgenthaler, 1997). The likelihood is a conditional probability $p(Y|x)$, the probability of Y given x. It means that the likelihood function is expressing the probability that observed data are a function of unknown parameters. Thus the «estimators» of the maximum likelihood are the $\hat{\beta}_0, \hat{\beta}_{1...}$ maximizing this function (the ^ symbol denotes the maximum likelihood estimate). Then $\hat{p}(x)$ is the maximum likelihood estimate of $p(x)$. Likelihood equations allowing to determine maximum likelihood estimates can be assessed thanks to iterative methods in numerical analysis[2]. In our context, this task is carried out by the GLMfit function supplied by the Matlab[3] software and implemented by Thierry Lassueur (2004).

Choice of the statistical test

To identify markers or alleles likely to be under natural selection, the relevance of the inclusion of examined geoenvironmental variables within models has to be assessed. To this end, the significance of the coefficients calculated with the GLM-fit function has to be evaluated by statistical tests. Their role is to answer the question to know whether a model including the examined variable tells us more about the response variable than a model that does not include that variable. The comparison of the observed values of the response variable with those predicted by each of two models (the model calculated with the examined variable, and the model calculated without it) is in a position to answer the question. In logistic regression, the comparison of observed to predicted values is based on the log likelihood function (Hosmer and Lemeshow, 2000, p.12). Three statistical tests are mentioned to be based on the likelihood ratio : G (also named likelihood ratio), Wald and Score. The latter is rather regarded as a multivariate test by Hosmer and Lemeshow and thus was not taken into account. It is important to note that, in this univariate model context, the quality of representation of the observed values by the predictive values (goodness-of-fit) was not considered[4].

It is in order to reinforce the robustness of the method that it was decided to run two statistical tests to achieve this comparison and to determine the significance of the models : a model is considered significant only if both test fail to reject the null hypothesis. These tests are described hereafter.

---

1. A loss function is a measure of fit between a mathematical model of data and the actual data.
2. Numerical analysis methods follow a series of steps. First, random initial estimates of the parameters are picked. Then the likelihood of the data given these parameter estimates is calculated. Then the parameter estimates are slightly improved and the likelihood of the data recalculated. This is repeated until the parameter estimates do not change much. The number of iterations is determined by the developers of the function. In the case of the GLMfit function, Matlab developers set the iteration limit to 100.
3. The MathWorks, Inc. Software, Natick, MA, United States.
4. For further details, see chapter 5 in Hosmer and Lemeshow (2000), and analyses carried out by Thierry Lassueur (2004) in the context of a diploma work at the GIS laboratory of the EPFL (Ecole Polytechnique Fédérale de Lausanne or Swiss Federal Institute of Technology).

Likelihood ratio or G statistic

G is defined as :

$$G = -2\ln\frac{L}{L'}$$

(EQ 7.2)

where L is the likelihood of the initial model (with a constant only) and L' the likelihood of the new model including the examined variable. If added parameters are equal to zero, this statistic is following a chi-square distribution with a number of degrees of freedom equal to the number of added parameters (see Hosmer and Lemeshow, 2000).

Thus the null hypothesis (H0) is «Added parameters are equal to zero».

Wald statistic

The Wald test is obtained by comparing the maximum likelihood estimate of the $\hat{\beta}_t$ parameter to an estimate of its standard error. Under the null hypothesis, the resulting ratio follows a standard normal distribution.

The method to assess the variance is conform to the theory of maximum likelihood (Hosmer and Lemeshow, 2000; Rao, 1973).

$$W = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}$$

(EQ 7.3)

The null hypothesis is that «the model with the examined variable doesn't explain the observed distribution better than a model with a constant only».

Multiple hypotheses testing and confidence level

For a given confidence level, both G and Wald statistics may simultaneously fall beyond the α significance threshold or not and thus the null hypothesis being rejected or not. If one of both tests fails to reject null hypothesis, the model is deemed not to be significant. This is the elementary reasoning on the basis of which significant *locus ↔ environmental variable* relationships are identified. But additional elements have to be considered.

For each practical case presented in the forthcoming sections, many univariate models had to be simultaneously run (more than 85'000 in the case of sheep microsatellites) in order to detect alleles and markers likely to be under natural selection. This is a multiple hypotheses testing context. It is recognized that when one wishes to simultaneously test several hypotheses at a common significance level α, the generalized Type I error probability (the probability of rejecting at least one of the hypotheses being tested that is in fact true) is typically much in excess of α (Shaffer, 1995). There are a large number of multiple testing procedures available, and we chose to apply the simple Bonferroni correction (Shaffer, 1995).

This correction implies to divide the wanted significance threshold α by the number of comparisons (the number of models simultaneously processed) to get a Bonferroni corrected significance threshold α.

Moreover, it was noticed that - despite the application of the Bonferroni correction - using a standard confidence level of 99.9% didn't allow to sufficiently discriminate very significant associations from others. Therefore it was decided to gradually increase the confidence level of one order of magnitude[1] - starting from 99.9% - and to sum the number of significant models at each order of magnitude increase (99.99%, then 99.999%, etc.) until their number doesn't evolve any more. This stabilization happens when the considered p-value is at least two orders of magnitude smaller[2] than the last level observed (see red circle and red arrow in table 7.2). At this significance level, deemed to be discriminant enough, alleles or markers are ranked according to the number of significant models they are involved in (see table 7.2).

This operation was carried out with the help of rejection tables (Joost, 2005; Lassueur *et al.*, accepted, Ecological Modelling) which are simple spreadsheet tools allowing to visualize whether models are significant or not while α is decreasing.

Table 7.2. Common frog data : a rejection table with p-values provided by G and Wald tests (see *Rana temporaria* on page 121). This didactic example shows how most significant models are identified. It is based on 1 environmental variable only (altitude). Sources : Aurélie Bonin, Laboratoire d'Ecologie Alpine, UJF, Grenoble.

Table 7.2 is a rejection table showing how markers or alleles possibly under natural selection have been identified in the case studies presented from page 121. This table displays results of the significance of models composed of one environmental variable only (altitude) and of one AFLP marker. The table displays 46 AFLP markers (on a total of 364) which are involved in at least one significant model. The yellow part shows the results for the G test, and the orange part for the Wald

---

1. Order of magnitude : a change in quantity or volume as measured by the decimal point. For example, from tens to hundreds is one order of magnitude. Tens to thousands is two orders of magnitude; tens to millions is three orders of magnitude, etc. Source : Computer Desktop Encyclopedia. Computer Language Company Inc., 2005. Answers.com, http://www.answers.com/topic/orders-of-magnitude (21.11.2005).
2. A p-value of 0.00000356 is at least 2 orders of magnitude smaller than 0.000356.

test. In both cases, for the correspondent significance threshold, a «1» (H0 rejected) is displayed when the presence of a given marker is significantly explained by altitude, and a «0» when it is not. The blue part reveals the definitive results, that is models significant for both statistical tests for the correspondent significance level displayed on the left. In this case, a «1» means that the H0 is rejected for both G and Wald tests.

Molecular markers and selection signatures : what about the neutral theory ?

The proposed Spatial Analysis Method (SAM) puts geoenvironmental information in touch with molecular data. While the description of geoenvironmental variables provided in chapter 4 is sufficient to grasp all implications of the results presented farther, additional details need to be exposed about molecular markers regarding the neutralist-selectionist debate (Avise, 2004), this given the goal of the analysis which is related to natural selection.

Since 1968 when Kimura first asserted that a majority of evolutionary changes at the molecular level were not caused by Darwinian selection but by random drift of selectively neutral mutants, the neutral theory gained rapidly «universal acceptance as a molecular evolution's gigantic null hypothesis» (Avise, 2004). The neutral theory - which has roots in the 1930s quantitative school of population genetics (Fisher, Wright) - does not deny the role of natural selection in determining the course of adaptive evolution, but assumes that only a tiny fraction of DNA changes in evolution are adaptive in nature. It claims that the great majority of molecular substitutions exert no significant influence on the survival or reproduction among species (Kimura, 1987). These elements are still considered as a basic theoretical construct whose prediction must be refuted before being allowed to invoke any form of selection.

But despite this important and rather consensual neutralist movement, things began to evolve from the 1980s when genetic markers became increasingly available. The more markers at disposal, the more chances to detect one of those small parts of DNA potentially under natural selection. Fine-scale molecular characterization of genes allowed to show the first footprints of natural selection in the middle 1980s for the earliest works, and later in the 1990s (Avise, 2004, and references therein). Nowadays several approaches are existing to deduce the action of various forms of natural selection on DNA sequences (see for instance Skøt *et al.*, 2002; Luikart *et al.*, 2003). Despite the neutral theory, markers can be employed to check potential association with environmental parameters and thus to locate genes under selection. The idea is that if genes under selection are existing in the studied genome, any of the utilized markers may be located next to them and show an outlier behavior when confronted to other neutral markers. This is particularly relevant for livestock which have a high degree of linkage desequilibrium (see explanations on page 46) in their genome due to population structure and to non random mating (Farnir *et al.*, 2000; Vallejo *et al.*, 2003)[1].

---

1. For example, the Black-and-White cattle breed population counts over 25 million animals worldwide. In the Netherlands only, the population of black-and-white cattle comprises 1.2 million of cows. But estimates of the theoretical population size yield numbers as low as 50 ! This is attributable to the widespread use of artificial insemination and to the intense artificial selection for increased milk production (to give an idea, the 10 top sires account for 40% of the inseminations in the Netherlands) (Farnir *et al.*, 2000 and references therein).

Thanks to a relatively new way of using the properties of linkage desequilibrium, it is possible to highlight signatures of recent positive selection using neutral markers (see details in Bamshad and Wooding, 2003, p. 104)[1].

But continuing uncertainties are existing about the relative roles of selected and neutral ***mutations*** in evolution. According to the neutral theory, molecular markers gain their informativeness by... being neutral. On the other hand, restrictively considering molecular markers according to the neutral point of view may lead to wrong interpretations. To illustrate this aspect, Avise (2004) mention that strong balancing selection (a polymorphism which is maintained within a population) obtained via heterosis[2] «can inhibit population differentiation in allele frequencies by random drift and result in uniform spatial patterns that could be misinterpreted as evidence for high ***gene flow*** under neutral models».

The neutralist-selectionist debate is still current. I hope that the method I propose and the results presented in the next section - carried out independently of any genetic model - may possibly contribute to make it progress notably thanks to perspectives offered by large parts of genome scans. Indeed, we often refer to *population genetic*s to describe the general context of molecular data analysis. This literally means that investigations are led at the gene level. But with the recent biotechnological developments, molecular markers data are increasingly becoming available at the genome level promoting the development of genome-wide tests for molecular adaptation and the identification of outliers. «Molecular technologies are bridging the gap between ***genotyping*** and genome typing» (Luikart *et al.*, 2003), favouring the emergence of *population genomics.* This approach makes it possible to simultaneously study numerous loci or genome regions to better understand the roles of evolutionary processes (***genetic drift***, ***gene flow***, natural selection). Population genomics relies on two main principles which are that neutral loci across the genome are similarly affected by demography and the evolutionary history of populations, and that loci under selection often behave differently and therefore reveal outlier patterns of variation (Luikart *et al.*, 2003).

---

1. See also Sabeti, P. C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419, pp. 832–837 and Nordborg, M. & Tavare, S. (2002) Linkage disequilibrium: what history has to tell us. Trends in Genetic 18, pp. 83–90.
2. Fitness superiority of heterozygotes.

INTRODUCTORY NOTE TO CASE STUDIES

Now that the theoretical framework of the method is well defined, it will be applied to four case studies. The first one is dedicated to the common frog (*Rana temporaria*) and involves only one environmental variable. It is therefore well adapted to demonstrate the approach in a didactic way. Moreover, each of those four case studies is containing a validation part as the results supplied by the SAM were compared to the ones calculated with the help of a standard genetic approach to detect F_{ST}-outlier loci (Beaumont and Nichols, 1996; Beaumont and Balding, 2004; Beaumont, 2005; Bonin *et al.*, in press). The first application on frog will also present this method and the genetic model it is relying on.

For each case study, the calculations on genetic data were carried out by different population geneticists which will be mentioned later on in the respective concerned section.

With the exception of the analysis dedicated to the Scandinavian Brown Bear, the applications presented hereafter fall under the competence of population genomics as they correspond to the ideal molecular approach described by Luikart *et al.* (2003), uncovering hundreds of polymorphic markers.

RANA TEMPORARIA

A research on the local adaptation of the common frog (*Rana temporaria*) along an altitude gradient was led by Aurélie Bonin of the Laboratoire d'Ecologie Alpine in Grenoble. On the basis of a genome scan implying 392 AFLP markers, the main goal of this study was to look for loci diverging from neutral expectations - thus potentially under natural selection - when comparing populations from different altitudes applying two outliers detection methods[1].

Geographical origin and sampling

Rana temporaria is the most abundant amphibian in Europe, occurring from Northern Spain to subarctic Scandinavia and from the sea level up to an altitude of 2'500 m in the Alps. Studied populations are living in the north of the French Alps, in a geographical area covering approximately 7'000 km2. Sampled individuals are distributed among three classes of altitude. They come from two *Low1* and *Low2* low altitude populations (around 400 m), two *Inter1* and *Inter2* intermediate altitude populations (1000 m) and two *High1* and *High2* high altitude populations (2000 m) in different mountain massifs. For each population, 28 to 34 adult frog fingers and/or tadpoles were sampled.

---

1. In this chapter we used the results supplied by one of the two methods evoked here (Beaumont and Nichols, 1996) as it was applied for all case studies. The results provided by the other method (Vitalis *et al.*, 2001; Vitalis *et al.*, 2003) will solely be mentioned in the frog case study. Both softwares were originally designed for codominant data and were modified in order to process AFLP data (Bonin *et al.*, in press).

Population genomics method

Two main statistical approaches exist to test for outlier loci. On the one hand a theoretical one which has recourse to simulated neutral distributions of F_{ST} and an empirical one based on observed data. The latter is used less often as it requires the ***genotyping*** of hundreds of loci across the genome to build a robust null distribution (Luikart *et al.*, 2003).

We will focus on the theoretical approach, particularly on a F_{ST}-outlier test developed by Beaumont and Nichols (Fdist software, 1996), which is using computer simulations to model neutral loci (a null distribution of 50'000 simulated loci was generated for *Rana temporaria*). The advantage of this approach is that different population structures can be simulated to assess the influence of different demographic situations (Luikart *et al.*, 2003). It is based on the principle that genetic differentiation between populations is expected to be higher for loci under selection than for the rest of the genome. Loci showing abnormal behavior and lying outside the neutral distribution are detected as outliers. Dfdist (modified from Beaumont and Balding, 2004) applies a hierarchical Bayesian approach to compute F_{ST} values conditional on heterozygosity in a subdivided population under the symmetrical island model[1] (Wright, 1951; Bonin *et al.*, in press, and references therein).

Spatial analysis method

In parallel with the intrinsic goals of the *Rana temporaria* study, we commonly decided to apply the SAM to attempt to detect outliers possibly selected by altitude among the 392 AFLP markers. This operation was likely to supply an opportunity to compare the results out of three different methods, the SAM one being hitherto totally novel and unpublished.

A data set containing on the one hand geographical coordinates[2] and altitude, and on the other hand AFLP markers information was prepared (see table 7.3 on page 123). For each marker, the sum of its presences among the 138 individuals was calculated. Then the markers have been sorted out according to their number of presences; those counting less than 5 presences and those counting more than 130 presences were removed from the data set (it corresponds to about 5% of the total effective on both sides of the frequency distribution). Indeed markers absent from almost all locations and those present at almost all locations are not informative on the natural selection point of view.

Then original data were arranged so that they could be imported in Matlab in order to run a procedure having recourse to the GLMfit function which has to solve the likelihood equations allowing to determine the maximum likelihood estimates of the β_1 parameter (altitude).

---

1. The island model deals with a species which is subdivided into a number of discrete finite populations among which migration occurs. If the number of populations is small, an assumption of equal rates of migration between each pair of populations may be reasonable approximation. ***Mutation*** at a constant rate to novel alleles may also be assumed (Latter, 1973).

2. Geographical coordinates are used either in order to retrieve altitude in case it couldn't be measured on the field, or to check the altitude, both operation being processed with the help of a digital elevation model. An additional role of geographical coordinates is to allow the geographical mapping of relevant molecular markers once the analysis is completed.

In addition, this procedure also :

- manages the number of models to be processed (the user has to indicate the number of markers and the number of environmental variables to be treated);
- calculates the p-values associated with both G and Wald statistical tests mentioned on page 117 for each model;
- stores temporary results within matrices;
- generates a graph with the sigmoid response curve for each model;
- exports tables in text format to be used in any spreadsheet or statistical software.

Idabs	Individuals	Longitude	Latitude	Altitude	Species	Marker 103	Marker 197	Marker 131	Marker 175	Marker 62
37	CU22	6.2775	45.3697	425	Rana temporaria	0	1	0	1	1
38	CU23	6.2775	45.3697	425	Rana temporaria	1	1	1	1	1
39	CU25	6.2775	45.3697	425	Rana temporaria	1	1	0	0	1
40	CU26	6.2775	45.3697	425	Rana temporaria	1	1	1	1	1
41	CU36	6.2775	45.3697	425	Rana temporaria	0	1	1	0	1
42	CU37	6.2775	45.3697	425	Rana temporaria	1	1	0	0	1
43	TI2	6.9022	45.9522	1082	Rana temporaria	1	0	1	1	0
44	TI3	6.9022	45.9522	1082	Rana temporaria	0	0	1	1	1
45	TI4	6.9022	45.9522	1082	Rana temporaria	1	0	1	1	0
46	TI5	6.9022	45.9522	1082	Rana temporaria	0	0	1	1	1
47	TI6	6.9022	45.9522	1082	Rana temporaria	0	1	0	1	0

Table 7.3. Table with data describing *Rana temporaria* geographic location, altitude and molecular information. On the right side are displayed five out of the 392 AFLP markers with their respective presences and absences.

The import is user-friendly as Matlab is directly reading Excel files. The format is also uncomplicated : on the left side *n* columns with environmental variables and on the right side *n* columns with binomial information, that is markers presence (1) or absence (0).
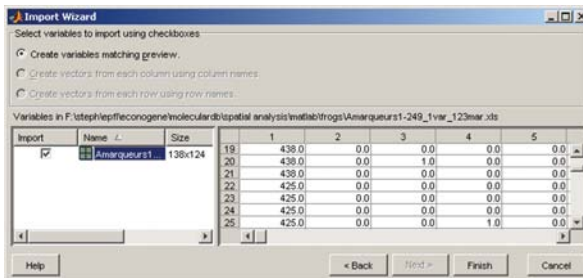
Fig 7.3. Data format when environmental and molecular data are imported into Matlab. Each line is an individual. The altitude is displayed in column 1, and the presence/absence data of markers appear in columns 2, 3, 4 and 5 (first 4 AFLP markers on 364).

The program distinguishes environmental from molecular data thanks to the information obtained from the user (see first point here above). The order with which the different variables are placed in the file is important as column names have to be removed before the import into Matlab, and as it will be necessary to identify the exported columns after processing. Their ID is corresponding to their frequency among sampling locations and this constitutes the order in which they are imported and exported into/from Matlab.

Results

Table 7.4 on page 125 is showing the results obtained with the SAM. 46 markers are displayed on the 364 which were analyzed, and appear to be significant for one single test at least with a significance threshold set to 2.747-E05 (this corresponds to a 99% confidence level including Bonferroni correction). Here are key-indications for the reading of this table and for similar ones displayed in the sections about sheep and bear :

- on the first line are displayed the identifiers of the markers with in black the ones which were also detected by mean of the Population Genomics Method (PGM), with the best candidates in bold. Markers in red were not detected by the PGM;
- on the second line their frequency at sampling locations;
- on the third the p-value (unreadable);
- on the 8 next lines the significance thresholds. It starts from a value corresponding to a confidence level of 99% which is increased of one order on each successive line;
- on the right of the thresholds, a «1» means that the tested variable is significantly participating in explaining the presence of an observed allele, and a «0» that it is not;
- the yellow part is about G test, the orange part concerns the Wald test;
- the blue part is the consolidated test : a blue «2» means that the model was significant with both tests;
- the «Total» line is showing the test-score - the sum of significant models on both tests - of the markers which are sorted from the left to the right according to this criteria;
- cells in grey are emphasizing models which have to be considered with caution given their low frequency among sampled locations;
- the last line is visually showing correspondences between both SAM and PGM. In dark green the best candidates, and other candidates in light green;
- the significance threshold for which the number of significant models stabilized is 2.74E-13; this line is not displayed in the table.

According to the SAM, marker 301 is very strongly standing out with all tests being very significant. Ignoring markers 320, 214, 337 and 357 (in grey because of few presences among samples, like 179 and 265), number 84 is the second serious candidate for being selected by altitude. But the consolidated significance of the latter is nevertheless 6 orders of magnitude lower than for marker 301. To conclude with those elementary observations, marker 250 is interesting because it shows solid results with the Wald test while no G test is significant. In fact, in this case the maximum number of iterations for the calculation of maximum likelihood estimates L (see equation 7.2 on page 117) was reached and the equation not solved. This is explaining why the G test produced no result.

As for PGM, Aurélie Bonin led the analysis on the one hand by mean of inter-altitude analyses for which 12 possible pairwise comparisons of the different populations where carried out, and on the other hand through a global analysis on all individuals, the latter corresponding to the general approach of the SAM. Table 7.5 on page 125 is revealing the results : 12 inter-altitude comparisons performed with Dfdist led to the characterization of 43 different loci. Among them, 29 loci appeared in only one analysis and were considered as false positives. 11 loci were related to one population in particular, and were classified into the category of outliers due to local effects. The three loci appearing in at least two independent comparisons (84, 248, and 301) were all also detected by the SAM although the 248 appeared to be selected by altitude in a less convincing manner.

Table 7.4. Results of the spatial analysis method run on *Rana temporaria* data displayed within a rejection table. 46 AFLP markers out of 364 are shown. Those markers are shown to be selected by altitude with one of both G or Wald statistical tests at least, and with a confidence level of at least 99%.

Software	Analysis type	Number of loci detected		Possible cause of outlier behaviour	Good candidates
Dfdist	Inter-altitude analyses (12 possible pair-wise comparisons)	43	29 loci appearing in only one analysis	False positives	-
			11 loci attached to one population in particular	Local effects	-
			3 loci appearing in at least two independent comparisons	Adaptation to altitude	Locus **84** (4/3) Locus **248** (3/3) Locus **301** (4/4)
	Global analysis on the three altitude categories	6		Adaptation to altitude	Locus **84** (1/1) Locus **97** (1/0) Locus **228** (1/0) Locus **250** (1/0) Locus **301** (1/1) Locus **388** (1/0)

Table 7.5. Outlier detection results using the Dfdist software. The number of significant analyses are indicated after the *locus* identifier, between brackets : the first number corresponds to a 95% confidence level and the second one to a 99% confidence level. Sources : Aurélie Bonin, Laboratoire d'Ecologie Alpine, UJF, Grenoble (Bonin *et al.*, in press).

125

Considering the global analysis results, that is to say in the same configuration as for the SAM, it is encouraging to observe that loci 84 and 301 appear to be the best candidates as also revealed by the SAM. Then loci 97, 228, 250 and 388 are all detected by the SAM and also seem to be less «sensitive» to the altitude parameter. Only one *locus* (301) is strongly revealed by the two PGM (Beaumont and Vitalis approaches) and by the SAM, what is conferring on him good support to its status of outlier because of adaptation to altitude.

.
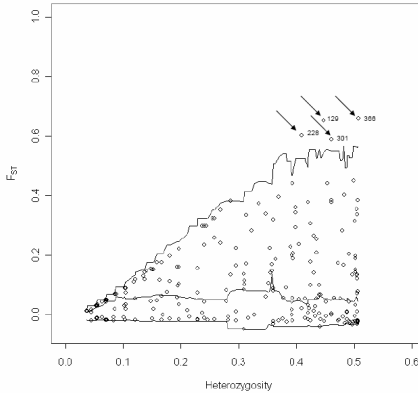
Fig 7.4. Plot of FST values against heterozygosity estimates for one inter-altitude analysis performed with Dfdist (population Low1 vs. population High1). Each dot represents an AFLP marker. The lower and the higher lines are including the 95% confidence intervals. The outlier loci situated out of the neutral envelope are pointed out by arrows and referred to by their identifier. Sources : Aurélie Bonin, Laboratoire d'Ecologie Alpine, UJF, Grenoble (Bonin *et al.*, in press).

This first example on *Rana temporaria* is showing that the proposed SAM is a possible alternative to the PGMs in order to detect natural selection signatures within genomes. Of course, one single example is not sufficient to rule on the relevance and on the effectiveness of a method. We will now assess its usefulness applied to a domestic species, while sophisticating the approach by measuring the simultaneous selectiveness of markers by several environmental variables.

ECONOGENE SHEEP AND MICROSATELLITES

This methodology was experimented on 1449 individuals among 48 european and middle Eastern autochthonous sheep breeds from marginal areas (Peter *et al.*, submitted to Journal of Heredity) with microsatellite markers. One initial requirement was to reencode microsatellite data - unlike AFLPs for which a marker is present or not (1/0) - as they define the length of the fragment amplified within the microsatellite *PCR* on each homologous chromosome of each individual (microsatellites are inherited codominantly, and therefore there is one allele which is derived from the father and one from the mother). After reencoding, these are alleles at specific loci whose presence was tested versus environmental parameters variation (see reencoding explanations on page 51). 744 alleles located at 31 loci were confronted to 118 topo-climatic variables for a total of 87'792 models to be run.
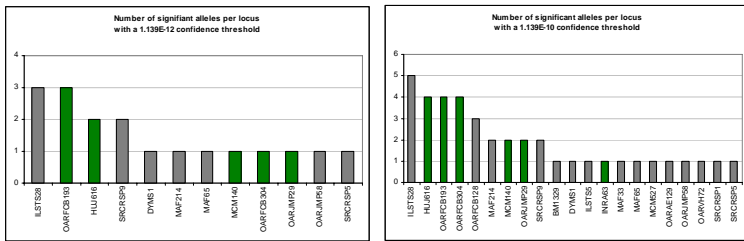
Fig 7.5. Loci and histograms of the frequencies of the related candidate-alleles with respective significance thresholds of 1.139E-12 (on the left) and 1.139E-10 (on the right). Loci which have been identified by both SAM and PGM are displayed in dark green.

This analysis allowed to highlight alleles possibly selected by environmental pressure according to the procedure described in the previous section. The loci to which they belong and their frequency are displayed in Figure 7.5. These loci were extracted from the rejection table 7.6 on page 128 which is presenting alleles implied in the most significant models representing 0.04% of the total processed models. The structure of the table is the same as for table 7.4 on page 125 excepted a few changes. The main change is that the presence of alleles is tested against 118 environmental variables; this means that the numbers displayed within the cells represent the number of environmental variables involved in significant models. Of course, only the ones implied in the most significant models will be identified and detailed later in the analysis. Another change is the red line showing the determining level of significance (repeat of the previous line). Then, the cell in grey highlights a case for which the allele frequency is lower than 1% of the sampled animals. The vertical lines are separating the groups of alleles according to the significance levels. In the consolidated test area in blue, a «1» means that the allele was involved in *n* models[1] significant with both tests. The alleles are displayed from the left to the right first according to a decreasing order of the number of consolidated models (models significant with both tests, line «Total Significance») and then according to the total number of significant models (these models may be significant with only one test, line «Total 2 models»). Thus the alle-

---

1. To know the number of significant models, one must sum the values of the corresponding cells (same significance threshold) for both yellow (G) and orange (Wald) areas.

les (and related loci) the most likely to be under natural selection are positioned on the left of the table. Their naming (line 2) is composed of the name of the locus, whether it is located on homologous chromosome 1 or 2, and the length of the microsatellite.

Table 7.6. Results of the spatial analysis method on microsatellites collected among 48 Econogene sheep breeds. The numbers in yellow and orange cells indicate the number of significant models for the corresponding significance level. Indeed, one rejection table has been calculated per significance level (11 for the G test and 11 for the Wald test). This table is a synthetic view of the 22 rejection tables required to obtain these results. The first-level rejection tables show the results of the tests for each «geoenvironmental variable x allele» model (see example in appendix 14).

As this analysis was also carried out with a PGM using the Fdist2 software (Beaumont and Baldings, 2004) to detect F_{ST}-outlier loci, the line containing the names of the alleles in table 7.6 is providing information about the correspondence between both analyses. In red are displayed alleles which have not been detected by the PGM, in bold black are displayed the best candidates and in black other candidates also detected by both approaches. In grey are displayed alleles which were detected by both methods, but for which the F_{ST}/heterozygosity plot points out that their values are underestimating the neutral distribution (see figure 7.6 on page 129). According to Beaumont and Balding (2004) those low outliers indicate loci subject to balancing selection which is working to maintain genetic polymorphisms (multiple alleles) from a generation to the other within a population. Whereas high outliers suggest adaptive or directional selection which favor a single allele. For instance adaptive selection would rather indicate a population in

which the effects of local climate have fixed one allele, while selection effects are weak or absent at that locus in the environment experienced by the other populations (Beaumont and Balding, 2004). We will come back to these aspects when assessing the distribution of candidate-loci frequencies within breeds and tackling breed effects.

The purple cell is pointing out the *DYMS1* locus (name in blue) which was detected only by the SAM, proven to be involved in parasite resistance (Buitkamp, 1996) and possibly related to an environmental parameter (frequency of precipitations has been highlighted in this analysis, see table 7.7 on page 130). As a relative summary (the threshold is arbitrary chosen), considering a significance threshold of 1.139E-14 allows to reach a convergence rate of 5/11 loci between both approaches.

Fig 7.6. Plot of FST values against heterozygosity estimates for Econogene sheep performed by Fdist2. The red and the green lines include the 99% confidence interval.

Environmental variables

The SAM permits to identify the environmental variables implied in the significant models. Considering the 5 alleles most likely to be selected by environmental pressure, table 7.7 on page 130 is showing climatic parameters involved in the corresponding models. This kind of information is likely to differentiate loci under selection whenever associated with genes playing a role in adaptation processes. Quantitative Traits Loci (QTL) are regions of the genome where one or more gene(s) influencing a trait is (are) located. Once identified and mapped, QTLs can be exploited to understand genetic bases of adaptation (Zeng, 2005) : some can be involved in controlling the resistance to parasites like *DYMS1* mentioned above, others can be responsible for the hair growth rate for instance.

This last example can be interesting in order to improve low temperature tolerance of a given breed as the valuable QTLs can be introgressed from one population into another. In a conservation perspective, this type of adaptive marker may be used to identify an appropriate source population from which to translocate individuals into small declining populations that require supplementation (Luikart *et al.*, 2003).

Allele	Test G : climatic variables	Wald test : climatic variables
SRCRSP9_134	Number of wet days (8 variables), relative humidity (4), sunshine (4)	Number of wet days (3 variables)
DYMS1_181	Number of wet days (4)	Number of wet days (3)
SRCRSP9_118	Number of wet days(1), wind (3)	Number of wet days (2), wind (1)
ILSTS28_127	Wind (1), number of wet days (1)	Number of wet days (1)
OARFCB304_171	Precipitations (1)	Precipitations (1)

Table 7.7. Five alleles constituting the most significant models and possibly under selection. The columns on the right are indicating climatic variables involved in those most significant models. The values between brackets are representing the number of variables involved in the models.

As an illustration, this kind of operation is mentioned about sheep resistance against nematode parasites as the latter have a large impact on the economy of sheep industries. «The identification of genes or linked markers that have a significant association with the variance of indicator traits of internal nematode resistance in sheep would facilitate the inclusion of nematode resistance in sheep breeding operations» (Dominik, 2004).

In our case, the association analysis led to the identification of variables related to rain for the 5 «top» loci (see table 7.7 on page 130). Precipitations in excess can cause problems regarding the wool quality. Sheep are producing a type of wool wax to protect against rain, but it was shown that continuously exposing animals to precipitations may bring to a significant loss of wool wax from the fleece (Hay and Mills, 1982). Moreover, the wax remaining in the fleece after wetting is altered in its chemical composition and assists the penetration by rain causing subsequent development of fleece rot.

Results provided here are perhaps an indication for identifying QTLs implied in wool wax production, or in any other aspect related to the adaptation to a rainy environment. We can imagine a case where the OARFCB304 locus - more specifically the allele OARFCB304_171 - plays a role about the production of effective wool wax. In this case, this valuable QTL could be introgressed from populations identified in figure 7.7 on page 131 to other breeds chronically suffering from fleece rot.

Several QTLs have been identified to play a role regarding wool production, but they are related to fibre diameter, crimp frequency, staple[1] length and strength (Purvis and Franklin, 2005 and references therein) and do not seem to have any relationship with rain or humidity. The present result may contribute to enrich the knowledge base usable to deliver molecular tools likely to facilitate enhanced genetic improvement programs for wool sheep.

Frequencies of detected loci and breed effect

The frequency of an allele identified as an outlier possibly under selection has to be measured within breeds in order to reveal a possible breed effect, or on the contrary show that the allele in question is well distributed among all breeds. A breed effect is likely to highlight the fact that the existence of an allele is related to a spe-

---

1. The fibre of wool graded as to length and fineness.

cific and local environmental stimulation, in which case the phenomenon would also be detectable on a map representing the geographical distribution of the studied marker.

Figure 7.7 shows the frequency of one of the 5 «top» alleles (*OARFCB304_171*) identified by the SAM. Obviously it is present in only 15 breeds with a maximum frequency of 55% for the Shkodrane. Moreover it is apparently especially concentrated in albanian breeds (Bardhoka, Ruda, Shkodrane), and thus shows a kind of regional breed effect.
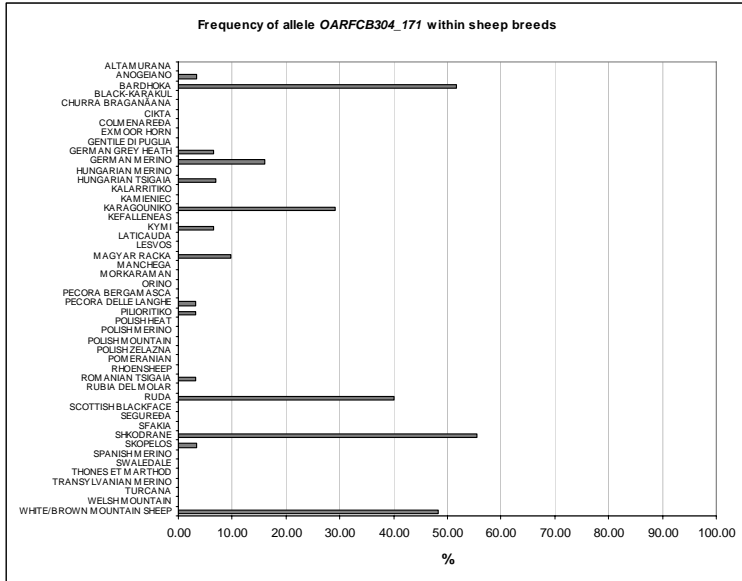
Fig 7.7. Frequency of the allele *OARFCB304_171* within sheep breeds analysed. All three albanian breeds show a frequency of at least 40% for this allele.

The frequency of detected markers within breeds has been analyzed in detail in the next section applied also to sheep but making use of AFLP markers. Histograms have been elaborated for all candidate loci detected with the help of the SAM, and one of them has been represented on a map to check for any geographic particularity.

Working with a different kind of molecular marker, this second case study confirms a convergence between results provided by the SAM and a PGM. Moreover, this application to a domestic species seems to show that despite the direct anthropic influence on breeds location and on their trade movements, the method is still sensitive enough to detect the influence of the environment on organisms. It is true that the example here above illustrate the case of autochthonous breeds only - thus geographically fixed and exposed to given ecological parameters for a long time - but next section will reveal that this observation is also applicable to cosmopolite breeds.

ECONOGENE SHEEP AND AFLPS
....

Sheep data investigated for this analysis were also captured in the context of the Econogene project. 624 animals were sampled representing 40 breeds (see appendix 8), from farms located in 10 european countries. As several animals (up to 3) could be sampled at the same location, we have 577 distinct locations characterized by 92 AFLP band presence/absence. 5% of the frequency distribution corresponding to very low values (markers absent almost everywhere) and 5% with very high values (markers present almost everywhere) were removed because they don't contribute in discriminating environmental information. Thus calculations were made for 62 AFLP markers. The same environmental variables as for microsatellites were used.

The SAM analysis identified 4 AFLP candidate-markers possibly involved in adaptation processes which are appearing in table 7.8. The reading of the table is globally the same as for sheep microsatellites. The only difference is regarding the last line where no color are displayed. Indeed, due to technical constraints[1], it was not possible to make use of Dfdist on the whole population and thus to assess the convergence level between both approaches. As results produced by the SAM were existing, it was decided to nevertheless use Dfdist and to look for outlier loci restrictively among breeds which are known to be differentiated (or contrasted)

---

1. In september 2005, the utility software conceived by Oscar Gagiotti (LECA, UJF, Grenoble) to prepare data to be imported in Dfdist was not able to support the number of samples we had to include in the analysis.

from an environmental point of view, as it was done - among other combinations - for the common frog when looking for outlier loci among high and low altitude populations. Climatic variables allowing to differentiate a wet profile against a dry one (see table 7.9) were used in order to chose contrasted breeds pairs whose genetic information could to be used in order to try to detect outlier loci with Dfdist.

Signif. threshold AFLP marker	Test G : climatic variables	Wald test : climatic variables
1.37E-13		
E35T32_32	Precipitations (1 variable), coefficient of variation of precipitations (1)	Coefficient of variation of precipitations (1 variable)
1.37E-11		
E35T32_32	Precipitations (2 variable), coefficient of variation of precipitations (1)	Precipitations (1), coefficient of variation of precipitations (1)
E35T38_16	Coefficient of variation of precipitations (1), number of wet days (3), relative humidity (5), sunshine (6), mean diurnal temperature range (1)	Number of wet days (1)

Table 7.9. Best candidate-markers possibly under natural selection, and climatic variables involved for two significance thresholds (1.37E-12 has been deliberately omitted).

Among the two contrasted pairs (Manchega-Welsh Mountain, Manchega-Shkodrane), and also in an analysis led across all three breeds, it was not possible to highlight any outlier behavior with a 95% confidence level. Figure 7.8 is showing results for the Manchega-Welsh Mountain contrast, that is the situation for which loci were located the nearer of the limits of the neutral distribution. The candidate-markers identified by the SAM are displayed in blue, the E35T32_32 being the nearest to the threshold.
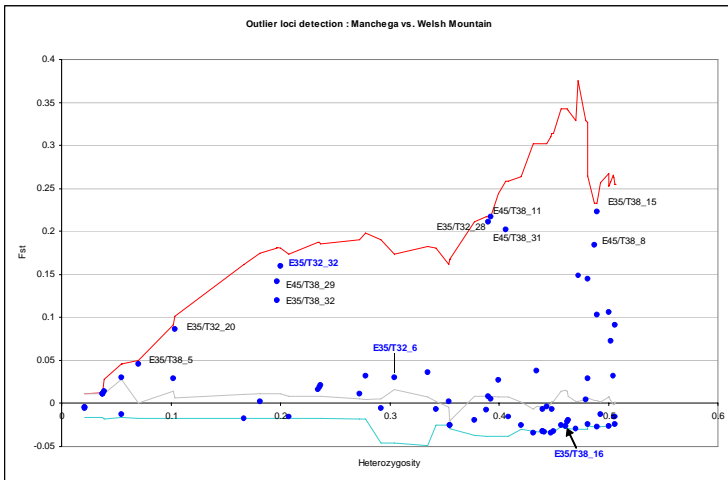
Fig 7.8. Plot of FST values against heterozygosity estimates in Econogene sheep as performed by Dfdist. The red and the green lines include the 99% confidence interval. Markers whose name is displayed in blue are the three most significant loci identified by the SAM.

Frequencies of detected AFLP markers within breeds surveyed

Figure 7.9 shows frequencies of the 4 candidates detected by the SAM within breeds surveyed. The E35T32_32 marker has the higher significance but a relatively low frequency among breeds studied (present in only 7.5% of the animals). Those 43 presences are distributed among the White and Brown Mountain Sheep, Polish Heat, Pecora delle Langhe, Pecora Bergamasca, and the Kameniec sheep. This is highlighting a breed effect deserving to be investigated.
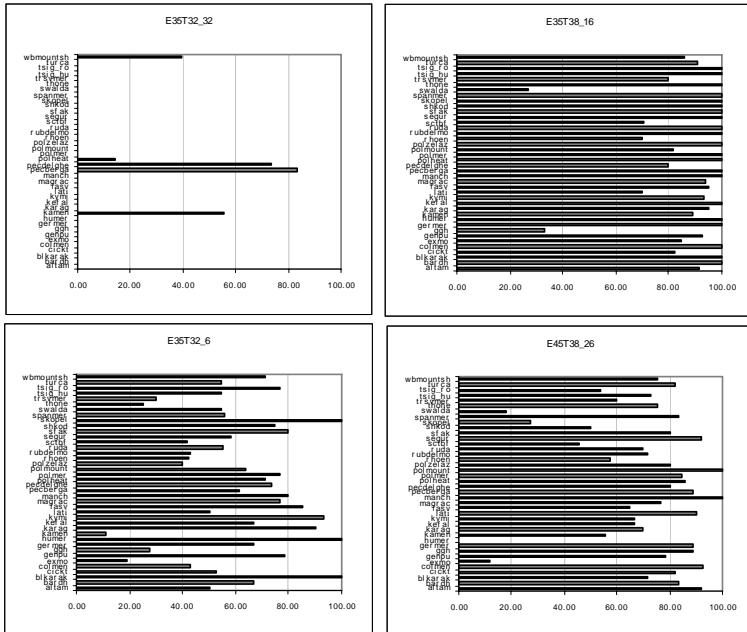
Fig 7.9. Frequencies within studied breeds of the 4 AFLP markers most likely to be under natural selection and identified by the SAM. See appendix 8 for breed abbreviations.

This effect can be deemed according to different points of view : first, and as mentioned in the previous section, it can be perceived as a factor reducing the relevance of the candidate-marker as its presence across the investigated area is restricted to specific zones. But on the other hand, it is precisely revealing where its presence is really necessary for any possible adaptation requirement (see figure 7.10 on page 135). We are possibly in a configuration for which adaptive selection by mean of the effects of local climate have fixed one allele. But selection effects are in this case absent at that locus in the environment experienced by the other populations.

A way to reinforce the relevance of this candidate-locus is to refine the scale of analysis in order to consider solely the region where the candidate-locus is observed, to make use of environmental data with a higher definition, to sample other breeds located within and outside the area in question, and to perform analyses to verify if the behavior of the marker is the same. In fact, I would even say

that this kind of configuration with a candidate-marker present among a few breeds in a localized area only is the best situation to identify a specific role related to adaptation.

In comparison, the other 3 candidate-markers are present within almost all analyzed breeds, Hungarian Merino being the only exception for marker E45T38_26.
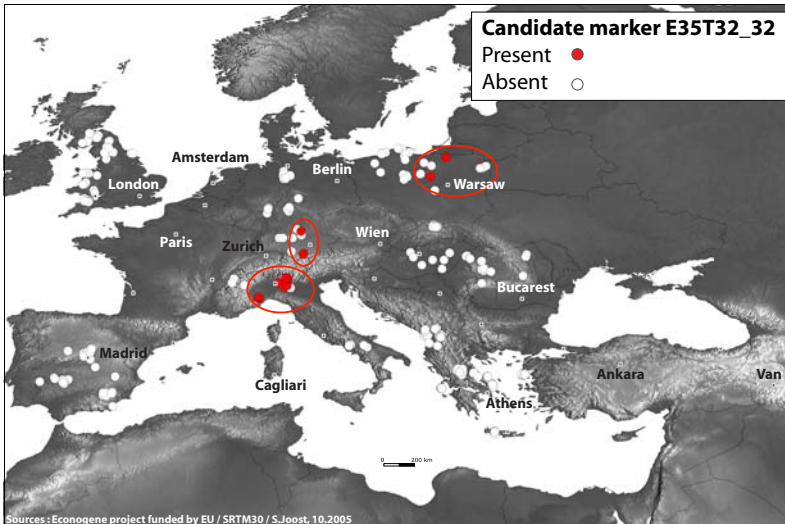
Fig 7.10. Map showing the geographic distribution of the E35T32_32 AFLP candidate-marker. This marker is present only in the Southern of the italian Alps, in Schwarzwald and in a part of polish low lands (red circles). Additional analyses would be important to investigate any ecological common point between those three locations (please refer to figure 5.2 on page 58 for breed identification).

About partial analyses restricted to contrasted populations

The analyses on common frog already showed that it was difficult to detect outlier loci restrictively on contrasted populations because of poorly reliable outliers, of local effect or because of false positives. However 3 outlier loci could nevertheless be identified in the analyses of contrasted populations, and the convergence analysis could be carried out at a global level between SAM and PGM.

In the context of this application to sheep AFLPs, the tentative to detect outlier loci with Dfdist and the restrictive recourse to pairs of contrasted populations failed. The achievement of the neutral distribution simulation on the basis of all 40 analyzed breeds only is likely to confirm the 4 markers highlighted by the SAM and possibly under selection. However, previous results obtained for both common frog and sheep microsatellites let us suppose that some of them are good candidates.

SCANDINAVIAN BROWN BEAR

The present outlier analysis led on Scandinavian Brown Bear microsatellite data was carried out in order to reinforce the SAM results mainly processed on domestic species. To assess its robustness, the method was applied to a wild species also to complete the common frog case study whose results were only based on the very general altitude environmental parameter, expressing combined effects of many other variables like temperature, precipitations, etc. Moreover, the Scandinavian Brown Bear is also an example of a possible alternative use of SAM results as environmental variables implied in the most significant models can be used to build potential habitat maps useful in conservation genetics (see general context on page 14).

The data consist of 17 microsatellite loci extracted from 728 Brown Bears sampled in Scandinavia from 1985 to 2004 (see Waits *et al.*, 2000; Manel *et al.*, 2004; Bellemain *et al.*, 2005). The frequencies of 181 alleles (mean of 10.6 alleles per locus) were compared to ten environmental variables, for a total of 1810 univariate models calculated to search for loci under selection. The yearly mean of the different topo-climatic parameters presented in table 4.1 on page 41 was used. Data sources are the same as described on page 40.
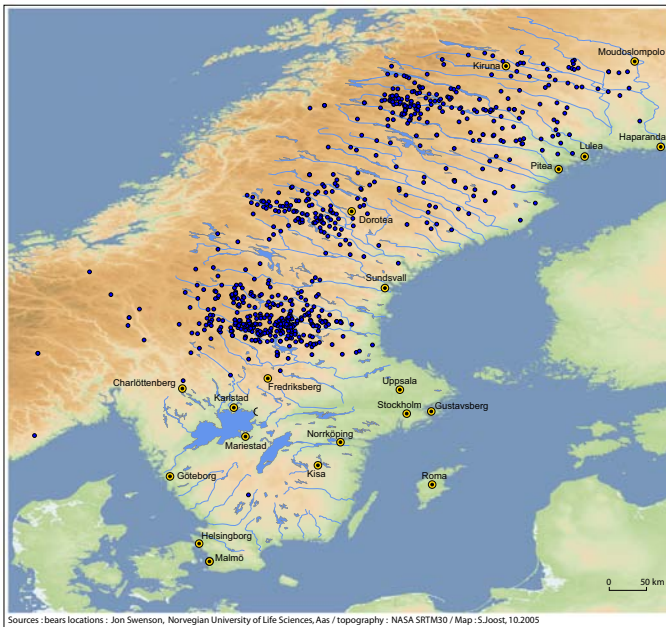
Fig 7.11. In blue, indication of the locations where the 728 Scandinavian Brown Bears used in this study were sampled.

SAM

Table 7.10 is showing the consolidated matrix of the results, that is models significant with both Wald and G tests (see appendix 10 for a comparison of statistical tests applied to this case study). The rejection table is displaying all relevant models at the highest significance threshold (5.52E-17) for which an important number are still significant. Almost half of these 58 pairs are distributed among the 5 first alleles, each of them counting a relevant frequency at sampling places. Table 7.10 also reveals that a wide range of environmental variables are involved in those significant models, from which are always excluded altitude, diurnal temperature range and relative humidity.

Table 7.10. Rejection table showing the 58 most significant models (blue cells) with both Wald **and** G statistical tests. 21 of the 181 alleles are implied in one significant model at least. The blue square points out alleles possibly implied in cold and humidity selection processes, and the red square indicates alleles that are rather related to light and temperature (see text on page 141).

Attention should be paid to the fact that only one significance level is represented in this table. The 58 models presented here are discriminated by allele frequency.

The percent of maximum possible sunshine (percent of day length) and temperature are responsible for 33% and 26 % percent of the significant models respectively, contributing to the detection of 20 out of 21 candidate-alleles (see figure 7.12 a).
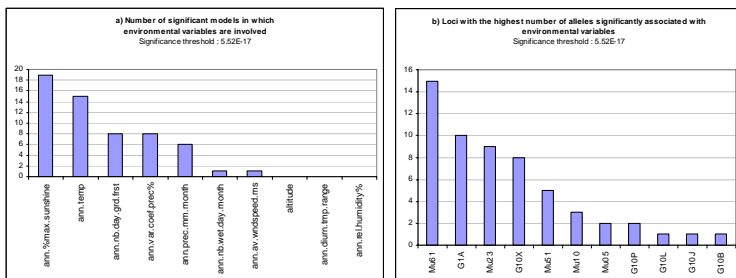
Fig 7.12. a) Number of significant models per environmental variable (left) and b) number of alleles implied in significant models per candidate-locus (right). The 21 alleles involved in the 58 consolidated and significant models shown in table 7.10 were aggregated per single locus to obtain this graph. Example : Mu61 = 7 + 5 + 2 + 1.

137

Aggregated alleles make Mu61 clearly stand out as a serious candidate-locus (see figure 7.12 b) to be involved in adaptation processes, its particularity being to be significantly selected by 7 different environmental variables; this kind of versatile behaviour was not noticed till now in the other case studies. The fact of considering yearly means instead of monthly values possibly favours this comportment, removing a part of the informativeness of the parameters.

Three other loci can be considered as candidate, as they respectively count 10, 9 and 8 alleles implied in significant models : G1A, Mu23 and G10X.

PGM

Fdist was used in order to detect outlier loci among the 17 analysed microsatellites. Unfortunately, it turned out that the number of microsatellites was not sufficient to reveal outliers (Bonin and Manel, personal communication[1]). Nevertheless, the analysis was completed and results supplied. They are shown in figure 7.13. Indeed, it appears that no locus is reaching the limit of the neutral envelope. But interestingly the 4 loci highlighted by the SAM are the same four that stand out in figure 7.13 because of a higher F_{ST}.

This «unfinished» investigation to check for convergence between the SAM and a PGM applied to a wild species nevertheless led to the idea of an alternative use of the results.
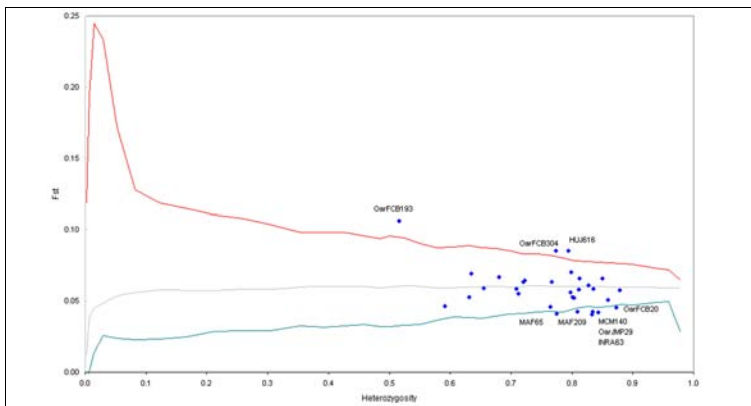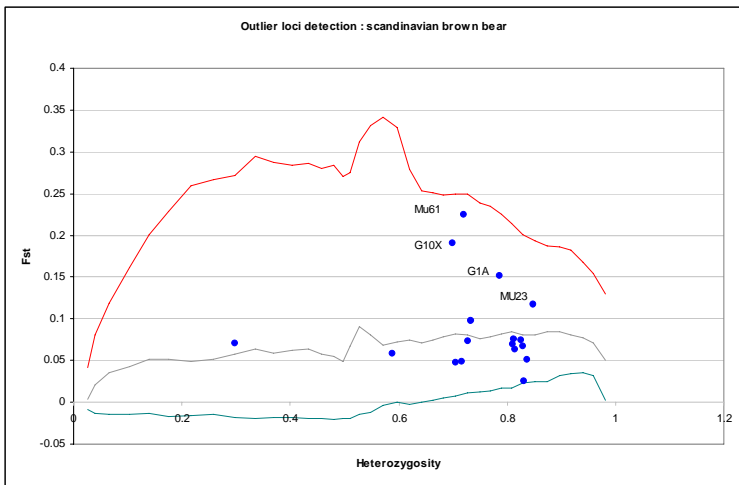

Fig 7.13. Plot of FST values against heterozygosity estimates for the Scandinavian Brown Bear performed by DFdist. The red and the green lines include the 99% confidence interval. The four loci whose name is displayed in blue are the most significant ones identified by the SAM.

Outlier loci detection applied to habitat modelling

In habitat modelling, an initial step consists in defining the different geoenvironmental variables to be included in a predictive model. According to Guisan and Zimmermann (2000), the selection of predictors can be made either arbitrarily, automatically by stepwise procedures[1], by following physiological principles (expert's decision), or by following shrinkage[2] rules (see also Guisan *et al.*, 2002 and references therein). Be that as it may, I propose an alternative to select predicitve variables based on a genetic criterion.

Focusing on environmental parameters involved in significant models

Indeed, it is possible to profit from indications provided by the results of the SAM and to exploit information about the environmental variables which turned out to significantly select parts of the genome analysed, the one of the Brown Bear in this case. As the presence of a series of alleles possibly correspond to their selection, one can postulate that those environmental parameters do play a role favouring Scandinavian Brown Bear presence, and that they participate in constituting an ideal habitat for the animal. According to this reasoning, it is then possible to build a map of the potential habitat of the plantigrade. Thus, the proposed method to select relevant predictive factors is composed of the following steps :

1 Run multiple univariate logistic regression models to assess alleles selectiveness by environmental pressure (see common frog and sheep case studies in this chapter);

2 Identify models for which the presence of an allele is significantly explained by a given geoenvironmental variables;

3 Extract geoenvironmental variables implied in those significant models;

4 Select the most significant among them;

5 Apply habitat modelling techniques on those variables only.

However, one should be certain that the detected candidate-alleles are confirmed to intervene in adaptive processes before modelling wildlife habitats according to this method. In the present case, we only have strong presumption as the PGM couldn't highlight loci with an outlier behaviour; and moreover, all outlier loci detected by this type of approach are not necessarily under selection. Nevertheless, the proposed method was applied to the Brown Bear case study in order to demonstrate the approach.

Application

The habitat modelling example presented hereunder was implemented above all to illustrate the reasoning exposed and not with the aim to perfom analysis : it is therefore perfectible. For instance, as at least 3 bears populations were genetically differentiated (Manel *et al.*, 2004), the most suitable solution would have been to calculate one habitat map per population and not one global map for all three

---

1. A set of rules for deriving a regression equation by adding or subtracting one variable at a time from the regression equation.
2. Also called «Ridge» or «Lasso» regression. «It keeps all the terms in the model, but shrinks their coefficients towards zero using a quadratic penalty term, such as a bound on the sum-of-squares of the coefficients. This has the effect of reducing the variance of the fit of the model, while increasing the bias» (Guisan *et al.*, 2002).

populations. Moreover, as ordinary multiple regression in its generalized form (GLM, see MacCullagh and Nelder, 1989; Guisan and Zimmermann, 2000) was used to process selected variables and to permit the production of maps showing a probability of presence[1], it was necessary to produce absence data to obtain a profile for the places where bears were not observed. The common practice in habitat modelling is to use existing available absence data (Broennimann, 2003) but unfortunately none was available. To fill in this lack, a number of random absences equalling the number of bear presences was generated (balanced method, see Broennimann, 2003). An alternative would have been to apply one of the modelling techniques developed to overcome the drawback of lack of accurate absence data by incorporating presence data only : environmental envelopes, genetic algorithms and ecological niche factor analysis (see Broennimann, 2003 and references therein).
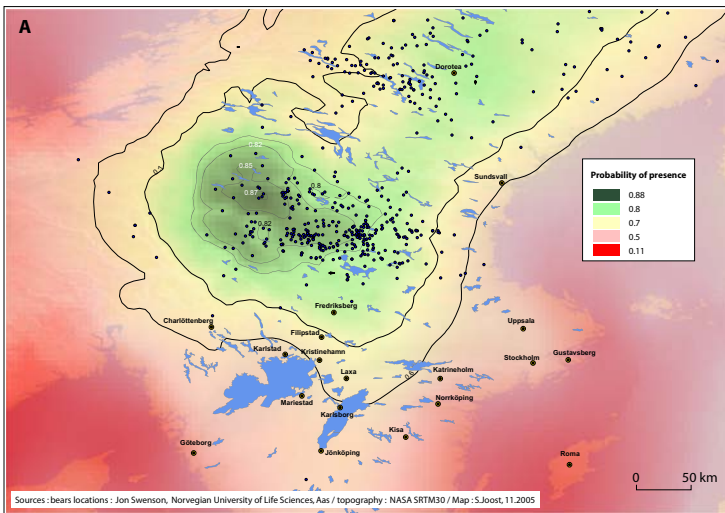
Fig 7.14. A map of the Scandinavian Brown Bear potential habitat calculated with a Generalized Linear Model (GLM) on the basis of 4 environmental variables identified by the SAM. The map corresponds to the area of the Southern population mentioned by Manel *et al.* (2004).

Selection of variables

I used the observed cross-check between the SAM and the PGM (see figure 7.14 on page 141 and figure 7.13 on page 140) to select 4 environmental variables in order to produce a potential habitat map of the Scandinavian Brown Bear. The most relevant way to select significant variables would have been to use the significance threshold criteria and to keep the ones corresponding to the last level. But it turns out that 7/10 variables are corresponding to this criteria. Then it was decided to chose the variables which are bringing about the higher number of alleles significant responses. These environmental parameters are the percent of maximum

---

1. GLMs are calculated on the basis of presences and absences data and it is the reason why «probability of presence» is used. We should talk about the probability of a cell (the corresponding area on the ground) to be the best combination of relevant environmental variables for the studied species to stay.

sunshine, temperature, and then the number of days with ground frost and the coefficient of variation of precipitations (see figure 7.13 on page 140).

Maps

These four selected variables were chosen to elaborate the map shown in figure 7.14 on page 140. Then the probability maps on figure 7.15 have been produced with respectively the first three (left) and the first two (right) environmental variables (according to figure 7.12 on page 137). On those maps, the probability value is displayed according to a cold to warm color scheme, highest probabilities being represented by warm colors.
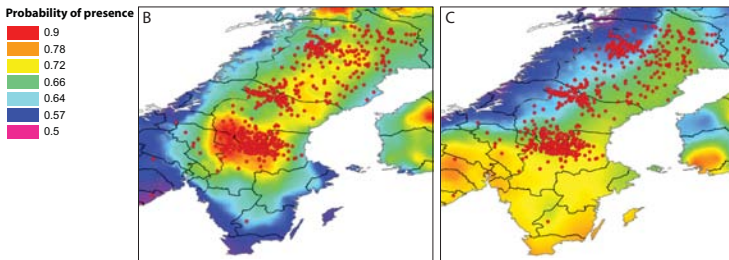
Fig 7.15. Probability that a cell is the most adequate combination of relevant environmental variables for the Scandinavian Brown Bear to stay here. On the left with 3 variables (percent of maximum sunshine, temperature, and number of days with ground frost) and on the right with 2 variables (percent of maximum sunshine and temperature). As a reminder, the resolution of the climatic grids is 10 minutes, thus about 9 kilometers at Stockholm's latitude.

One can on the one hand observe (map B) that the influence of presence data on the probabilities distribution is possibly not balanced enough by random absences. Indeed we can see that the highest probability of presence is where most of the bears were sampled. It is also posssible that the model in B more accurately fits the Southern population's characteristics. On the other hand (map C), keeping only sunshine duration and temperature naturally shifts southwards the optimal potential habitat niches. This map also reinforce the hypothesis of the predominance of criteria fitting the Southern population as probabilities of presence in map C are rather low for the central and the Northern population. Located farther towards Northern areas, those populations may have developed particular abilities with regard to cold (here materialized by the ground frost variable) in comparison with the Southern population more sensitive to heat-related variables. This has to be considered while having a look on table 7.10 on page 137, as it is possible to identify the group of alleles implied in cold/humidity resistance and influencing the appearance of maps A (blue square on figure 7.14 on page 140) and B in comparison with the group of alleles which would be - by deduction - rather related with processes in which light and temperature are intervening (red square on figure 7.14 on page 140). This is endowing the first group of alleles with a more versatile role. Be that as it may, these raw observations strengthen the fact that separate habitat maps should be built for each population.

Despite its elementariness, this example shows that it is worth basing habitat modelling on genetic criteria in the context of conservation actions. It appears that many combinations - involving different alleles supposed to be implied in natural

selection processes - can be used to produce maps supporting various ecological scenarios likely to be evaluated by experts. This constitutes a concrete application showing that the results of the SAM can be used in a different way than «solely» detecting loci under selection.

SUMMARY
....

A Spatial Analysis Method (SAM) used to identify alleles possibly involved in adaptation processes was presented in this chapter. It relies on a precise conception of environmental modelling, mistrusting sophisticated models implying numerous parameters - expression of the present dominant pragmatic realism - and is advocating simple models as the best way to reveal and understand the main dimensions of reality.

This original approach applied to molecular biology turned out to be efficient and supplied results converging on the ones provided by a standard population genomics method. It was possible to observe that the SAM tends to supply an important number of loci possibly under natural selection among which solely the cross-check with population genomics or deepened researches in the existing literature are allowing to detect pertinent candidates.

It is balanced by the fact that the SAM method allows to reveal novel information about the genome : more and more markers can be placed on genome maps and are likely to contribute in identifying candidate-genes. Finally, this method implies less modelling contraints than population genomics ones, also leads to the identification of environmental parameters providing relevant ecological informations, and permits to analyse the molecular «response» beyond the locus level by comparing alleles selectiveness by environmental pressure.

As an alternative application, this last property of the method makes it possible to apply the results to elaborate maps of habitat modelling, and to select the necessary environmental predictors on the basis of this genetic criterion.

TOWARDS LANDSCAPE GENOMICS

ASSESSING THE CONTRIBUTION OF GISCIENCE

This tentative application of diverse facets of GIScience's to problems in molecular biology addressed three distinct issues, namely, the investigation of large genetic spatial data sets, representation of molecular geodata, and detection of natural selection signatures. These demonstrations provide initial answers to the questions raised early (see on page 16) regarding the extent to which an adequate exploitation of its geographic dimension is likely to favour progress in molecular biology.

Exploring the geographic dimension of large genetic data sets

It has long been clear that geography aids our understanding of how genetic resources function (see population genetics first steps on page 22). Most often the analysis of a precise topic (*genetic drift* for instance) revealed spatial characteristics (distance, isolation), which were then taken into account and exploited in subsequent researches.

The exercise conducted in the present research was slightly different from those historical approaches, whereas Exploratory Spatial Data Analysis (ESDA) or Geographic visualization (GVIS) approaches propose to start from the spatial angle to investigate the variation of different molecular data variables.

Exploratory analysis of spatial information facilitates the investigation of data, provides a visual support realized by geo-graphic representations, offers a dynamic and interactive access to both elementary and more sophisticated statistical tools, and finally allows the combination of these two elements, what is even more relevant. Thus, ESDA facilitates the manipulation of data, and favours the emergence of numerous working hypotheses to be quickly verified by multiple possibilities of comparison.

Advanced expertise is, naturally, still required for the understanding of the distribution of genetic resources, and for proposing explanations and interpretations. The application of such exploratory tools to the classification of goat breeds was unilaterally directed from the GIScience perspective, thus being limited itself to rather simple observations and interpretation proposals. In contrast, it is clear that for a biologist who commits a whole day to applying the appropriate methodology, *molecular data wandering* (Banos, 2001) is likely to be a high-performance and

high-return approach. The enthusiasm spatial exploratory tools have already inspired among a few molecular ecologists indicates that this technological appropriation may occur in the near future. Moreover this «event» could also be assisted by the development perspectives presented in the «Perspective overview» section on page 146.

From molecular data visualization to cartography

In a majority of specialized papers or books, the quality of the graphic representation of spatial genetic data and processes is often poor, frequently consisting of raw colorless depictions of population or individual locations. Costs only partly explain this observation.

It is worth applying cartography to molecular biology as its very potential is not yet used to advantage. The development and the generalization of its use on the one hand support spatial analysis in one of its simplest form, thanks notably to an adequate usage of colors and a better recourse to the geographical context. On the other hand, resorting to high-performance cartography clearly improves the transmission of the results and their communication between researchers themselves, or between researchers and the general public. This is of primary importance in the modern context, considering the impacts of genetics research on the civil society (regarding genetically modified organisms for instance).

Landscape genomics

A further question broached the issue of natural selection processes whose prevailing scientific interest and stakes deserve a more detailed discussion.

Owing to the development of biotechnologies, the increasing availability of molecular markers promoted the «outbreak of *population genomics*» (Luikart *et al.*, 2003; Bonin *et al.*, submitted to Molecular Evolution and Ecology). This favoured the development of genome-wide tests for the identification of outlier loci by assessing their compared response to demography and to neutral history of populations. The existing population genomics approaches are restrictively based on former population genetics models and on the neutral theory. The outliers these methods detect are believed to be involved in selection processes; these presumptions are solidly argued (see for instance Beaumont and Balding, 2004) and are firmly established. But environmental information - the effective selective pressure - is not directly[1] taken into account and not included in those models. This is precisely the kind of gap that the Spatial Analysis Method (SAM) proposed in chapter 7 can fill as shown by the results supplied here, which indicate global convergence with results calculated by standard outlier loci detection methods. Selection processes are viewed from the geoenvironmental angle, permitting a highlight of not only loci but also specific ecological parameters. This makes it possible to offer new means of interpreting the role specific regions of the genome may play. However, this aspect can still be improved by testing supplementary environmental parameters and by enriching the existing collections.

---

1. Although environmental information is evoked and discussed in the previously mentioned literature.

The main characteristics of the SAM's functioning and of the results it generates are that :

- it permits the identification of environmental parameters causing loci response;
- it *may* highlight an important number of loci possibly under selection making it necessary to discriminate between good candidates and possible «noise statistics» ones. This was particularly noticeable in the Scandinavian Brown Bear case study where generalized environmental variables (annual means) were used. Thus the method seems very sensitive to variation of environmental parameters;
- the method is independent of any genetic model and thus does not imply modelling constraints like Hardy-Weinberg equilibrium assumption for example, which is far from being verified when considering markers involved in selection processes;
- the SAM allows detection of both loci under balancing selection (underestimating the neutral F_{ST}/heterozygosity distribution) as well as ones submitted to directional (adaptive) selection (overestimation);
- there is, as of yet, no means of differentiating them in the event that it is not possible to cross-check with standard population genetics methods.

Moreover, considering the fact that the SAM relies on large genome scans to identify alleles whose existence at a specific geographic location is selected by a given pressure of environmental variables, one can sense here the advent of *landscape genomics*. The elaboration of this method based on the processing and the comparison of many simultaneous models was made possible thanks to the recent availability of large data sets. The latter are generated by large genome scans, and reveal the evolution of molecular biology towards population genomics. The SAM may also be an adapted statistical tool among others able to «fully exploit the explosion of data that are becoming available for many species» (Luikart *et al.*, 2003).

Regarding the definition formulated by Manel *et al.* (2003), landscape genetics «aims to provide information about the interaction between landscape features and microevolutionary processes, such as *gene flow, genetic drift* and selection». Therefore, one can present landscape genomics as landscape genetics transferred to the genome level[1]. This is potentially the logical evolution of landscape genetics given simultaneous technical improvements in DNA *genotyping*, constant enhancement of computer capabilities regarding processing the immense quantities of information, and interdisciplinary efforts favouring collaborative research in molecular biology, landscape ecology and GIScience.

Which profit for GIScience ?

Interdisciplinarity is peculiar to GIScience. This is a domain whose intrinsic research and development is profitable for any other scientific or applied field resorting to its functionality. Still, this technological enhancement is only made possible by the application of GIScience to the frontier challenges in various disciplines. This is what causes evolution in the techniques of geographical information processing.

---

1. The «microevolutionary processes» mentioned in the landscape genetics definition may be replaced by «evolutionary processes» provided the fact that according to a landscape genomics approach many of these micro-processes are comparable to many landscape features.

This application to molecular biology mainly led to the development of a spatial analysis method bringing into play many simultaneous univariate logistic regression models. The method was first adapted to a precise conception of environmental modelling advocating the implementation of simple models in order to better grasp the function of natural processes, recognizing an inescapable uncertainty. It turned out to be conclusive, showing that research is also an occasion to make use of and to defend fundamental and philosophical approaches and to demonstrate their applicability.

Furthermore, the SAM method suited the particular configuration of thousands of molecular markers spread over Europe, to be compared with thousands of environmental variables. Other domains and problems in environmental modelling certainly correspond to this kind of configuration and could benefit from the spatial analysis in this work. It is clear that the ultimate realisation would consist of a computer engineering phase to implement this multiple GLM[1] processing and assessment procedure into a function usable in a widely distributed GIS software. This is still conceivable in a further stage, but in spite of the encouraging results obtained in this research, the robustness and the relevance of the approach needs at first to be confirmed by means of application to complementary case studies. For example, an appropriate task would consist in testing the approach on a well-known gene.

The general results have revealed relevance in turning to GIS to manage, investigate and show spatial molecular data, as well as to divulge some of the existing environmental footprints within the genome. Highlighting a relationship between an ecological parameter and a specific location within DNA may be a starting point to understand an evolutionary process, providing an information likely to contribute to the understanding of the role of selection in the evolution of genomes and populations. More pragmatically, the detection of this kind of association is really important for gene-assisted selection in farm animals.

Finally, aside from the technical aspects, the main lesson for GIScience is that given the present threat to worldwide animal and plant biodiversity, research will continue with sampling campaigns and further spatial molecular data production. In the context of genetic resource conservation applied to domestic and wild animals as well as to plants, needs for management and analysis of spatial genetic data will rapidly expand. It is thus time to come closer to conservation genetics, to make GIS students aware and attune to this discipline, and to train them with basic skills in molecular biology : there is more work to be done[2].

PERSPECTIVES OVERVIEW

There are two kinds of perspectives in spatial molecular biology that we will quickly review. On the one hand, are applications directly related to the matter investigated in this research, and on the other hand is a different concept of «biol-oGIS» experimentation, whose initial steps have already been completed and whose potential is acutely challenging, especially for the GIScience community.

---

1. Generalized Linear Model.
2. In this perspective, it is also important to develop GIS education for students in molecular biology.

Remote exploratory analysis of spatial data

As mentioned above, GVIS software offer performance tools that promise to be all the more useful if directly mastered by biologists. It was noted during the Econogene project that they could be used as long as a GIS specialist was present. The data format constraints for example and the many features offered make them inefficient for use by biologists, even in an easy-to-use environment. In the meantime, training periods typical of collaborative research projects are insufficient to enable full exploitation of the potential of those exploratory tools. To become truly effective, these applications should be joined by the following :

- available online through dedicated internet platforms;
- easy to reach (such as by URL) and implying no requirement of any additional software implementation;
- simplified enough, easy-to-use, with especially designed functions, sophisticated enough to permit high-level investigations;
- customizable, alterable by developers.

CommonGIS (see on page 59) already proposes a web version but rests on Java technology which requires the installation of a Java Virtual Machine. It was shown not to have been adapted in the present context.

On the other hand, promising work is emerging at the GIS Laboratory[1] of EPFL[2], currently being applied to anthropology by Abram Pointet (unpublished). A dedicated and simplified WebGIS interface is made available, containing only required vectorial geographic and statistical attributive data, together with necessary functions, and supplying dynamic and interactive maps and graphs generated by the user. In addition, the display of a geographical context is possible, transforming the GVIS application in a real «middleware», also producing geographic ephemeral views for display of neat - printable - thematic maps.

This kind of shared development will make it possible to fully profit from molecular biology specialists in the early stage of analysis, possibly identifying more relevant observations that could enhance on-going and planned investigations. Its application in the context of any future research project involving both GIS scientists and molecular biologists should reveal many lessons to be drawn.

Spatio-temporal modelling

It has not escaped our notice that the very high variability of microsatellite markers (see on page 48) coupled with a spatial context is an ideal case study for spatio-temporal modelling in GIScience. Indeed microsatellite alleles change rapidly and evolve over time, from generation to generation. The study of genetic spatial patterns together with their variation in time is likely to provide relevant elements to progress in the understanding of evolution processes. This could be for instance the study of a microsatellite *mutation* rate (time) according to wild populations movings in the landscape, even taking into account the characteristics of the geographical space. It is supposed to involve a whole and important current of GIScience in a concrete and stimulating application.

---

1. http://lasig.epfl.ch (30.11.2005).
2. Ecole Polytechnique Fédérale de Lausanne or Swiss Federal Institute of Technology.

Moving the georeference system

Until now we solely evoked coordinate systems whose reference was geographic. Making use of the GIS technology, a research team of the National Center for Geographic Information and Analysis[1] (NCGIA, University of Maine, USA) developed a Genome Spatial Information System[2] (GenoSIS) at the end of the 1990s. This «GIS» for within-genome analysis used a system where the chromosome defines a genome coordinate space. The project was an application of the concepts and tools of geographic and spatial information science to the interpretation and modelling of genome data, with a major assumption being that the organization of genome features has biological significance.

According to the same reasoning, that is to place a reference coordinate system within the organism, *epigenetics* is a field possibly offering an application domain to specialists in GIS technologies. Epigenetics studies how gene regulatory information that *is not expressed in DNA sequences* is transmitted from one generation of cells or organisms to the next. Among other topics, it was discovered that the location of chromosomes in the cell's nucleus is influential in that chromosomal function and that the position of a gene within a chromosome *territory* may «influence its access to the machinery responsible for specific nuclear functions, such as transcription and splicing» (Cremer, 2001). This is not a matter of geography, but spatial association processes are involved and it may prove to be relevant, in another interdisciplinary attempt, to apply approaches of spatial analysis to the epigenetics context.

A NEW APPROACH TO RESPOND TO A RECENT PLEA

The present period seems to be adequate for calling molecular biology concepts and techniques into question. In 2004, Karl Woese published in the Microbiology and Molecular Biology Review a major paper entitled «A new biology for a new century» (Woese, 2004; see also Dyson, 2005) in which he states that the «traditional» reductionist biology - the molecular paradigm based on genes and molecules - was henceforth obsolete. He advocates a more invigorating biology «seeking a new and inspiring vision of the living world» rather than a biology continuing to follow its usual comfortable path[3]. A major change in the way to tackle the discipline is likely to engender progress.

Latterly, and in a similar way, Michel Morange also claims in «Les secrets du vivant, contre la pensée unique en biologie» (Morange, 2005) that new approaches and important innovations in molecular biology are required in order to make «une nouvelle lumière[4]» emerge. For him, those important innovations will only

---

1. http://www.ncgia.maine.edu/~cbult/project.html (18.11.2005)
2. http://gis.esri.com/library/userconf/proc02/pap0719/p0719.htm (18.11.2005)
3. «A society that permits biology to become an engineering discipline, that allows that science to slip into the role of changing the living world without trying to understand it, is a danger to itself. Modern society knows that it desperately needs to learn how to live in harmony with the biosphere. Today more than ever we are in need of a science of biology that helps us to do this, shows the way. An engineering biology might still show us how to get there; it just doesn't know where "there" is» (Woese, 2004).
4. Literally «a new light».

appear in the context of collaborative and interdisciplinary efforts, thanks to the interaction of contributions from several different disciplines revolving around physics and biology. It is a long time that chemistry, physics and biology (including all its derivative fields) are involved together to make life sciences progress. And the comment about the necessity of involving various disciplines and different approaches in the study of life is not new : in 1932 Niels Bohr already advocated the application of his complementarity principle when exposing the implications of the new physics[1] on the common perception of the nature of living beings (Morange, 2003)[2].

Experiences show that interdisciplinary objectives often fail due to several reasons discussed by Sillitoe (2004)[3]. But it may also work[4], and then GIScience thanks to the unifying dimension of geography - or of spatial processes to include every possible configuration - is an auxiliary but useful discipline to contribute in any impending major progress in molecular biology. The intrusion of such an unexpected field on the comfortable path of reductionist biology is hopefully a contribution to this major change demanded by Woese.

---

1. At this time, recent results obtained in quantum mechanics.
2. Bohr, N. (1933) Light and life, Nature, vol. 131, pp. 421-423 and 547-459.
3. For example one can mention the tendency of researchers to be superficial in their own discipline to keep it accessible to partners.
4. Indeed molecular biology is already an example of a success of interdisciplinarity as this discipline is the result of efforts in (bio)chemistry, genetics, biology, and physics.

TOWARDS LANDSCAPE GENOMICS
A new approach to respond to a recent plea

AFTERWORD

The last paragraph of the conclusion demands a personal comment about the question of a paradigm change raised by Karl Woese. It makes no doubt that in order to become really efficient, this kind of external novelty like the discussed GIScience «impingement» requires to come along with a molecular biology more drastic self-questioning.

There is a gap between the plea expressed by Woese or Morange and the present reality of research. Let us consider the neutral theory (Kimura, 1968) on which research in population genetics is resting since almost 40 years. For a GIS scientist immersed into molecular genetics, it looked like a systematic brake put on many avenues for research proposed. Well, the neutral theory appeared to me like an old unwavering dogma, whereas Darwin's much older works on natural selection and adaptation didn't even disturb me...

I'm not a biologist and I'm probably ignoring a lot of elements likely to make such a speech somewhat out of place. But I was really surprised by this general inertia in comparison with new proposals well suited to Woese's demand. For example, the idea formulated by Freeman Dyson is probably intentionally rather provocative, but I cannot resist the temptation to describe it here. Dyson is professor emeritus of physics at the Institute for Advanced Study in Princeton, New Jersey. His research has focused on the internal physics of stars, subatomic-particle beams, and the origin of life. He considers the Darwinian era to be over (Dyson, 2005). Indeed, on the basis of Woese's proposal to answer the question to know when did Darwinian evolution begin (Woese, 2004), Dyson argues for a cycle going from a pre-Darwinian «bio-marxism» - an age during which horizontal gene transfer was universal and separate species did not exist - to a post-Darwinian period, through «The Darwinian Interlude» when selection was operating. Dyson considers that the present domination of *Homo sapiens* on the biosphere materializes the return to a period of horizontal transfer because the cultural evolution which is implied - also evoked by Christian de Duve (2005) - has replaced the biological one as the driving force of change. This cultural evolution is not Darwinian, because cultures are spread by horizontal transfer of ideas more than by genetic inheritance...

This is really likely to stir up the hornets' nest !

Literature cited

· ·
· · · ·
· · ·

Adams, J., Kelly, B. and Waits, L., (2003) Using faecal DNA sampling and GIS to monitor hybridization between red wolves (Canis rufus) and coyotes (Canis latrans), Molecular Ecology, 12 (8):2175-2186*.

Ajmone-Marsan, P., (2005) 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, FAO, Torino, pp. 89-96.

Ajmone-Marsan, P., Negrini, R., Milanesi, E., Bozzi, R., Nijman, I.J., Buntjer, J.B., Valentini, A., et al., (2002) Genetic distances within and across cattle breeds as indicated by biallelic AFLP markers, Animal Genetics, 33 (4):280-286.

Ajmone-Marsan, P., Negrini, R., Crepaldi, P., Milanesi, E., Gorni, C., Valentini, A. and Cicogna, M., (2001) Assessing genetic diversity in Italian goat populations using AFLP (R) markers, Animal Genetics, 32 (5):281-288.

Ajmone-Marsan, P., Valentini, A., Cassandro, M., Vecchiotti-Antaldi, G., Bertoni, G. and Kuiper, M., (1997) AFLP (TM) markers for DNA fingerprinting in cattle, Animal Genetics, 28 (6):418-426.

Apps, C.D., McLellan, B.N., Woods, J.G. and Proctor, M.F., (2004) Estimating grizzly bear distribution and abundance relative to habitat and human influence, Journal of Wildlife Management, 68 (1):138-152.

Arnaud, J.-F., (2003) Metapopulation genetic structure and migration pathways in the land snail Helix aspersa: influence of landscape heterogeneity, Landscape Ecology, 18:333-346.

Aspinall, R.J., (1999) GIS and landscape conservation, In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds), Geographical Information Systems, John Wiley & Sons, New York, pp. 967-980.

Avise, J.C., (2004) Molecular Markers, Natural History, and Evolution, Sinauer, Sunderland.

Bamshad, M. and Wooding, S.P., (2003) Signatures of natural selection in the human genome, Nature Reviews Genetics, 4:99-111.

Banos, A., (2001) A propos de l'analyse spatiale exploratoire des données, Cybergeo (197), http://193.55.107.45/MODELIS/banos/article.htm (last consulted on the 30.11.2005).

Beaumont, M.A., (2005) Adaptation and speciation: what can F-st tell us?, Trends in Ecology & Evolution, 20 (8):435-440.

Beaumont, M.A. and Balding, D.J., (2004) Identifying adaptive genetic divergence among populations from genome scans, Molecular Ecology, 13 (4):969-980.

Beaumont, M.A. and Nichols, R.A., (1996) Evaluating loci for use in the genetic analysis of population structure, Proceedings of the Royal Society of London Series B-Biological Sciences, 263 (1377):1619-1626.

Bellemain, E., Swenson, J.E., Tallmon, O., Brunberg, S. and Taberlet, P., (2005) Estimating population size of elusive animals with DNA from hunter-collected feces: Four methods for brown bears, Conservation Biology, 19 (1):150-161.

153

Berry, R.J., (1989) Ecology : where genes and geography meet, Journal of Animal Ecology, 58:733-759.

Bertaglia, M., (2004) Livestock biodiversity conservation: the case of sheep and goat breeds in european marginal areas, Institut für Ernährungswirtschaft und Verbrauchslehre, Christian-Albrechts University, Kiel.

Bertin, J., (1983) Semiology of graphics : diagrams, networks, maps, University of Wisconsin Press, Madison, London.

Beven, K.J., (2002) Towards a coherent philosophy for environmental modelling, Proceedings of the Royal Society of London, Royal Society of London, London, pp. 2465-2484.

Beven, K.J., (2001) Rainfall-Runoff Modeling: The Primer, Wiley, Chichester.

Beven, K.J., (1998) Generalised Likelihood Uncertainty Estimation (GLUE), About GLUE, Environmental Science Department at Lancaster University, Lancaster.

Beven, K.J. and Binley, A.M., (1992) The future of distributed models - model calibration and uncertainty prediction, Hydrological processes 6(3):279-298.

Bian, L., (1997) Multiscale Nature of Spatial Data in Scaling up Environmental Models, In: Goodchild, M.F., Quattrochi, D.A. (eds), Scale in Remote Sensing and GIS, Lewis Publishers, Boca Raton, pp. 13-26.

Blott, S.C., Williams, J.L. and Haley, C.S., (1999) Discriminating among cattle breeds using genetic markers, Heredity, 82:613-619.

Boichard, D., Le Roy, P., Levéziel, H. and Elsen, J.-M., (1998) Utilisation des marqueurs moléculaires en génétique animale, INRA Productions Animales, 11:67-80.

Bonin, A., Taberlet, P., Miaud, C., Pompanon, F., (In press) Explorative genome scan to reveal adaptive divergence along a gradient of altitude in the Common Frog (Rana temporaria), Molecular Ecology and Evolution.

Bowmaker, J.K. and Dartnall, H.J.A., (1980) Visual pigments of rods and cones in a human retina, Journal of Physiology, 298:501-511.

Brewer, C.A., (1999) Color Use Guidelines for Data Representation, Proceedings of the Section on Statistical Graphics, American Statistical Association, Alexandria VA, pp. 55-60.

Brodlie, K., (1994) A typology for scientific visualization, In: Hearnshaw, H.M. and Unwin, D.J. (eds), Visualization in Geographical Information Systems, Wiley, pp. 34-41.

Broennimann, O., (2003) Modelling the potential distribution of rare and endangered plant species, Department of Ecology and Evolution, University of Lausanne, Lausanne.

Brooke, C.H. and Ryder, M.L. (eds), (1979) Declining breeds of mediterranean sheep, Food and Agriculture Organization of the United Nations (FAO), Rome.

Bruford, M. and the Econogene Consortium, (2005) Strategies for integrating husbandry, genetics, geographic and socio-economic data for sustainable conservation, International Workshop on the role of biotechnology for the characterisation and conservation of crop, forestry, animal and fishery genetic resources, FAO, Torino, pp. 117-120.

Bruford, M., Bradley, D. and Luikart, G., (2003) DNA markers reveal the complexity of livestock domestication, Nature Reviews Genetics, 4 (11):900-910.

Brunet, R., (1987) La carte mode d'emploi, Fayard, Paris.

Bucci, G. and Vendramin, G., (2000) Delineation of genetic zones in the European Norway spruce natural range: preliminary evidence, Molecular Ecology, 9 (7):923-934.

Buitkamp, J., Filmether, P., Stear, M. and Epplen, J., (1996) Class I and class II major histocompatibility complex alleles are associated with faecal egg counts following natural, predominantly Ostertagia circumcincta infection, Parasitology Research, 82 (693).

Burrough, P.A. and McDonnell, R.A., (1998) Principles of Geographical Information Systems, Clarendon, Oxford.

Caloz, R., (2005) Réflexions sur les incertitudes et leurs propagations en analyse spatiale, Revue Internationale de Géomatique, 15 (3):303-319.

Caloz, R. and Collet, C., (1997) Geographic information systems (GIS) and remote sensing in aquatic botany: methodological aspects, Aquatic Botany, 58 (3/4):209-228.

Caloz, R., Puech, C., (1996) Hydrologie et imagerie satellitaire, In: Bonn, F. (eds), Précis de télédétection, applications thématiques, Presses de l'Université du Québec, Sainte-Foy.

Cavalli-Sforza, L., Menozzi, P., Piazza, A., (1994) The history and geography of human genes, Princeton University Press, Princeton, New Jersey.

Clinton, B., (1996) President Bill Clinton accepts his nomination at the Democratic National Convention, Chicago, http://www.pbs.org/newshour/convention96/floor_speeches/clinton_8-29.html (consulted on the 30.11.2005).

Conway Morris, S., (2003) Life's solution, Inevitable humans in a lonely universe, Cambridge University Press, Cambridge, UK.

Cramer, J.S., (2002) The Origins of Logistic Regression, 4 (119), Tinbergen Institute.

Cremer, T. and Cremer, C., (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells, Nature Reviews Genetics, 2 (4):292-301.

Crow, J.F., (1986) Basic concepts in population, quantitative and evolutionary genetics, W.H. Freeman and Company, New York.

Cyranoski, D., (2004) Gene-ecology agreement circles the globe, Nature, 428 (6):6.

Dangermond, J., (1993) The role of software vendors in integrating GIS and environmental modeling, In: Goodchild, F., Parks, B.O., Steyaert, L.T. (eds), Environmental Modeling with GIS, Oxford University Press, New York.

Darwin, C., (1989) The voyage of the Beagle, Charles Darwin's journal of researches, Penguin Books, London.

Darwin, C., (1985) On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, Penguin Classics, London.

Dawkins, R., (2004) The devil's chaplain, Phoenix, London.

Dawkins, R., (1989) The selfish gene, Oxford University Press, Oxford.

de Duve, C., (2005a) La dynamique du hasard contraint, Les dossiers de La Recherche, Mai (19):22-26.

de Duve, C., (2005b) A l'écoute du vivant, Odile Jacob, Paris.

de Saussure, F., (1995) Cours de linguistique générale, Payot, Paris.

Domingos, P., (2000) Bayesian Averaging of Classifers and the Overfitting Problem, 17th International Conference on Machine Learning, Center for the Study of Language and Information, Standford University, Stanford, pp. 223-230.

Dominik, S., (2005) Quantitative trait loci for internal nematode resistance in sheep: a review, Genetics Selection Evolution, 37, pp. 83-96

Dyson, F.J., (2005) The Darwinian Interlude, The Technology Review.com (March), http://www.technologyreview.com (consulted on the 30.11.2005).

Dyson, F.J., (1999) The Sun, the Genome and the Internet : tools of scientific revolutions, Oxford University Press, Inc., New York.

Eco, U., (1985) How culture conditions the colours we see, In: Blonsky, M. (eds), On signs, The John Hopkins University Press, Baltimore.

Eddy, J.A., (1993) Environmental Research,what we must do, In: Goodchild, F., Parks, B.O., Steyaert, L.T. (eds), Environmental Modeling with GIS, Oxford University Press, New York.

Epperson, K.E., (2003) Geographical Genetics, Princeton University Press, Princeton.

Ertz, O., Joost, S. and Rappo, D., (2001) Towards Geoservices Portals MEDIAMAPS: WGIS Trends for Business Applications, In: Claramunt, C., Winiwarter, W., Kambayashi, Y. and Zhang, Y. (eds), Second International Conference on Web Information Systems Enginee-ring (WISE 2001), IEEE Computer Society Press, Kyoto, pp. 102-108.

Escudero, A., Iriondo, J. and Torres, M., (2003) Spatial analysis of genetic diversity as a tool for plant conservation, Biological Conservation, 113 (3):351-365.

Farnir, F., Coppieters, W., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., et al., (2000) Extensive genome-wide linkage disequilibrium in cattle, Genome research, 10 (2):220-227.

Faucher, D., (2002) Initiation à la cartographie automatique, Département de géographie, PRCD, Université de Toulouse Le Mirail, Toulouse.

Fedra, K., (1993) GIS and environmental modeling. In: Goodchild, F., Parks, B.O., Steyaert, L.T. (eds), Environmental Modeling with GIS, Oxford University Press, New York, pp. 36-50.

Friendly, M. and Denis, D.J., (2005) Milestones in the history of thematic cartography, statistical graphics, and data visualization, Statistical Consulting Service, York University, Canada, http://www.math.yorku.ca/SCS/Gallery/milestone/ (consulted on the 30.11.2005).

Garon, J.M., (2001) VT Conference Puts New Research Area on the Map; GIS Expert Michael Goodchild Echoes Its Value., Virginia Tech, Blacksburg (VA), https://www.vbi.vt.edu/arti-cle/articleview/48/1/15/ (consulted on the 30.11.2005).

Goodchild, M.F., Haining, R.P., (2004) GIS and spatial analysis : Converging perspectives, Papers in Regional Science, 83:363-385.

Goodchild, M.F., Quattrochi, D.A., (1997) Scale, Multiscaling, Remote Sensing, and GIS, In: Goo-dchild, M.F., Quattrochi, D.A. (eds), Scale in Remote Sensing and GIS, Lewis Publishers, Boca Raton, pp. 1-11.

Goodchild, M.F., (1996) Geographic Information Systems and spatial analysis in the social scien-ces, In: Aldenderfer, M. and Maschner, H.D.G. (eds), Anthropology, space, and Geogra-phic Information Systems, Oxford University Press, New York, pp. 214-250.

Goodchild, F., Parks, B.O., Steyaert, L.T. (1993) Environmental Modeling with GIS, Oxford Uni-versity Press, New York.

Goodchild, M.F., (1992) Geographical information science, International Journal of Geographical Information Systems, 6 (1):31-45.

Gould, S.J., (2002) I have landed, Harmony books, New York.

Gould, S.J., (1996) The mismeasure of man, W.W. Norton & Company, New York.

Gould, S.J., (1994) Curveball, The New Yorker (28.11.1994).

Gould, S.J., (1989) Wonderful Life, W.W. Norton & Company, New York.

Grant, V., (2003) Incongruence between cladistic and taxonomic systems, American Journal of Botany, 90 (9):1263-1270.

Greene, S., Gritsenko, M. and Vandemark, G., (2004) Relating morphologic and RAPD marker variation to collection site environment in wild populations of red clover (Trifolium pra-tense L.), Genetic resources and crop evolution, 51 (6):643-653.

Guarino, L., Jarvis, A., Hijmans, R.J., Maxted, N., (2002) Geographic Information Systems (GIS) and the Conservation and Use of Plant Genetic Resources, In: Engels, J., Ramanatha Rao, R., Brown, A.H.D., Jackson, M., (eds), Managing Plant Diversity, CABI Publishing, Cambridge, p. 512.

Guisan, A., Edwards, T.C. and Hastie, T., (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene, Ecological Modelling, 157 (2-3):89-100.

Guisan, A., Zimmermann, N., (2000) Predictive habitat distribution models in ecology, Ecological Modelling, 135:147-186.

Haining, R., (2003) Spatial data analyis, theory and practice, Cambridge University Press, Cambridge.

Hamann, A., Aitken, S.N., Yanchuk, A.D., (2004) Cataloguing in situ protection of genetic resources for major commercial forest trees in British Columbia, Forest Ecology and Management, 197:295-305.

Hamann, A., Koshy, M., Namkoong, G. and Ying, C., (2000) Genotype x environment interactions in Alnus rubra: developing seed zones and seed-transfer guidelines with spatial statistics and GIS, Forest Ecology and Management, 136 (1-3):107-119.

Hamann, A., El-Kassaby, Y., Koshy, M. and Namkoong, G., (1998) Multivariate analysis of allozymic and quantitative trait variation in Alnus rubra: geographic patterns and evolutionary implications, Canadian Journal of Forest Research, 28 (10):1557-1565.

Hanotte, O. and Jianlin, H., (2005) Genetic characterization of livestock populations and its use in conservation decision-making, International Workshop on the Role of Biotechnology for the Characterisation and Conservation of Crop, Forestry, Animal and Fishery Genetic Resources., The FAO Working Group on Biotechnology, Torino, pp. 131-136.

Hardy, O.J. and Vekemans, X., (2002) SPAGeDI : a versatile computer program to analyse spatial genetic structure at the individual or population levels, Molecular Ecology Notes, 2:618-620.

Hay, J.B. and Mills, S.C., (1982) Chemical changes in the wool wax of adult Merino sheep during prolonged wetting and prior to development of fleece rot, Australian Journal of Agricultural Research, 33 (2):335-346.

Hieter, P., Boguski, M., (1997) Functional Genomics: It's All How You Read It, Science, 278 (5338):601-602.

Hijmans, R.J., Guarino, L., Cruz, M., Rojas, E., (2001) Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS, Plant Genetic resources Newsletter, 127:15-19.

Hijmans, R., Schreuder, M., De la Cruz, J. and Guarino, L., (1999) Using GIS to check co-ordinates of genebank accessions, Genetic Resources and Crop Evolution, 46 (3):291-296.

Hirao, A.S., Kudo, G., (2004) Landscape genetics of alpine-snowbed plants: comparisons along geographic and snowmelt gradients. , Heredity, 93 (3):290-298.

Hirzel, A., Hausser, J., Chessel, D., Perrin, N., (2002) Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data?, Ecology, 83 (7):2027-2036.

Hirzel, A., Helfer, V., Metral, F., (2001) Assessing habitat-suitability models with a virtual species, Ecological Modelling, 145:111-121.

Hirzel, A., (2001) When GIS come to life. Linking landscape- and population ecology for large population management modeling: the case of Ibex (Capra ibex) in Switzerland, Ph.D. Thesis, Departement of Ecology and Evolution, University of Lausanne, Lausanne.

Hoffmann, M.H., Glass, A., Tomiuk, J., Schmuths, H., Fritsch, R.M., Bachmann, K., (2003) Analysis of molecular data of Arabidopsis thaliana (L.) Heynh. (Brassicaceae) with Geographical Information Systems (GIS), Molecular Ecology, 12:1007-1019.

Hosmer, D.W., Lemeshow, S., (2000) Applied logistic regression, John Wiley & Sons, New York.

Houle, J.L., Cadigan, W., Henry, S., Pinnamanenib, A. and Lundahl, S., (2000) Database Mining in the Human Genome Initiative, Amita Corporation, http://www.biodatabases.com/whitepaper01.html (consulted on the 30.11.2005).

Hunsaker, C.T., Nisbet, R.A., Lam, D.C.L., Browder, J.A., Baker, W.L., Turner, M.G., Botkin, D.B., (1993) In: Goodchild, F., Parks, B.O., Steyaert, L.T. (eds), Environmental Modeling with GIS, Oxford University Press, New York.

Hussy, C., (1998) Signifier and signified: Between insignificance and operability, Semiotica, 122 (3-4):297-308.

Inselberg, A., (1985) The Plane with Parallel Coordinates, The Visual Computer, 1:69-91.

Jaccard, P., (1908) Nouvelles recherches sur la distribution florale, Bulletin de la Société Vaudoise des Sciences Naturelles, 44:223-270.

Jacquard, A., (1998) L'équation du nénuphar, Calmann-Lévy, Paris.

Jacquez, G.M., Greiling, D.A., Kaufmann, A., (In press) Design and implementation of space-time information systems, Journal of Geographical Systems.

Jacquez, G., (2004) Current practices in the spatial analysis of cancer: flies in the ointment, International Journal of Health Geographics, 3 (22):1-10.

Jaynes, E.T., (2003) Probability Theory : The Logic of Science, Cambridge University Press, Cambridge.

Jelinski, D., (1997) On genes and geography: a landscape perspective on genetic variation in natural plant populations, Landscape and Urban Planning, 39 (1):11-23.

Ji, W., Leberg, P., (2002) A GIS-Based approach for assessing the regional conservation status of genetic diversity: an example from the Southern Appalachians, Environmental Management, 4 (29):531-544.

Jones, P.G., Guarino, l., Jarvis, A., (2002) Computer tools for spatial analysis of plant genetic resources data: 2. FloraMap, Plant Genetic resources Newsletter, 130:1-6.

Jones, P.G., Beebe, S.E., Tohme, J., (1997) The use of geographical information systems in biodiversity exploration and conservation, Biodiversity and Conservation, 6:947-958.

Joost, S. and the Econogene Consortium, (2005) Combining biotechnologies and GIScience to contribute to sheep and goat genetic resources conservation., International Workshop on the role of biotechnology for the characterisation and conservation of crop, forestry, animal and fishery genetic resources, FAO, Torino, pp. 109-116.

Karp, A., Edwards, K.J., Bruford, M., Funk, S., Vosman, B., Morgante, M., Seberg, O., et al., (1997) Molecular technologies for biodiversity evaluation: Opportunities and challenges, Nature Biotechnology, 15 (7):625-628.

Karp, A., Seberg, O. and Buiatti, M., (1996) Molecular Techniques in the Assessment of Botanical Diversity, Annals of Botany, 78:143-149.

Kidd, D.M., Ritchie, M.G., (2000) Inferring the patterns and causes of geographic variation in Ephippiger ephippiger (Orthoptera, Tettigoniidae) using geographical information systems (GIS), Biological Journal of the Linnean Society, 71:269-295.

Kimura, M., (1991) Recent Development of the Neutral Theory Viewed from the Wrightian Tradition of Theoretical Population-Genetics, Proceedings of the National Academy of Sciences of the United States of America, 88 (14):5969-5973.

Kimura, M., (1968) Genetic Variability Maintained in a Finite Population Due to Mutational Production of Neutral and Nearly Neutral Isoalleles, Genetical Research, 11 (3).

Lassueur, T., (2004) Modélisation spatiale de l'habitat d'espèces végétales: apports du modèle numérique d'altitude à très haute résolution, diploma work, SSIE, EPFL, Lausanne.

Lassueur, T., Joost, S. and Randin, C., (Accepted) Very high resolution digital elevation models: do they improve models of plant species distribution? Ecological Modelling.

Latter, B.D.H., (1973) The island model of population differentiation : a general solution, Genetics, 73:147-157.

Leach, S. and de Gennes, P.-G., (2005) Je suis un physico-chimico-biologiste !, La recherche (387):59-62.

Lehmann, A., Lachavanne, J.B., (1997) Geographic information systems and remote sensing in aquatic botany Aquatic Botany, 58:195-207.

Lehmann, A., Jaquet, J.-M., Lachavanne, J.-B., (1997) A GIS approach of aquatic plant spatial heterogeneity in relation to sediment and depth gradients, Lake Geneva, Switzerland, Aquatic Botany, 58:347-361.

Lenstra, J.A. and the Econogene Consortium, (2005) Evolutionary and demographic historay of sheep and goats suggested by nuclear, mtDNA and Y-chromosome markers, International Workshop on the role of biotechnology for the characterisation and conservation of crop, forestry, animal and fishery genetic resources, FAO, Torino, pp. 97-100.

Levin, S.A., (1992) The problem of pattern and scale in ecology, Ecology, 73 (6):1943-1967.

Loftus, R.T., Ertugrul, O., Harba, A.H., El-Barody, M.A.A., MacHugh, D.E., Park, S.D.E. and Bradley, D.G., (1999) A microsatellite survey of cattle from a centre of origin: the Near East, Molecular Ecology, 8:2015-2022.

Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., (2001) Geographic Information Systems and Science, Wiley, Chichester.

Loytonen, M., (1998) GIS, time geography and health, In: Gatrell, T., Loytonen, M. (eds), GIS and health, Taylor and Francis, London.

Luikart, G., England, P.R., Tallmon, D., Jordan, S. and Taberlet, P., (2003) The power and promise of population genomics: From genotyping to genome typing, Nature Reviews Genetics, 4 (12):981-994.

Luikart, G., Gielly, L., Excoffier, L., Vigne, J.-D., Bouvet, J. and Taberlet, P., (2001) Multiple maternal origins and weak phylogeographic structure in domestic goats, Proceedings of the National Academy of Sciences of United States of America, 98:5927-5932.

Luria, S., Neri, D.F. and Jacobsen, A.R., (1986) The effect of set size on color matching using SRT displays, Human Factors, 28 (1):49-61.

Lush, J.L., (1948) The Genetics of Populations. Original and Revised Notes (1948-94). Iowa State University Press, Ames.

MacArthur, R.H. and Wilson, E.O., (2001) The theory of island biogeography, Princeton University Press, Princeton.

MacCullagh, P. and Nelder, J.A., (1989) Generalized Linear Models, Chapman & Hall/CRC, London.

MacEachren, A.M. and Kraak, M.-J., (2001) Research challenges in geovisualization, Cartography and Geographic Information Science, 28 (1).

MacEachren, A.M., (1995) How maps work : representation, visualization and design, Guildford Press, New York.

MacEachren, A.M. and Taylor, D.R.F., (1994) Visualization in modern cartography, Pergamon, Oxford.

MacHugh, D.E. and Bradley, D.G., (2001) Livestock genetic origins: goats buck the trend, Proceedings of the National Academy of Sciences of the United States of America, 98 (10):5382-5384.

Mackey, B.G., (1996) The role of GIS and environmental modelling in the conservation of biodiversity, 3rd International Conference on Integrating GIS and environmental modeling, National Center for Geographic Information and Analysis, Santa Fe.

MacKendry, J.E., Machlis, G.E., (1991) The role of geography in extending biodiversity gap analysis, Applied Geography, 11:135-152.

MacNoleg, O., (2003) An account of the origins of conceptual models of geographic space, Computers, Environment and Urban Systems (27):1-3.

MacNoleg, O., (1998) Professor Oleg McNolegs Guide to the successful use of Geographical Information Systems, International Journal of Geographical Information Science, 12 (5):429-430.

MacQueen, J.B., (1967) Some Methods for classification and Analysis of Multivariate Observations, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, pp. 281-297.

Maidment, D.R., (1996) Environmental Modeling within GIS, In: Goodchild, F., Steyaert, L.T., Parks, B.O., et al. (eds), GIS and Environment Modeling : Progress and Research Issues, GIS World Books, Fort Collins.

Manel, S., Bellemain, E., Swenson, J.E., François, O., (2004) Assumed and inferred spatial structure of populations : the Scandinavian brown bears revisited, Molecular Ecology (13):1327-1331.

Manel, S., Schwartz, M., Luikart, G. and Taberlet, P., (2003) Landscape genetics: combining landscape ecology and population genetics, Trends in Ecology & Evolution, 18 (4):189-197.

Mayr, E. and Bock, W.J., (2002) Classifications and other ordering systems, Journal of Zoological Systematics and Evolutionary Research, 40 (4):169-194.

Michels, E., Cottenie, K., Neys, L., De Gelas, K., Coppin, P., De Meester, L., (2001) Geographical and genetic distances among zooplankton populations in a set of interconnected ponds: a plea for using GIS modelling of the effective geographical distance, Molecular Ecology, 10 (8):1929-1938.

Moazami-Goudarzi, K. and Laloe, D., (2002) Is a multivariate consensus representation of genetic relationships among populations always meaningful?, Genetics, 162 (1):473-484.

Moonan, P.K., Bayona, M., Quitugua, T.N., Oppong, J., Dunbar, D., Jost, K.C., Burgess, G., Singh, K.P., Weis, S.E., (2004) Using GIS technology to identify areas of tuberculosis transmission and incidence, International Journal of Health Geographics, 3 (23):1-10.

Morange, M., (2005) Les secrets du vivant, contre la pensée unique en biologie, La Découverte, Paris.

Morange, M., (2003) Histoire de la biologie moléculaire, La Découverte, Paris.

Morgenthaler, S., (1997) Introduction à la statistique, méthodes mathématiques pour l'ingénieur, Presses Polytechniques et universitaires romandes, Lausanne.

Mourant, A., (1954) The distribution of the human blood groups, Blackwell Scientific, Oxford.

Mueller, U.G. and Wolfenbarger, L.L., (1999) AFLP genotyping and fingerprinting, Trends in Ecology & Evolution, 14 (10):389-394.

Mukesh, M., Sodhi, M., Bhatia, S. and Mishra, B.P., (2004) Genetic diversity of Indian native cattle breeds as analysed with 20 microsatellite loci, Journal of Animal Breeding and Genetics, 121 (6):416.

Munch, Z., Van Lill, S.W.P., Booysen, C.N., Zietsman, H.L., Enarson, D.A., Beyers, N., (2003) Tuberculosis transmission patterns in a high-incidence area: a spatial analysis, International Journal of Tuberculosis and Lung Disease, 7 (3):271-277.

Negrini, R., Joost, S., Milanesi, E., Bernardi, J., Pellecchia, M., Patrini, M., Caloz, R., et al., (2004) Genetic diversity of european goats as measured by AFLP markers, European Association for Animal Production (EAAP) 55th meeting, Bled, Slovenia.

Pakniyat, H., Powell, W., Baird, E., Handley, L.L., Robinson, D., Scrimgeour, C.M., Nevo, E., Hackett, C.A., Caligari, P.D.S., Forster, B.P., (1997) AFLP variation in wild barley (Hordeum spontaneum C. Koch) with reference to salt tolerance and associated ecogeography, Genome, 3 (40):332-341.

Paquet, S., (2005) Towards Solving the Interdisciplinary Language Barrier Problem, http://www.iro.umontreal.ca/~paquetse/knoweb/000_INTRODUCTION.html, (consulted on the 20.02.2006).

Palumbi, S.R., Gaines, S.D., Leslie, H., Warner, R.R., (2003) New wave: high-tech tools to help marine reserve research, frontiers in Ecology and the Environment, 1 (2):73-79.

Patthey, P., (2003) Habitat and corridor selection of an expanding red deer (*Vervus elaphus*) population, Ph.D. Thesis, Departement of Ecology and Evolution, University of Lausanne, Lausanne.

Parks, B.O., (1993) The need for integration. In: Goodchild, F., Parks, B.O., Steyaert, L.T. (eds), Environmental Modeling with GIS, Oxford University Press, New York, pp. 31-34.

Pellegrini, P., (2005) De l'idée de race animale et de son évolution dans le milieu de l'élevage, Ruralia, http://ruralia.revues.org/document112.html (consulted on the 30.11.2005).

Petit, R.J., Bialozy, R., Brewer, S., Cheddadi, R. and Comps, B., (2001) From spatial patterns of genetic diversity to postglacial migration processes in forest trees, In: Silvertown, J. and Antonovics, J. (eds), Integrating ecology and evolution in a spatial context, British Ecological Society, pp. 295-318.

Pickles, J., (1997) Tool or science? GIS, technoscience, and the theoretical turn, Annals of the Association of American Geographers, 87 (2):363-372.

Popper, K.R.S., (2002) The logic of scientific discovery, Routledge, London.

Pringle, H., (1998) Reading the Signs of Ancient Animal Domestication, Science, 282:1448.

Pritchard, J.K., Stephens, M. and Donnelly, P., (2000) Inference of population structure using multilocus genotype data, Genetics, 155 (2):945-959.

Purvis, I.W. and Franklin, I.R., (2005) Major genes and QTL influencing wool production and quality: a review, Genetics Selection Evolution, 37:S97-S107.

Rao, C.R., (1973) Linear Statistical Inference and Its Application, Wiley, New York.

Rappo, D., (1996) Cartographie thématique et sémiologie graphique, Cours de Systèmes d'Information Géographique, University of Lausanne, Lausanne.

Rappo, D., (1994) Géomatique et infographie : la problématique de l'intermédiaire "géo-graphique", Institute of Geography, University of Lausanne, Lausanne.

Ray, N., (2005) PATHMATRIX: a geographical information system tool to compute effective distances among samples, Molecular Ecology Notes, 5 (1):177-180.

Reichert, P. and Omlin, M., (1997) On the usefulness of overparameterized ecological models, Ecological Modelling, 95 (2-3):289-299.

Reynolds, J., Weir, B.S. and Cockerham, C., (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance, Genetics Society of America, 105:767-779.

Richardson, M., van Lill, S.W.R., van der Spuy, G.D., Munch, Z., Booysen, C.N., Beyers, N., van Helden, P.D., Warren, R.M., (2002) Historic and recent events contribute to the disease dynamics of Beijing-like Mycobacterium tuberculosis isolates in a high incidence region, International Journal of Tuberculosis and Lung Disease, 6 (11):1001-1011.

Ritchie, M., Kidd, D. and Gleason, J., (2001) Mitochondrial DNA variation and GIS analysis confirm a secondary origin of geographical variation in the bushcricket Ephippiger ephippiger (Orthoptera : Tettigonioidea), and resurrect two subspecies, Molecular Ecology, 10 (3):603-611.

Roekaerts, M., (2002) The Biogeographical Regions Map of Europe, European Environment Agency, Copenhagen http://dataservice.eea.eu.int/ (consulted on the 30.11.2005).

Romney, D., Thorne, P., Lukuyu, B. and Thornton, P., (2003) Maize as food and feed in intensive smallholder systems: management options for improved integration in mixed farming systems of east and Southern Africa, Field Crops Research, 84 (1-2):159-168.

Rose, S., (2003) Lifelines : Life beyond the Genes, Oxford University Press Inc., New York.

Rose, S., (2001) Moving on from old dichotomies : beyond nature-nurture towards a lifeline perspective, British Journal of Psychiatry, 178:s3-s7.

Schellenberg, J.A., Newell, J.N., Snow, R.W., Mung'ala, V., Marsh, K., Smith, P.G., Hayes, R.J., (1998) An analysis of the geographical distribution of severe malaria in children in Kilifi District, Kenya, International Journal of Epidemiology, 27 (2):323-329.

Scott, J.M., (1993) A geographical approach to protection of biological diversity, Wildlife-Monographs, 123:1-41.

Shaffer, J.P., (1995) Multiple Hypothesis Testing, Annual Review of Psychology, 46:561-584.

Sheppard, E., McMaster, R.B., (2004) Scale and Geographic inquiry : Nature, Society, and Method, Blackwell Publishing, Malden.

Sillitoe, P., (2004) Interdisciplinary exprexiences : working with indigenous knowledge in development, Interdisciplinary Science Reviews, 29 (1):6-23.

Skøt, L., Hamilton, N.R.S., Mizen, S., *et al.*, (2002) Molecular genecology of temperature response in Lolium perenne: 2. association of AFLP markers with ecogeography, Molecular Ecology 9 (11):1865-1876.

Slocum, T.A., McMaster, R.B., Kessler, F.C. and Howard, H.H., (2005) Thematic cartography and geographic visualization, Pearson Prentice Hall, Upper Saddle River.

Smith, J.M. and Szathmary, E., (1999) The origins of life : from the birth of life to the origin of language, Oxford University Press, Oxford.

Smith, P. and Gaffney, P., (2000) Toothfish stock structure revealed with DNA methods, NIWA Water & Atmosphere, 8(4).

Sokal, R.R., Oden, N.L., Rosenberg, M.S., Thomson, B.A., (2000) Cancer incidences in Europe related to mortalities, and ethnohistoric, genetic, and geographic distances, Proceedings of the National Academy of Sciences, 97:6067-6072.

Spear, S.F., Peterson, C.R., Matocq, M.D., Storfer, A., (2005) Landscape genetics of the blotched tiger salamander (Ambystoma tigrinum melanostictum), Molecular Ecology, 14:2553–2564.

Stoelhorst, J.W., (2002) The Naturalist View of Universal Darwinism: An Application to the Evolutionary Theory of the Firm, EAEPE 2002 Conference, European Association for Evolutionary Political Economy / Social Science Research Network Electronic Library, Aix-en-Provence.

Subramanian, S., (1995) The Story in Our Genes, The Time Magazine (16.1.1995).

Suzuki, D.T., Griffiths, A.J.F., Miller, J.H. and Lewontin, R.C., (1991) Introduction à l'analyse génétique, De Boeck, Bruxelles.

Taberlet, P., Swenson, J.E., Sandegren, F. and Bjarvall, A., (1995) Localization of a Contact Zone between 2 Highly Divergent Mitochondrial-DNA Lineages of the Brown Bear Ursus-Arctos in Scandinavia, Conservation Biology, 9(5):1255-1261.

Tait, N., Durr, P.A., Zheng, P., (2004) Linking R and ArcGIS - Developing a spatial statistical tool-kit for epidemiologists, GISVET 04 Conference, Department of Epidemiology of the Veterinary Laboratories Agency at Weybridge, England, University of Guelph, Ontario.

Thrupp, L.A., (1998) Linking Biodiversity and Agriculture: Challenges and Opportunities for Sustainable Food Security. World Resources Institute, Washington DC, March 1997.

Tobon, C., (2001) Visual and interactive exploration of point data Centre for Advanced Spatial Analysis, UCL, Working Paper (Series Paper 31).

Troy, C.S., MacHugh, D.E., Bailey, J.F., Magee, D.A., Loftus, R.T., Cunningham, P., Chamberlain, A.T., *et al.*, (2001) Genetic evidence for Near-Eastern origins of European cattle, Nature, 410:1088-1091.

Tufte, E.R., (1983) The visual display of quantitative information, Graphics Press, Cheshire.

Tukey, J.W., (1977) Exploratory Data Analysis Addison-Wesley, Reading.

Valavanis, V.D., Georgakarakos, S., Kapantagakis, A., Palialexis, A., Katara, I., (2004) A GIS environmental modelling approach to essential fish habitat designation, Ecological Modelling, 178:417-427.

Vallejo, R.L., Li, Y.L. and Rogers, G.W., (2003) Genetic diversity and background linkage disequilibrium in the North-American Holstein cattle population, Journal of Dairy Science, 86 (12):4137-4147.

Vitalis, R., Dawson, K., Boursot, P. and Belkhir, K., (2003) DetSel 1.0: A computer program to detect markers responding to selection, Journal of Heredity, 94 (5):429-431.

Vitalis, R., Dawson, K. and Boursot, P., (2001) Interpretation of variation across marker loci as evidence of selection, Genetics, 158 (4):1811-1823.

Vuilleumier, S., (2003) Dispersal modelling : integrating landscape features, behaviour and meta-populations, Institute of Environmental Science and Technology (ISTE), Thesis no 2878, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne.

Vuilleumier, S. and Metzger, R., (In press) Animal dispersal modelling : handling landscape features and related animal choices, Ecological Modelling.

Vuilleumier, S. and Perrin, N., (In press) Effects of cognitive abilities on metapopulation connectivity, Oikos.

Waits, L., Taberlet, P., Swenson, J.E., Sandegren, F., Franzen, R., (2000) Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (Ursus arctos), Molecular Ecology (9):421-431.

Walton, M., (2005) Finding the roots of modern humans (14.04.2005), CNN http://cnnstudentnews.cnn.com/2005/TECH/science/04/12/genographic/ (consulted on the 30.11.2005).

Watts, P., Rouquette, J., Saccheri, J., Kemp, S. and Thompson, D., (2004) Molecular and ecological evidence for small-scale isolation by distance in an endangered damselfly, Coenagrion mercuriale, Molecular Ecology, 13 (10):2931-2945.

Wei, J., (2002) A GIS-Based approach for assessing the regional conservation status of genetic diversity : an example from the Southern Appalachians, Environmental Management, 29 (4):531.544.

Wiens, J.A., (1989) Spatial scaling in ecology, Functional Ecology, 3:385-397.

Wiens, J.A., Milne, B.T., (1989) Scaling of landscapes in landscape ecology, or, landscape ecology from a beetles perpective, Landscape Ecology, 3 (2):87-96.

Wilson, P.J., Grewal, S., Rodgers, A., Rempel, R., Saquet, J., Hristienko, H., Burrows, F., Peterson, R., White, B.N., (2003) Genetic variation and population structure of moose (Alces alces) at neutral and functional DNA loci, Canadian Journal of Zoology, 81 (4):670-683.

Woese, C.R., (2004) A new biology for a new century, Microbiology and Molecular Biology Reviews, 68 (2):173-186.

Wonnacott, T.H. and Wonnacott, R.J., (1995) Statistique, Economica, Paris.

Wood, M., (1994) The traditional map as a visualization technique, In: Hearnshaw, H.M. and Unwin, D.J. (eds), Visualization in Geographical Information Systems, Wiley, pp. 9-17.

Wright, D., Goodchild, M. and Proctor, J., (1997) Demystifying the persistent ambiguity of GIS as "tool" versus "science", Annals of the Association of American Geographers, 87 (2):346-362.

Wright, S., (1951) The genetical structure of populations, Annals of Human Genetics, 15:323-354.

Zeder, M.A. and Hesse, B., (2000) The initial domestication of goats (Capra hircus) in the Zagros mountains 10,000 years ago, Science, 287 (5461):2254-2257.

Zeng, Z.B., (2005) QTL mapping and the genetic basis of adaptation: recent developments, Genetica, 123 (1-2):25-37.

* Indications of pages are given for papers and for book sections. The number of pages of books is not appearing.

GLOSSARY

. . . .
. . .
. . .
. . .

GENETIC TERMS

The following concise definitions are based on Suzuki (1991), and on the Glossary of Genetic Terms, Division of Intramural Research, National Human Genome Research Institute : http://www.genome.gov/glossary.cfm (25.11.2005). When not based on these sources, the reference is mentioned after the definition.

Allele

One of the variant forms of a gene at a particular ***locus***, or location, on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant one) may be expressed more than another form (the recessive one).

AnGR

Animal Genetic Resources. FAO suggests to develop and use more AnGR to identify and understand the genetic resources of each important farm animal species, and to prioritize and conserve unique AnGR.

Centimorgan (cM)

A measure of genetic distance that tells how far apart two genes are. Generally one centimorgan equals about 1 million base pairs.

Conservation biology

A branch of biology that is concerned with preserving genetic diversity in plants and animals. This scientific field evolved to study the complex problems surrounding habitat destruction and species protection. The objectives of conservation biologists are to understand how humans affect biodiversity and to provide potential solutions that benefit both humans and non-human species.

DNA (Deoxyribonucleic Acid)

The chemical inside the nucleus of a cell that carries the genetic instructions for making living organisms. Deoxyribonucleic acid is a nucleic acid which carries genetic instructions for the biological development of all cellular forms of life and many viruses. During reproduction, it is replicated and transmitted to offspring. Most of the DNA is found in the chromosomes, which are located in the cell nucleus.

Enzyme

A ***protein*** that encourages a biochemical reaction, usually speeding it up. Organisms could not function if they had no enzymes.

Gamet

A gamete is a haploid cell that has half the genetic information than its parent cell possessed. When two join up, for instance human male sperm and female and egg gametes, the genetic information is linked together to make a diploid zygote.e

Gene flow

Gene flow is the transfer of genes from one population to another.

Genetic drift

Genetic drift is a mechanism of evolution that acts in concert with natural selection to change the characteristics of species over time. It is a stochastic effect that arises from the role of random sampling in the production of offspring. Like selection, it acts on populations, altering the frequency of alleles and the predominance of traits amongst members of a population, and changing the diversity of the group. Drift is observed most strongly in small populations and results in changes that need not be adaptive.

Gene

The functional and physical unit of heredity passed from parent to offspring. Genes are pieces of DNA, and most genes contain the information for making a specific *protein*.

Genetics

The study of heredity, or how the characteristics of living things are transmitted from one generation to the next.

Genetic marker

A segment of DNA with an identifiable physical location on a chromosome and whose inheritance can be followed. A marker can be a gene, or it can be some section of DNA with no known function. Because DNA segments that lie near each other on a chromosome tend to be inherited together, markers are often used as indirect ways of tracking the inheritance pattern of a gene that has not yet been identified, but whose approximate location is known.

Genome

All the DNA contained in an organism or a cell, which includes both the chromosomes within the nucleus and the DNA in mitochondria.

Genomics

Genomics is the study of an organism's genome and the use of the genes. It deals with the systematic use of genome information, associated with other data, to provide answers in biology, medicine, and industry.

Genotype

The genotype is the specific genetic make-up of an individual, usually in the form of DNA. It codes for the *phenotype* of that individual. Typically, one refers to an individual's genotype with regard to a particular gene of interest and, in polyploid individuals, it refers to what combination of alleles the individual carries.

Genotyping

Testing that reveals the specific alleles inherited by an individual.

Haploid

The number of chromosomes in a sperm or egg cell, half the diploid number.

Haplogroup

A collection of closely related haplotypes.

Haplotype

A set of closely linked alleles (genes or DNA polymorphisms) inherited as a unit. A contraction of the phrase «haploid genotype». Different combinations of polymorphisms are known as haplotypes. Collectively the results from several loci could be referred to as a haplotype. «Haplo» comes from the Greek word for «single».

Locus (plural = loci)

The place on a chromosome where a specific gene is located, a kind of address for the gene.

Mantel test

This randomization test allows a measure of correlation between dissimilarity matrices and evaluates the significance of the statistic. By permuting rows and columns in one of the matrices, it determines the distribution of the measure of association. It is used when problems involve the consideration of possible relationships between distance matrices and when observations are not independent and spatially autocorrelated.

Meiosis

The process of cell division in sexually reproducing organisms that reduces the number of chromosomes in reproductive cells from diploid to haploid, leading to the production of gametes in animals and spores in plants.

Microsatellites

Repetitive stretches of short sequences of DNA used as genetic markers to track inheritance in families.

Mitochondrial (mt)DNA

The genetic material of the mitochondria, the organelles that generate energy for the cell. MtDNA is typically passed on only from the mother during sexual reproduction. This means that there is little change in the mtDNA from generation to generation, unlike nuclear DNA which changes by 50% each generation.

Molecular biology

Molecular biology is the study of biology at a molecular level. The field overlaps with other areas of biology, particularly genetics and biochemistry. Molecular biology chiefly concerns itself with understanding the interactions between the various systems of a cell, including the interrelationship of DNA, *RNA* and *protein* synthesis and learning how these interactions are regulated.

Molecular genetics

Molecular genetics is the field of biology which studies the structure and function of genes at a molecular level. Molecular genetics employs the methods of genetics and molecular biology. It is so-called to differentiate it from other sub fields of genetics such as ecological genetics and population genetics.

Molecular marker

DNA sequences that can be identified by a simple assay, allowing the presence or absence of neighbouring stretches of the genome to be inferred.

Mutation

A permanent structural alteration in DNA. In most cases, DNA changes either have no effect or cause harm, but occasionally a mutation can improve an organism's chance of surviving and passing the beneficial change on to its descendants.

Nucleotide

One of the structural components, or building blocks, of DNA and *RNA*. A nucleotide consists of a base (one of four chemicals: adenine, thymine, guanine, and cytosine) plus a molecule of sugar and one of phosphoric acid.

Population genetics

Population genetics is the study of the distribution of and change in allele frequencies under the influence of the five evolutionary forces: natural selection, genetic drift, mutation, migration and nonrandom mating. It also takes account of population subdivision and population structure in space. As such, it attempts to explain such phenomena as adaptation and speciation
[Gillespie, J. (1998) Population Genetics: A Concise Guide, Johns Hopkins Press]

Populations genomics

Large-scale comparison of DNA sequences in comparison with population genetics, [Jorde, L.Bb., *et al.*, (2001) Population genomics: a bridge from evolutionary history to genetic medicine, Human Molecular Genetics, Vol. 10, No. 20 2199-2207], «an organism's complete genetic code».

PCR (Polymerase Chain Reaction)

A fast, inexpensive technique for making an unlimited number of copies of any piece of DNA. Sometimes called «molecular photocopying,» PCR has had an immense impact on biology and medicine, especially genetic research.

Phenotype

The observable traits or characteristics of an organism, for example hair color, weight, or the presence or absence of a disease. Phenotypic traits are not necessarily genetic.

Phylogeography

Field of study concerned with principles and processes governing the geographic distribution of genealogical lineages, especially those within and among closely related species.

Protein

A large complex molecule made up of one or more chains of amino acids. Proteins perform a wide variety of activities in the cell.

RNA

Ribonucleic acid (RNA) is a nucleic acid consisting of a string of covalently-bound nucleotides. It is biochemically distinguished from DNA by the presence of an additional hydroxyl group, attached to each pentose ring; as well as by the use of uracil, instead of thymine. RNA transmits genetic information from DNA (via transcription) into proteins (by translation).

SNP (Single Nucleotide Polymorphism)

DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. SNPs promise to significantly advance our ability to understand and treat human disease [http://www.biochem.northwestern.edu/holmgren/Glossary/index.html (5.12.2005)].

Y chromosome

The male sex chromosome in species in which males have two sex chromosomes that differ to one another.

GISCIENCE TERMS

The following definitions are based the on-line dictionary of GIS terms developed by the Association for Geographic Information and the University Of Edinburgh Department of Geography (http://www.geo.ed.ac.uk/agidict/welcome.html), «A Practitioner's Guide to GIS Terminology» by Stearns J. Wood (http://www.geocities.com/gisdatawest), and http://www.gisdevelopment.net/glossary/ (07.02.2006).

Arc second

An arc second, or a second of arc, is specified as a unit of angular measure, especially in astronomy and in global positioning. It is equal to exactly 1/3600 of an angular degree or 1/1'296'000 of a circle. Sixty arc seconds comprise an arc minute; 60 arc minutes comprise an angular degree. One arc second of latitude at the earth's surface corresponds to a North-South distance of about 31 m.

Digital Elevation Model (DEM)

The digital cartographic representative of the surface of the earth or a subsurface feature through a series of three-dimensional coordinate values: a continuous variable over a two-dimensional surface by a regular array of z values referenced to a common datum. Digital elevation models are typically used to represent terrain relief; a model of terrain relief in the form of a matrix consisting of a data file of a topographic surface arranged as a set of regularly spaced X,Y,Z coordinate locations where Z represents surface elevation. Abbreviated DEM. The grid is defined by identifying one of its corners (lower left usually), the distance between nodes in both the X and Y directions, the number of nodes in both the X and Y directions, and the grid orientation.

Geodetic reference system

The true technical name for a datum. The combination of an ellipsoid, which specifies the size and shape of the earth, and a base point from which the latitude and longitude of all other points are referenced.

GPS

Global Positioning System. A network of radio-emitting satellites deployed by the US Department of Defense. Ground-based GPS receivers can automatically derive accurate surface coordinates for all kinds of GIS, mapping, and surveying data collection.

Raster

A method for the storage, processing and display of spatial data. Each given area is divided into rows and columns, which form a regular grid structure. Each cell must be rectangular in shape, although not necessarily square. Each cell within this matrix contains an attribute value as well as location co-ordinates. The spatial location of each cell is implicitly contained within the ordering of the matrix, unlike a vector structure which stores topology explicitly. Areas containing the same attribute value are recognized as such, however, raster structures cannot identify the boundaries of such areas as polygons. Also raster structures may lead to increased storage in certain situations, since they store each cell in the matrix regardless of whether it is a feature or simply 'empty' space.

Standardization

Data Standardization is the process of making all data of the same type or class conform to an established convention or procedure to ensure consistency and comparability across different databases. This is especially important and necessary in a data warehouse environment that contains information from many sources. Without data standardization, no relationship can be established between the various data sources to produce reports that include information from multiple data sets within the data warehouse.

Topology

The relative location of geographic phenomena independent of their exact position. In digital data, topological relationships such as connectivity, adjacency and relative position are usually expressed as relationships between nodes, links and polygons. For example, the topology of a line includes its from- and to-nodes, and its left and right polygons.

Topology is useful in GIS because many spatial modelling operations don't require co-ordinates, only topological information. For example, to find an optimal path between two points requires a list of the lines or arcs that connect to each other and the cost to traverse each line in each direction. Co-ordinates are only needed for drawing the path after it is calculated.

Vector

An abstraction of the real world where positional data is represented in the form of co-ordinates. In vector data, the basic units of spatial information are points, lines and polygons. Each of these units is composed simply as a series of one or more co-ordinate points, for example, a line is a collection of related points, and a polygon is a collection of related lines.

Voronoi polygon

A polygon bounding the region closer to a point than to any adjacent point. The polygons are drawn so that the lines are of equal distance between two adjacent points. They are sometimes used as a crude form of interpolation.

LIST OF TABLES

LIST OF FIGURES

INDEX

APPENDIX 1

· ·
                                            ·
                                            ·

Environmental characterization of breeds

Yearly mean of topo-climatic variables are used (see codebook in appendix 4).

Goats

Breed name	longitude	latitude	altitude	wnd	dtr	frs	pr	prcv	tmp	rdo	reh	sun
ABAZA	42.90	40.95	2083.40	2.39	12.32	15.25	59.23	54.75	4.03	11.41	70.47	58.27
ALPINE FR	5.64	45.19	800.00	3.54	8.38	7.98	84.83	58.17	9.31	12.56	72.43	45.11
ALPINE CH	7.74	46.68	779.14	2.81	7.31	12.50	100.00	55.61	6.23	15.63	75.53	37.54
ANGORA	32.15	39.63	809.62	1.95	11.96	7.38	32.99	67.45	11.74	8.38	61.83	62.79
ARGENTATA DELL'ETNA	15.00	37.99	749.00	4.22	6.63	2.43	58.96	86.40	14.93	7.67	70.80	55.25
BALADI	35.64	31.32	300.00	3.15	12.65	3.85	21.40	226.78	17.22	3.11	51.58	75.63
BIONDA DELL'ADAMELLO	10.30	45.97	761.27	3.14	7.10	15.85	67.95	67.82	4.74	14.53	71.73	42.61
BRAVIA	-8.01	41.56	923.11	4.36	9.51	6.07	119.63	74.00	11.20	13.23	72.42	51.03
CABRA DEL GUADARRAMA	-4.11	40.55	769.50	3.12	10.73	6.38	40.62	84.02	12.90	10.28	61.31	56.94
CAMOSCIATA DELLE ALPI	9.82	46.01	553.10	2.29	7.70	12.26	87.45	65.32	7.98	13.49	73.62	42.65
CAPORE	20.62	40.88	858.82	2.19	10.24	7.08	70.76	68.25	9.87	10.98	70.42	49.54
CARPATHIAN	23.24	46.53	424.64	2.77	9.54	11.35	60.73	54.86	7.88	12.70	79.96	41.49
CORSICAN	9.21	42.14	340.00	4.45	4.95	5.92	69.65	82.28	11.05	10.99	75.79	53.25
DUKATI	19.50	40.29	520.44	2.59	9.67	4.44	98.87	70.94	12.74	10.53	71.75	52.97
FLORIDA	-5.16	37.82	324.50	3.93	13.02	3.55	51.49	112.74	16.29	9.89	63.70	62.11
GERMAN ALPINE	9.08	49.87	240.55	2.93	7.85	9.11	62.79	51.81	8.20	14.96	78.38	32.74
GIRGENTANA	13.87	37.51	410.63	4.27	7.03	1.61	39.26	107.90	15.83	6.65	72.18	58.02
GREEK GOAT	24.04	38.86	629.27	2.87	9.14	5.59	67.46	81.29	11.84	8.44	67.44	59.62
GRIGIA MOLISANA	14.40	41.52	479.17	3.60	7.14	5.17	61.36	73.78	11.63	10.41	71.54	46.84
GRISONS STRIPED	9.57	46.67	1111.55	3.97	6.03	17.68	123.47	58.34	2.41	17.01	74.66	40.17
GURCU	43.24	40.38	2209.00	2.38	12.69	15.59	46.97	57.60	3.67	11.50	69.60	58.57
HAIR	38.42	38.74	1327.58	2.38	11.49	8.62	45.11	82.08	10.63	8.55	61.68	65.05
HASI	20.41	42.18	729.55	2.62	10.21	7.24	83.92	63.23	10.30	11.04	68.54	42.76
HUNGARIAN NATIVE	19.90	47.02	107.80	2.71	9.77	9.45	44.12	63.11	10.84	11.09	74.61	40.77
MALAGUEĐA	-4.36	36.87	348.60	4.25	10.45	3.32	50.33	117.79	15.62	9.27	64.81	62.35
MUZHAKE	20.21	40.09	472.80	1.97	10.82	5.94	107.28	68.99	11.51	10.98	70.48	53.91
ORINO	20.92	39.41	627.00	1.48	10.51	5.38	88.16	71.02	11.76	10.38	67.93	60.62
OROBICA	9.56	45.88	526.73	2.47	7.29	12.57	97.39	63.26	7.73	14.10	73.97	42.13
PAYOYA	-5.39	36.84	504.10	4.63	10.65	3.94	66.40	113.36	15.01	9.19	67.21	62.06
PEACOCK GOAT	9.00	46.42	625.00	2.51	7.28	11.83	100.00	53.86	6.99	15.67	76.06	36.08
PINZGAUER	12.77	47.35	946.50	3.16	8.15	14.51	100.89	54.04	4.29	15.85	74.09	40.31
POLISH FAWN COLOURED GOAT	18.20	52.47	120.10	3.71	7.76	9.91	52.34	53.65	7.50	14.49	80.88	31.86
PYRENEAN	1.02	43.34	430.00	3.88	9.60	7.49	84.61	65.17	10.15	12.23	72.48	48.89
ROVE	4.99	44.21	507.22	4.02	8.69	4.64	64.07	69.18	11.76	10.15	71.41	53.78
SARDA	9.16	39.55	402.00	3.95	7.36	3.69	63.02	92.42	13.85	10.56	76.11	55.92
SKOPELOS	23.31	39.33	310.91	1.86	9.17	2.53	56.24	78.19	14.26	7.27	69.26	61.42
ST. GALLEN BOOTED GOAT	9.26	47.21	723.09	2.71	7.30	12.62	97.34	54.76	6.43	15.67	75.18	37.13
TAUERNSCHECKEN	13.11	47.27	998.11	3.41	7.64	14.84	111.56	53.49	3.72	15.89	75.32	40.66
THURINGIAN FOREST GOAT	10.31	50.84	307.00	3.20	7.92	9.20	55.69	52.57	8.15	14.68	78.84	32.09
VALAIS BLACK NECK	8.00	46.33	983.36	3.71	6.34	16.52	116.71	59.87	3.25	16.24	74.30	40.24
VALDOSTANA	7.42	45.69	749.18	3.47	6.74	14.41	99.11	63.98	4.89	14.78	73.32	41.53
VERATA	-5.64	40.11	803.60	3.36	11.22	5.29	35.68	89.66	14.06	10.44	62.35	59.11
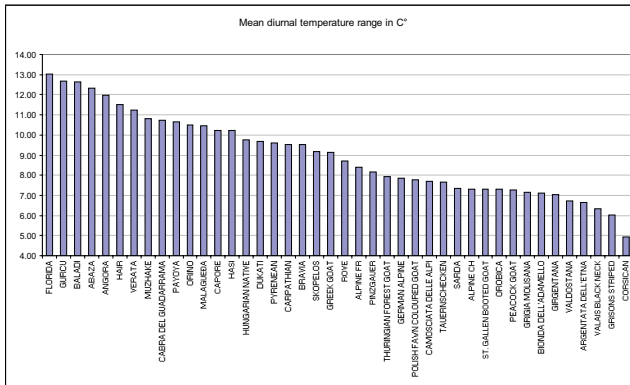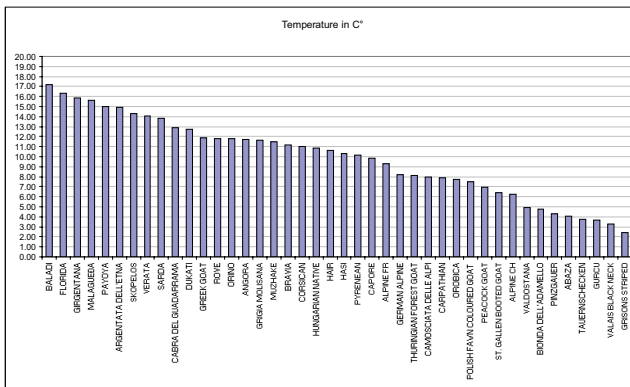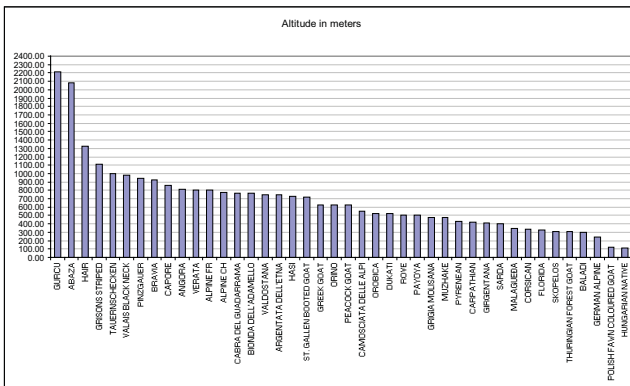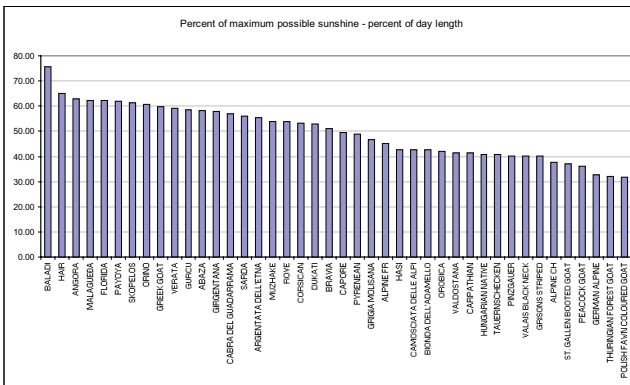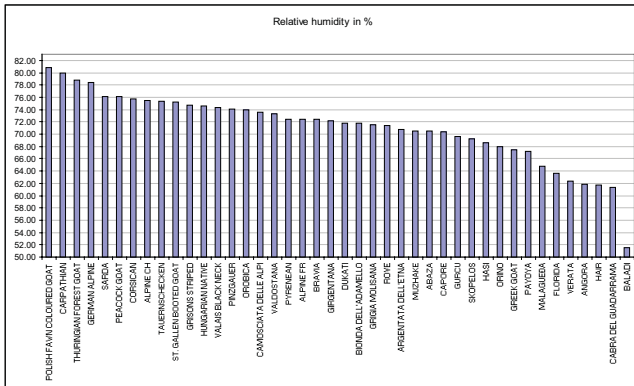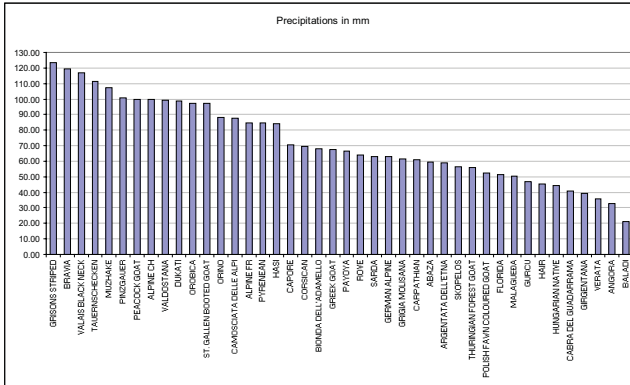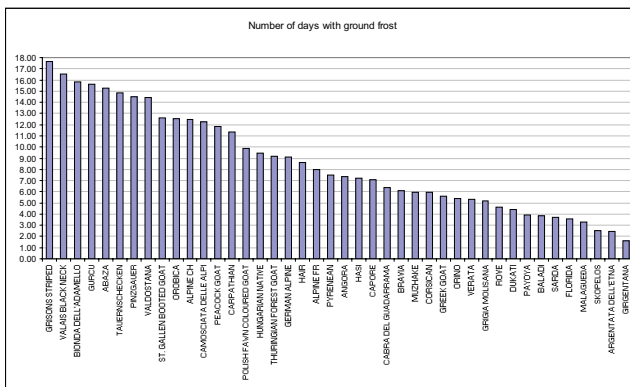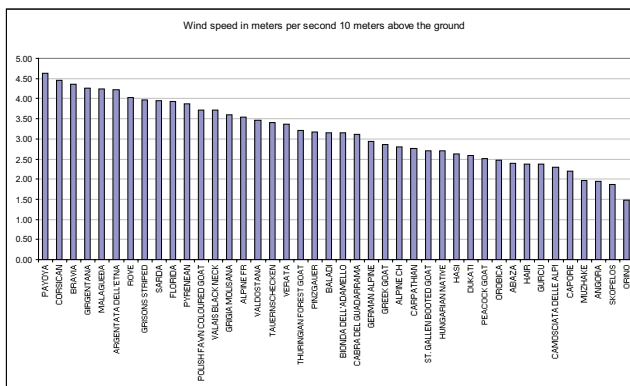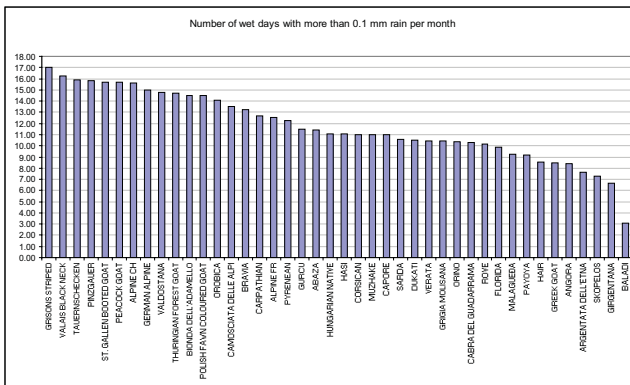
Appendix 1

Altitude in meters

Temperature in C°

Mean diurnal temperature range in C°

II

Goats

Precipitations in mm

Relative humidity in %

Percent of maximum possible sunshine - percent of day length

Appendix 1

Number of wet days with more than 0.1 mm rain per month

Wind speed in meters per second 10 meters above the ground

Number of days with ground frost

Environmental characterization of breeds

Yearly mean of topo-climatic variables are used.

Sheep

Breed name	longitude	latitude	altitude	wnd	dtr	frs	pr	prcv	tmp	rdo	reh	sun
AKKARAMAN	40.31	39.86	1633.67	2.17	12.59	13.60	42.18	64.28	5.45	10.51	66.55	59.89
ALTAMURANA	15.48	41.13	218.00	3.68	7.53	4.00	42.86	76.60	13.12	10.26	70.18	51.49
ANOGEIANO	24.82	35.25	1424.36	4.26	7.19	3.97	92.61	125.55	13.69	6.61	62.40	62.33
AWASSI	35.87	32.58	300.00	3.25	12.09	2.18	34.56	207.29	17.98	4.62	58.06	70.93
BARDHOKA	19.73	42.02	158.23	2.12	10.00	3.87	117.86	64.21	13.44	10.71	69.10	50.25
BLACK KARAKUL	26.40	47.40	161.40	2.89	10.53	11.16	46.93	62.17	8.17	11.63	80.40	39.89
CHURRA BRAGANÇANA	-8.34	41.78	565.40	4.47	9.28	3.86	133.11	73.79	12.63	13.93	74.74	49.23
CIKTA	18.45	47.62	274.00	3.11	9.14	9.77	48.68	60.63	9.99	11.10	72.95	38.83
COLMENAREÑA	-3.75	40.67	734.20	3.08	10.95	6.89	46.45	81.97	12.43	10.67	62.47	54.97
EXMOOR HORN	-3.75	51.16	296.45	5.01	6.08	8.08	117.84	48.95	8.99	17.93	88.17	30.69
GENTILE DI PUGLIA	14.74	41.27	442.75	2.86	8.94	3.28	64.74	73.10	13.94	10.69	72.06	50.16
GERMAN GREY HEATH	9.87	52.85	22.44	3.98	7.75	8.44	59.00	50.53	8.65	15.00	82.37	32.63
GERMAN MERINO	10.09	49.16	401.20	2.63	8.39	9.74	62.20	51.66	8.26	14.93	78.27	34.44
HUNGARIAN MERINO	19.78	46.76	113.91	2.74	9.68	9.38	44.96	62.53	10.83	11.04	74.64	41.10
KALARRITIKO	21.80	39.66	296.27	1.02	11.65	4.28	55.11	77.72	13.77	9.15	65.91	58.85
KAMIENIEC	20.30	53.62	106.40	3.68	7.70	10.23	52.39	53.15	7.13	14.57	80.99	32.42
KARAGOUNIKO	21.83	39.48	203.55	0.98	11.56	3.82	59.77	77.45	14.06	9.12	65.41	60.92
KEFALLENEAS	20.57	38.32	195.78	2.42	9.04	2.63	85.96	83.54	15.03	9.07	69.48	62.85
KYMI	24.11	38.50	148.55	2.46	7.94	1.29	49.87	92.58	15.75	6.22	68.31	63.66
LATICAUDA	14.24	41.20	218.67	2.63	9.38	3.03	72.08	71.11	14.28	10.86	72.65	49.99
LESVOS	26.19	39.23	202.55	4.06	7.59	1.61	54.01	100.80	15.23	7.26	68.59	67.33
MAGYAR RACKA	20.88	47.04	92.50	2.58	10.18	9.08	45.76	62.07	10.66	10.87	75.83	41.70
MANCHEGA	-3.69	39.54	685.60	3.28	12.26	6.63	38.82	89.33	13.45	10.31	63.18	58.59
MORKARAMAN	42.71	39.91	2069.77	2.10	12.20	13.48	40.87	65.14	5.94	9.97	64.50	59.86
ORINO	20.76	39.70	713.30	1.50	10.64	5.65	89.46	70.35	11.57	10.50	68.22	59.49
PECORA BERGAMASCA	9.78	45.45	304.90	1.59	8.74	9.14	83.88	65.78	10.57	11.97	74.91	43.84
PECORA DELLE LANGHE	8.04	44.49	728.45	2.86	6.66	5.96	65.38	77.97	11.70	10.72	72.67	47.44
PILIORITIKO	23.02	39.31	311.09	1.67	10.15	4.01	50.64	78.23	12.94	8.08	66.42	60.48
POLISH HEAT	19.81	53.33	160.10	3.71	8.20	11.77	49.66	53.03	6.64	14.36	81.23	33.04
POLISH MERINO	18.59	52.88	92.45	3.31	8.31	9.55	44.18	51.65	8.04	14.15	80.47	31.75
POLISH MOUNTAIN	20.14	49.38	685.73	3.85	9.37	13.73	75.65	50.94	6.20	14.32	77.94	35.79
POMERANIAN	17.41	54.45	109.27	4.61	6.72	9.46	54.79	54.55	7.08	14.82	82.73	31.70
RHOENSHEEP	9.85	50.58	419.89	3.15	7.84	9.75	62.76	51.79	7.51	15.03	79.25	32.56
RUBIA DEL MOLAR	-3.82	40.69	564.46	3.16	8.76	9.13	59.06	60.93	8.81	13.74	74.48	39.44
RUDA	20.28	41.50	1148.18	3.18	9.58	9.54	92.52	63.34	6.98	12.00	72.18	45.85
SCOTTISH BLACKFACE	-3.16	55.50	219.45	4.52	7.19	11.68	105.33	48.79	6.81	18.34	87.44	25.63
SEGUREÑA	-2.32	37.55	964.30	3.77	11.23	6.91	44.63	102.78	13.00	10.06	63.75	61.52
SFAKIA	24.21	35.27	349.91	3.94	7.62	2.19	70.73	128.95	15.80	6.36	63.22	61.56
SHKODRANE	19.43	42.27	67.43	2.06	9.67	3.75	133.67	63.80	13.95	10.74	67.91	50.96
SKOPELOS	23.23	39.16	121.91	1.84	9.05	1.88	56.79	79.28	14.85	7.03	69.60	61.97
SPANISH MERINO	-5.99	39.16	354.00	3.56	12.14	3.68	42.59	100.71	15.85	10.29	61.68	61.02
SWALEDALE	-1.79	54.34	289.82	4.93	6.63	11.01	104.27	50.16	7.05	18.42	87.46	26.54
THONES ET MARTHOD	6.62	45.62	996.64	3.74	7.24	13.89	99.50	60.24	4.83	14.66	72.96	41.66
TRANSYLVANIAN MERINO	22.42	47.30	148.55	2.46	9.97	9.18	51.33	56.41	9.83	11.20	77.63	42.77
TSIGAIA HUNGARY	18.94	47.02	267.20	2.89	9.31	11.01	54.83	57.95	8.79	12.24	77.42	40.83
TSIGAIA ROMANIA	24.78	46.19	597.00	2.78	10.99	13.47	57.47	59.17	6.33	12.52	81.30	38.96
TURCANA	23.66	46.26	496.36	3.45	9.08	12.53	67.63	55.25	6.61	13.27	81.18	40.32
WELSH MOUNTAIN	-3.63	52.40	263.55	5.12	6.79	9.59	128.44	47.69	8.17	17.51	87.37	27.71
WHITE/BROWN MOUNTAIN SHEEP	11.14	48.13	651.45	2.66	8.29	11.58	81.11	52.29	7.29	15.60	76.05	38.10
ZELAZNA	21.68	52.63	192.50	3.49	8.31	11.17	48.51	53.18	7.10	14.21	80.88	32.57

Appendix 1

Sheep

Altitude in meters

Temperature in C°

Mean diurnal temperature range in C°

Sheep

Precipitations in mm

Relative humidity in %

Percent of maximum possible sunshine - percent of day length

Appendix 1

Number of wet days with more than 0.1 mm rain per month

Wind speed in meters per second 10 meters above the ground

Mean diurnal temperature range in C°
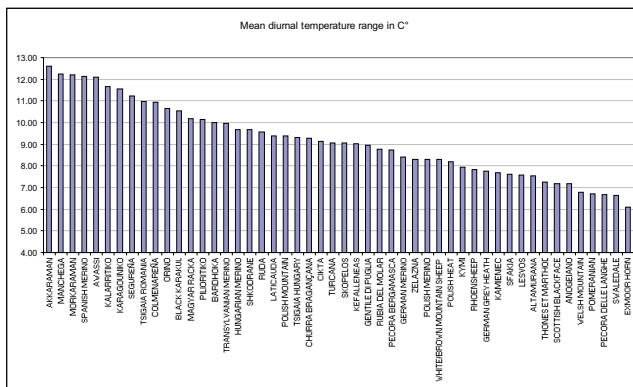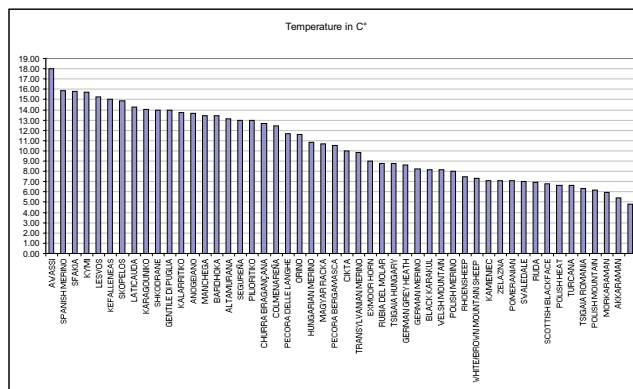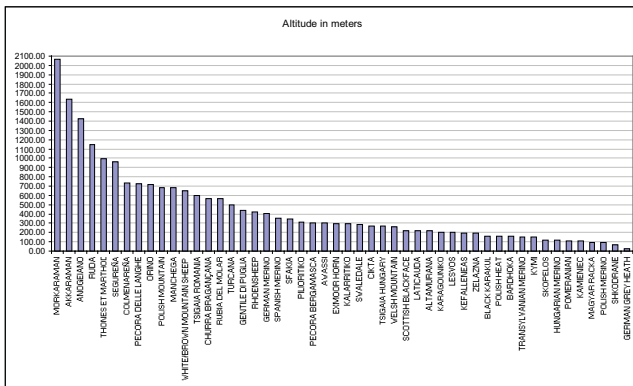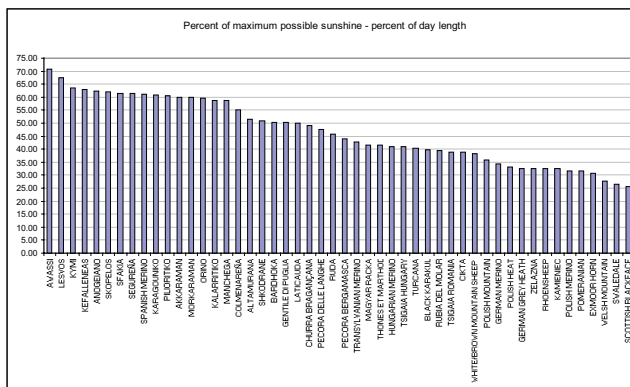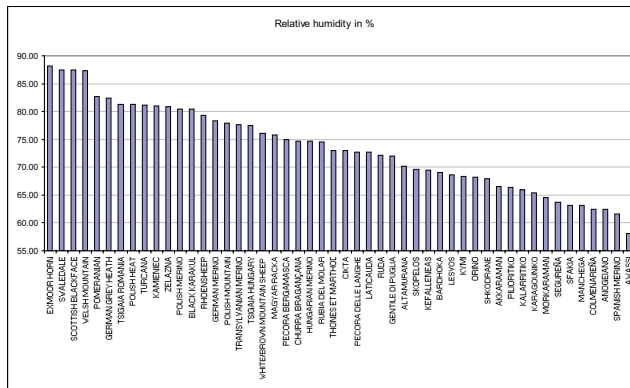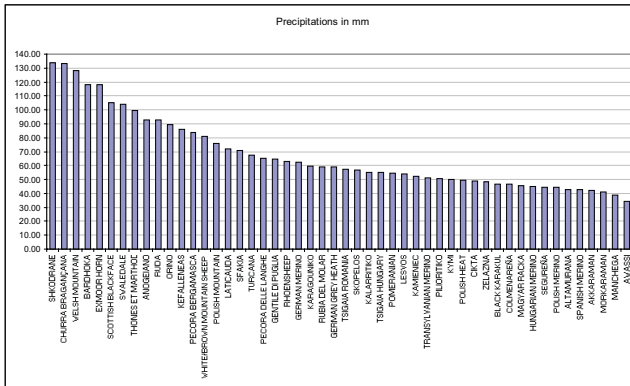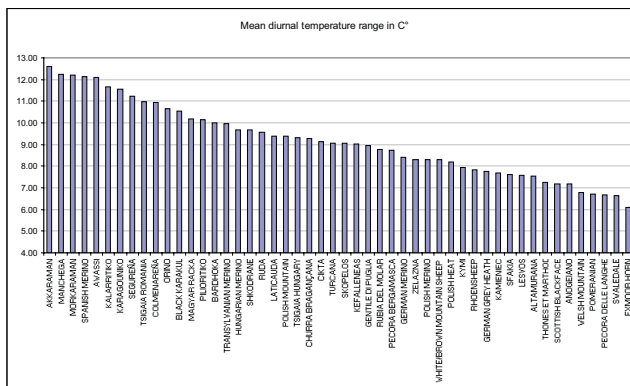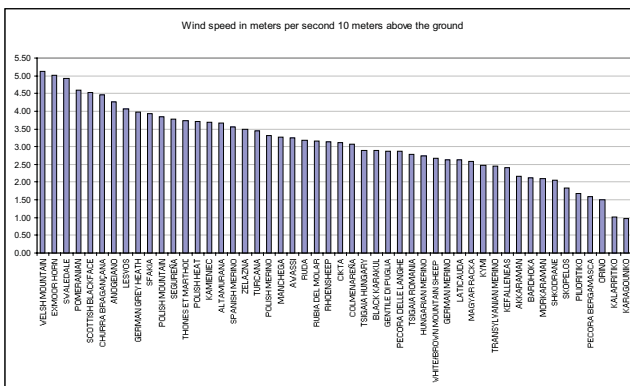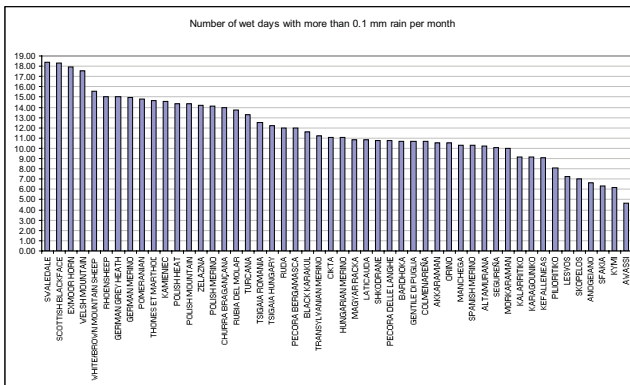
APPENDIX 2

Polymerase Chain Reaction
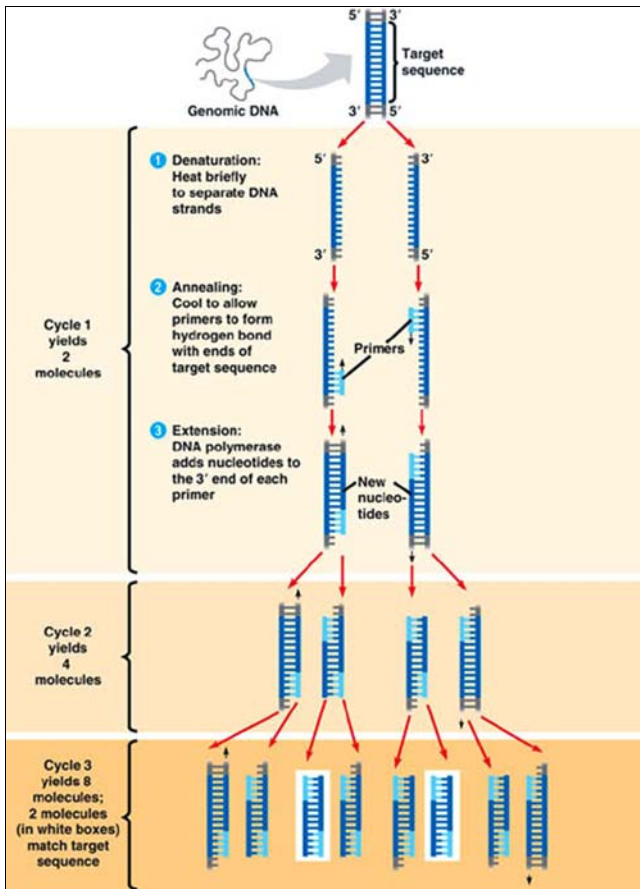
© Charles H. Mallery, Department of Biology, University of Miami

Appendix 2

APPENDIX 3

· ·

Production of AFLP markers

a) AFLP template preparation
Whole genomic DNA

Restriction enzymes
(*Mse*I and *Eco*RI)

*Mse*I adaptor
G
C A T

*Eco*RI adaptor
A A T T G
C

b) Restriction and ligation

*Mse*I cut   *Eco*RI cut

5'-   T   T A A   G   A A T T C
      A A T   T   C T T A A   G
3'-

*Mse*I adaptor   *Eco*RI adaptor
      G   A A T T G
      C A T   C

      G T A A   G A A T T G
      C A T T   C T T A A C

c) Selective amplification (one of many primer combinations)

*Mse*I primer 1
      C A T T G T A

5'-   G T A A C A T   C G A G A A T T G
3'-   C A T T G T A   G C T C T T A A C

      C G A G A A T T G
      *Eco*RI primer 1

Copyright © Trends in Ecology and Evolution, vol. 14, no. 10, october 1999

a) Extremely small amounts of DNA (~50 ng) are digested with two restriction enzymes;
b) AFLP adaptors are joined (ligated) to these ends;

The end sequences of each adapted fragment now consist of the adaptor sequence (in red) and the remaining part of the restriction sequence (in blue and green). These known end sequences serve as priming sites in the subsequent AFLP–PCR.

c) To achieve selective amplification of a subset of these fragments, primers are extended into the unknown part of the fragments (underlined base pairs), one to three arbitrarily chosen bases beyond the restriction site (in black);

Adapted from Mueller, U.G. and Wolfenbarger, L.L., (1999) AFLP genotyping and fingerprinting, Trends in Ecology & Evolution, 14 (10):389-394.

Appendix 3

APPENDIX 4

. . . .
. . .
. . .
. . .

Codebook of molecular and geoenvironmental variables

Molecular variables

Variable name	*Description*
aflppca1 *	AFLP PCA - contribution on factor 1
aflppca2 *	AFLP PCA - contribution on factor 2
mtdna_nbhaplotypes	Mitochondrial DNA number of haplotypes
mtdna_haplogrA_freq	MtDNA frequency of haplogroup A
mtdna_haplogrB_freq	MtDNA frequency of haplogroup B
mtdna_haplogrC_freq	MtDNA frequency of haplogroup C
mtdna_haplogrD_freq	MtDNA frequency of haplogroup D
mtdna_nucl_div	MtDNA nucleotides diversity
ych_nbhaplotypes	Y chromosome number of haplotypes
ych_haplogrA_freq	Y chromosome frequency of haplogroup A
ych_haplogrB_freq	Y chromosome frequency of haplogroup B
ych_haplogrC_freq	Y chromosome frequency of haplogroup Cs
microsat_ht	Microsatellites heterozygosity
microsat_Fis	Microsatellites inbreeding coefficient due to non random mating
microsat_mean_nb_alleles	Microsatellites mean number of alleles
aflp_average_exp_ht	AFLP average expected heterozigosity
aflp_se_exp_ht	AFLP standard error expected heterozigosity
aflp_av_jac_simindex	AFLP average Jaccard similarity index
aflp_nb_polym_loci	AFLP number of polymorphic loci
aflp_prc_polym_loci	AFLP percentage of polymorphic loci
aflp_freq_recess_genotype	AFLP frequency of recessive genotype

Geoenvironmental variables

longitude *	Longitude
latitude *	Latitude
altitude	Mean altitude of the breed farms (recorded on the field, Econogene)
wnd	Wind speed in m/s 10 meters above the ground
dtr	Mean diurnal temperature range in deg C
frs	Number of days with ground-frost
pr	Precipitations in mm/month
prcv	Coefficient of variation of monthly precipitation in percent
tmp	Mean temperature in deg C
rdo	Number of wet-days - number of days with >0.1mm rain per month
reh	Relative humidity in percent
sun	Percent of maximum possible sunshine (percent of day length)

* Those variables have not been included in the cluster analysis

Molecular variables are described in chapter 4.

Appendix 4

APPENDIX 5

. . . .
                                                    .
                                                    .
                                                    .

Exploratory Data Analysis : interactive correlation table
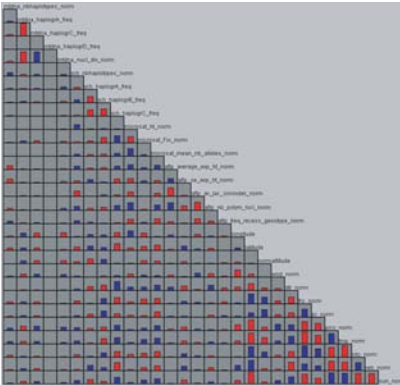
Goats with all correlations
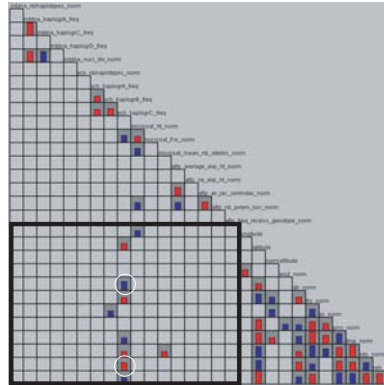
Goats with correlations > 0.44
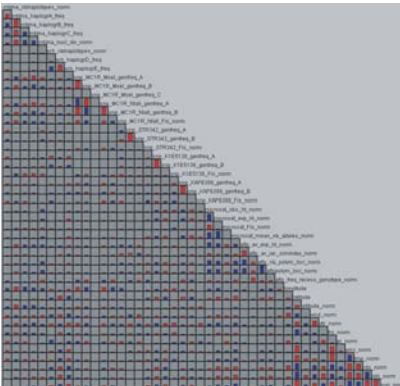
Sheep with all correlations
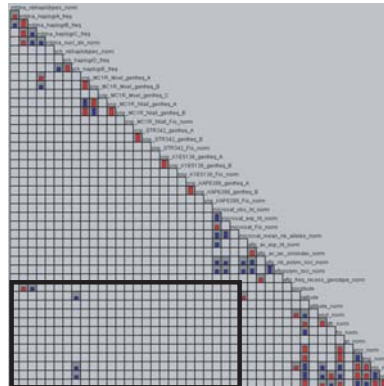
Sheep with correlations > 0.44

Interactive correlation table in CommonGIS (http://www.commongis.com/ - [23.11.2005]). Positive correlations are displayed in blue, and negative ones in red. The height of the bars is proportional to the correlation coefficient. The black rectangles are delimitating correlations between molecular and geoenvironmental data on which we focus. An interactive potentiometer allows to visualise remaining coefficients when increasing a correlation-threshold value.

Appendix 5

APPENDIX 6

. . . .
. . .
. . .
. . .

Clustering configurations from K = 2 to K = 7
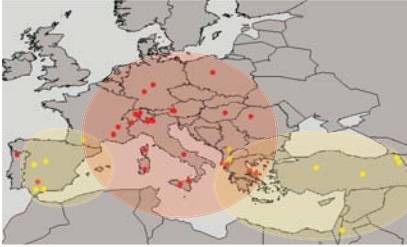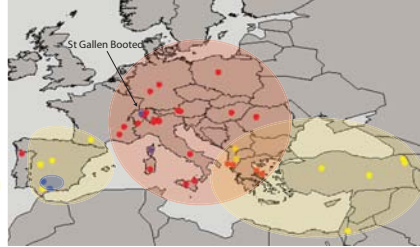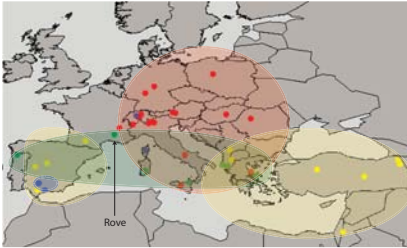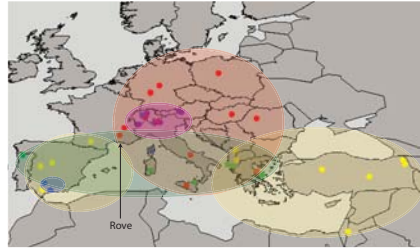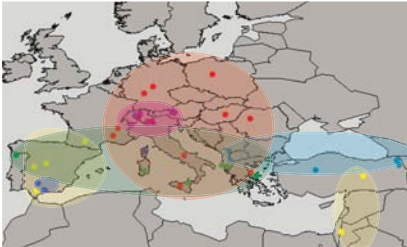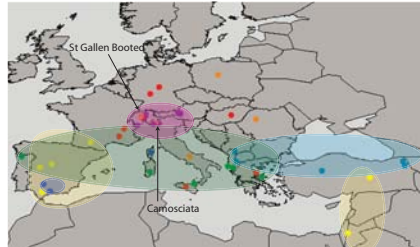
The large colored areas are displayed to make the main geographic structure of breed groups stand out.

K = 2 : two rather homogeneous groups, one Central European, and one South-Western and South-Eastern. In the Western part, the exceptions are Brava (Portugal) and Florida (South of Spain). In the Eastern part, it is possible to observe two Albanian breeds member of the yellow group (Hasi and Capore).

K = 3 : the blue group stands out with Florida, Malaguena, St Gallen Booted (see arrow) and Corsican breeds (mtDNA haplogroup C and mtDNA high diversity of nucleotides).

K = 4 : the green class appears with Brava (Portugal), Sarda (Sardegna), Argentata dell'Etna (Sicily), Dukati and Muzhake (the south Albanian), as well as Greek goat breeds. The Rove also joins this group (see arrow).

*K = 5 is detailed in chapter 5. Emergence of the purple Alpine group with Valdostana, Orobica (Italy), Peacock Goat (Switzerland), Pinzgauer and Tauern-schecken (Austria). With the emergence of the purple class, the Rove is back to the red class.

K = 6 : the yellow group splits and breeds of Northern Turkey (Angora, Abaza, Gurcu) join together with Northern Albanian breeds (Hasi and Capore).

K = 7 : the blue class looses the St Gallen Booted (Switzerland) which joins Camosciata and Grigia Molisana (Italy), Polish Fawn Colored (Poland) and the Carpathian in a new orange group.

APPENDIX 7

. . . .

NeighborNet graph of Nei standard genetic distances
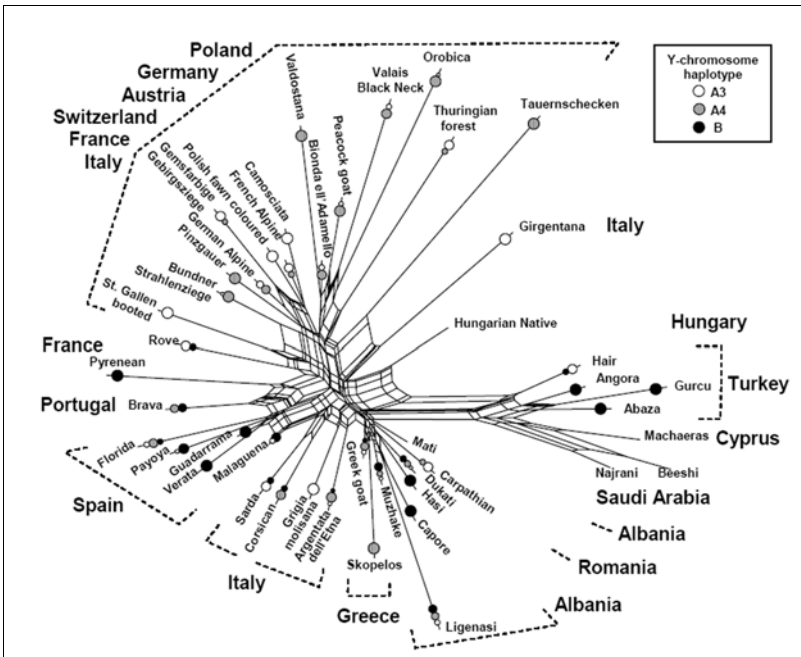
NeighborNet graphs [1] permits to visualize genetic distances of goat breeds (Lenstra, 2005). Model-based clustering [2] of goat microsatellite genotypes (Econogene data) reveals at least four discrete clusters : East-Mediterranean, Central Mediterranean, West-Mediterranean, and Central with Northern Europe, respectively. In breeds from the last cluster, the average number of microsatellite alleles is clearly less than in the Mediterranean breeds.

The sizes of the open, hatched and filled circle indicate frequencies of Y-chromosomal haplotypes A3, A4 and B, respectively.

[1] Bryant D, Moulton V. (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21, 255-65.

[2] Pritchard JK, Stephens M, Donnelly P. (2000) Inference of population structure using multilocus genotype data. Genetics 155, 945-59.

Appendix 7

APPENDIX 8

. . . .
. . .
. . .
. . .

40 sheep breeds used for the detection of natural selection signatures (AFLP)

Abbreviation	Breed Name	Country
altam	ALTAMURANA	Italy
bardh	BARDHOKA	Albania
blkarak	BLACK-KARAKUL	Romania
cickt	CIKTA	Hungary
colmen	COLMENARENA	Spain
exmo	EXMOOR HORN	UK
genpu	GENTILE DI PUGLIA	Italy
ggh	GERMAN GREY HEATH	Germany
germer	GERMAN MERINO	Germany
humer	HUNGARIAN MERINO	Hungary
kamen	KAMIENIEC	Poland
karag	KARAGOUNIKO	Greece
kefal	KEFALLENEAS	Greece
kymi	KYMI	Greece
lati	LATICAUDA	Italy
lasv	LESVOS	Greece
magrac	MAGYAR RACKA	Hungary
manch	MANCHEGA	Spain
pecberga	PECORA BERGAMASCA	Italy
pecdelghe	PECORA DELLE LANGHE	Italy
polheat	POLISH HEAT	Poland
polmer	POLISH MERINO	Poland
polmount	POLISH MOUNTAIN	Poland
polzelaz	POLISH ZELAZNA	Poland
rhoen	RHOENSHEEP	Germany
rubdelmo	RUBIA DEL MOLAR	Spain
ruda	RUDA	Albania
sctbf	SCOTTISH BLACKFACE	UK
segur	SEGURENA	Spain
sfak	SFAKIA	Greece
shkod	SHKODRANE	Albania
skopel	SKOPELOS	Greece
spanmer	SPANISH MERINO	Spain
swalda	SWALEDALE	UK
thone	THONES ET MARTHOD	France
trsvmer	TRANSYLVANIAN MERINO	Romania
tsig_hu	TSIGAIA HUNGARY	Hungary
tsig_ro	TSIGAIA ROMANIA	Romania
turca	TURCANA	Romania
wbmountsh	WHITE/BROWN MOUNTAIN SHEEP	Germany

Appendix 8

Complete list of Econogene sheep breeds

Breedname	Country
Bardhoka	Albania
Ruda	
Shkodrane	
Thone et Martod	France
Grey Heath	Germany
Rhöensheep	
Brown/White Mountain Sheep	
German merino	
Orino	Greece
Sfakia	
Anogeiano	
Kalarritiko	
Pilioritiko	
Kefalleneas	
Karagouniko	
Lesvos	
Kymi	
Skopelos	
Racka (black and white)	Hungary
Tsigaia	
Cikta	
Hungarian merino	
Gentile di Puglia	Italy
Laticauda	
Altamurana	
Bergamasca	
Delle Langhe	
Awassi	Jordan
Zelazna (Polish lowland)	Poland
Pomorska	
Kamieniec	
Polish mountain	
Polish heat	
Polish merino	
Churra Bragancana	Portugal
Turcanà	Romania
Tsigaia	
Transilvanian Merino	
Black Karakul	
Merinos	Spain
Merinos 2	
Manchega	
Colmenareña	
Segureña	
Rubia del Molar	
Daglic	Turkey
Akkaraman (White Karaman)	
Morkaraman (Red Karaman)	
Karayaka	
Scottish blackface	United Kingdom
Swaledale	
Welsh Mountain	
Exmoor Horn	

APPENDIX 9

. . . .
. . . .
. . .
. . .

List of Econogene goat breeds used in this research

Breedname	Country
Liqenasi	Albania
Hasi	
Mati	
Capore	
Muzhake	
Dukati	
Pintzgauer	Austria
Tauernschecken	
Rove	France
Pyrenean	
Corsican	
French Alpine	
Thuringian forest goat	Germany
German Alpine goat	
Skopelos	Greece
Greek goat	
Girgentana	Italy
Grigia molisana	
Bionda dell'Adamello	
Orobica	
Valdostana	
Camosciata (Italian Alpine)	
Argentata dell'Etna	
Sarda	
Polish fawn coloured goat	Poland
Brava	Portugal
Carpathian	Romania
Verata	Spain
Payoya	
Florida	
Guadarrama	
Malaguena	
St Gallen booted Goat	Switzerland
Peacock goat	
Alpine	
Grisons striped	
Angora	Turkey
Hair	
Gurcu	
Abaza	

Appendix 9

APPENDIX 10

Comparative behavior of Wald and G tests on 3 case studies

Sheep breeds - microsatellites
Number of significant models : Wald and G tests
Significance threshold :1.139E-17

Scandinavian Brown Bear - microsatellites
Number of significant models : Wald and G tests
Significance threshold : 5.52E-17

Sheep breed - AFLPs
Number of significant models : Wald and G tests
Significance threshold : 1.37E-09
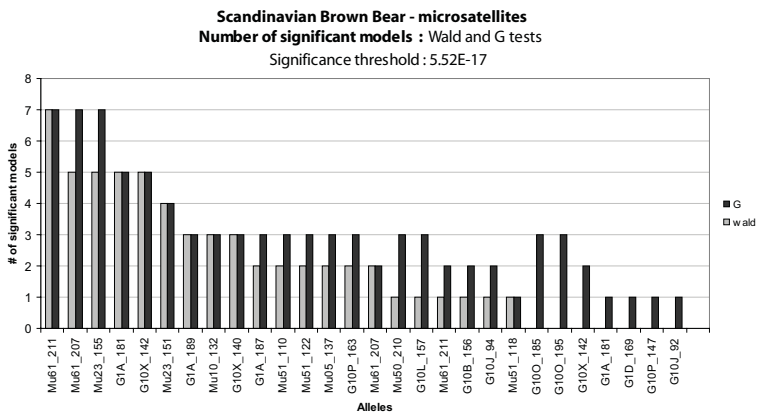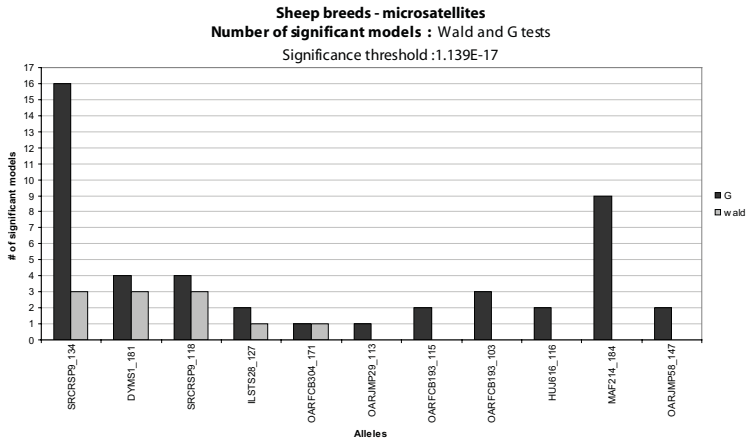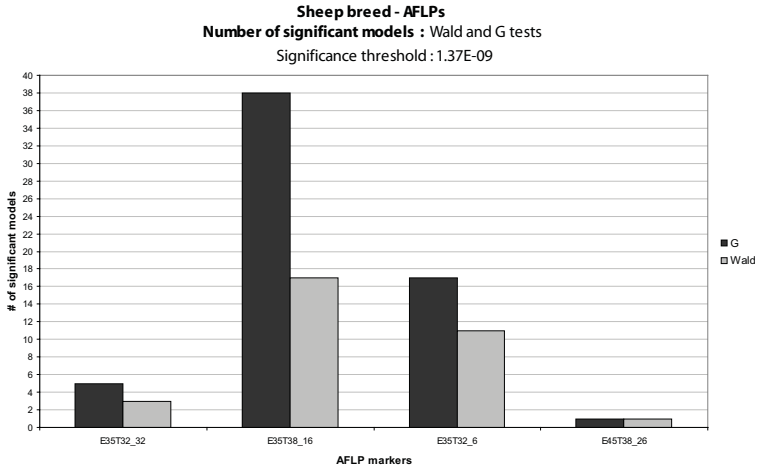
The 3 examples here above show that the Wald statistical test is globally more restrictive than the G test (likelihood ratio). Informations found in the literature are contradictory about their respective efficiency. For example, Thu *et al.* [1] state that the likelihood ratio test slightly outperforms the Wald test while the performance of the latter is satisfactory, especially when the size of the samples is large (what is the case with Econogene data sets). Agresti [2] confirms this last point : the likelihood ratio is better when the size of the sample is small. Whereas Conte and de Maio stipulate that «(...) the performance assessment, conducted also in comparison with the generalized likelihood ratio test (...) shows that the Wald test outperforms the others and is very effective (...)[3].

A reason likely to explain the conservatism of the Wald test compared to G is that in case of large logit coefficients, the standard error is inflated, and this lowers the Wald statistic and leads to Type II errors, that is false negatives : thinking the effect is not significant when it is [4].

[1] Tu, W., and Zhou, Z-H. (1999) A Wald test comparing medical costs based on log-normal distributions with zero valued costs. Statistics in Medicine 18, pp. 2749-2761.

[2] Agresti A. (1990) Categorical Data Analysis, John Wiley and Sons, New York.

[3] Conte, E. and de Maio, A. (2003) Distributed Target Detection in Compound-Gaussian Noise with Rao and Wald Tests, IEEE Transactions on Aerospace and Electronic Systems, 39 (2).

[4] Menard, S. (2002) Applied logistic regression analysis, 2nd Edition. Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106. (First ed., 1995).

APPENDIX 11

Molecular data set for sheep breeds

ID	breed	breed_code	country	longitude	latitude	microsatREYpca1	microsatREYpca2	mtdna_nbhaplotypes	mtdna_haploidA_freq	mtdna_haploidG_freq	mtdna_haploidC_freq	mtdna_nucl_div	Ych_nbhaplotypes	Ych_haploidG_freq	Ych_haploidY_freq	microsat_obs_ht	microsat_exp_ht	microsat_Fis	microsat_mean_nb_alleles	aHp_av_exp_ht	aHp_ae_exp_ht	aHp_av_Jac_simindex	aHp_nb_polym_loci	aHp_polym_loci	aHp_freq_recess_genotype
1	CHURRA BRAGANZANA	OAPTCHU	Portugal	-6.34	41.78	0.88	0.37	4	1.00	0.00	0.00	0.01	2	0.5	0.5	0.68	0.75	0.09	8.35	0.28	0.02	0.72	69	82.6	0.43
2	SPANISH MERINO	OAESMRS	Spain	-5.59	39.16	0.32	0.15	3	0.94	0.06	0.00	0.00	1	1	0	0.67	0.73	0.08	8.26	0.28	0.02	0.69	76	77.2	0.44
3	RUBIA DEL MOLAR	OAESRBM	Spain	-3.82	40.69	0.73	0.21	2	1.00	0.00	0.00	0.00	1	1	0	0.73	0.70	0.05	6.84	0.28	0.02	0.72	71	77.2	0.44
4	COLMENAREÑA	OAESCLM	Spain	-3.75	40.67	0.80	0.19	3	0.97	0.03	0.00	0.02	2	0.86	0.14	0.61	0.68	0.12	7.48	0.29	0.02	0.68	71	77.2	0.45
5	EXMOOR HORN	OAGBEXH	UK	-3.75	51.16	0.32	0.58	2	0.94	0.06	0.00	0.03	1	0	1	0.62	0.63	0.07	6.00	0.28	0.02	0.71	72	79.3	0.42
6	MANCHEGA	OAESMCG	Spain	-3.69	39.54	0.83	0.08	7	0.94	0.06	0.00	0.06	1	1	0	0.67	0.72	0.07	7.71	0.28	0.02	0.67	73	78.3	0.43
7	WELSH MOUNTAIN	OAGBWMT	UK	-3.63	52.40	0.84	0.41	3	1.00	0.00	0.00	0.02	1	1	0	0.67	0.68	0.03	7.03	0.29	0.02	0.69	72	78.3	0.45
8	SCOTTISH BLACKFACE	OAGBSBF	UK	-3.16	55.50	0.70	0.39	8	0.94	0.03	0.03	0.02	2	0.86	0.14	0.67	0.73	0.07	8.55	0.31	0.02	0.69	73	79.3	0.43
9	SEGUREÑA	OAESSGR	Spain	-2.32	37.55	0.93	0.15	3	1.00	0.00	0.00	0.00	2	0.5	0.5	0.69	0.69	0.03	6.90	0.31	0.02	0.69	70	76.1	0.42
10	SWALEDALE	OAGBSDL	UK	-1.79	54.34	0.57	0.36	4	2.00	0.00	0.00	0.00	2	0.33	0.67	0.63	0.68	0.07	8.74	0.29	0.02	0.67	76	80.4	0.45
11	THONES ET MARTHOD	OAFRTHM	France	6.62	45.62	0.72	0.44	4	0.97	0.03	0.00	0.03	1	0.67	0	0.60	0.65	0.09	6.35	0.29	0.02	0.68	74	76.1	0.47
12	PECORA DELLE LANGHE	OAITLAN	Italy	9.04	44.49	0.67	0.27	5	0.97	0.03	0.00	0.03	1	0	0	0.60	0.70	0.09	7.16	0.28	0.02	0.66	70	76.1	0.44
13	PECORA BERGAMASCA	OAITBER	Italy	9.76	45.43	0.13	0.27	5	0.97	0.03	0.00	0.03	1	0	0	0.60	0.65	0.09	6.03	0.30	0.02	0.66	69	76.1	0.44
14	RHOENSHEEP	OADERHO	Germany	9.87	50.86	0.57	0.41	12	0.90	0.10	0.00	0.03	1	0	0	0.60	0.64	0.12	6.10	0.27	0.02	0.64	74	78.3	0.47
15	GERMAN GREY HEATH	OADEGGH	Germany	9.87	52.85	0.12	-0.22	16	0.90	0.10	0.00	0.06	1	0	0	0.60	0.64	0.06	5.74	0.27	0.02	0.69	73	78.3	0.43
16	GERMAN MERINO	OADEMLS	Germany	10.09	49.16	0.80	-0.15	9	0.94	0.06	0.00	0.10	1	0	0	0.67	0.73	0.07	7.00	0.28	0.02	0.69	72	80.4	0.46
17	WHITE/BROWN MOUNTAIN SHEEP	OADEWBSBBS	Germany	11.14	48.13	0.86	-0.08	7	0.94	0.06	0.00	0.03	1	0	0	0.64	0.69	0.07	7.87	0.30	0.02	0.63	76	82.6	0.48
18	LATICAUDA	OAITLAT	Italy	14.24	41.20	0.82	0.13	6	0.97	0.03	0.00	0.03	2	0.68	0.32	0.58	0.75	0.15	7.39	0.30	0.02	0.73	73	79.3	0.41
19	GENTILE DI PUGLIA	OAITGDP	Italy	14.74	41.27	0.93	0.14	7	0.90	0.10	0.00	0.10	2	0.32	0.68	0.71	0.76	0.07	7.65	0.29	0.02	0.69	74	79.3	0.41
20	ALTAMURANA	OAITALT	Italy	15.48	41.13	0.87	0.29	5	0.94	0.06	0.00	0.06	2	0.32	0.68	0.67	0.74	0.09	6.39	0.30	0.02	0.71	71	80.4	0.44
21	POMERANIAN	OAPLPOM	Poland	17.41	54.45	0.78	0.00	6	0.94	0.06	0.00	0.06	2	0.33	0.67	0.67	0.70	0.04	8.19	0.33	0.02	0.68	72	77.2	0.43
22	CIKTA	OAHUCK	Hungary	18.45	47.62	0.78	-0.39	6	0.81	0.13	0.06	0.13	2	0.04	0.04	0.81	0.77	0.03	8.13	0.30	0.02	0.67	80	78.3	0.44
23	POLISH MERINO	OAPLMPO	Poland	18.59	52.88	0.95	-0.15	3	0.88	0.06	0.06	0.06	2	0.67	0.33	0.71	0.76	0.07	8.58	0.30	0.02	0.62	71	83.7	0.42
24	TSIGAIA	OAHUTSI	Hungary	19.03	47.10	0.93	-0.15	3	0.87	0.13	0.00	0.13	2	0.32	0.32	0.67	0.72	0.08	8.84	0.31	0.02	0.67	72	77.8	0.47
25	SHKODRANE	OAALSKO	Albania	19.43	42.27	0.72	0.28	9	1.00	0.00	0.00	0.06	2	0.29	0.29	0.70	0.72	0.02	8.06	0.28	0.02	0.69	73	78.9	0.42
26	BARDHOKA	OAALBAR	Albania	19.73	42.02	0.87	-0.03	10	0.94	0.06	0.00	0.03	2	0.71	0.71	0.69	0.73	0.04	8.42	0.30	0.02	0.71	73	79.3	0.43
27	HUNGARIAN MERINO	OAHUMER	Hungary	19.78	46.76	0.72	-0.13	5	0.97	0.03	0.00	0.03	2	0.71	0.71	0.67	0.75	0.10	8.84	0.29	0.02	0.71	78	84.8	0.42
28	POLISH HEATH	OAPLWRZ	Poland	19.81	53.33	0.87	-0.16	6	1.00	0.00	0.00	0.03	2	0.08	0.08	0.75	0.77	0.02	8.44	0.30	0.02	0.70	74	82.6	0.41
29	POLISH MOUNTAIN	OAPLGOR	Poland	19.81	49.38	0.93	-0.11	4	0.97	0.03	0.00	0.06	2	0.67	0.67	0.70	0.73	0.04	8.19	0.31	0.02	0.67	76	82.6	0.43
30	RUDA	OAALRUD	Albania	20.28	41.50	0.91	0.03	7	0.97	0.03	0.00	0.03	2	0.30	0.30	0.71	0.77	0.07	7.58	0.28	0.02	0.67	77	78.3	0.42
31	KAMIENIEC	OAPLKAM	Poland	20.30	53.62	0.53	0.23	9	0.87	0.13	0.00	0.19	2	0.62	0.62	0.70	0.71	0.03	7.52	0.32	0.02	0.62	73	77.8	0.43
32	KEFALLENEAS	OAGRKEF	Greece	20.57	38.32	0.91	-0.02	5	0.87	0.13	0.00	0.13	2	0.32	0.32	0.67	0.72	0.08	7.52	0.32	0.02	0.70	74	82.2	0.42
33	ORINO	OAGRORI	Greece	20.75	39.70	0.86	-0.03	10	0.94	0.06	0.00	0.06	2	0.29	0.29	0.67	0.72	0.08	8.42	0.29	0.02	0.69	74	79.3	0.42
34	MAGYAR RACKA	OAHURAC	Hungary	20.88	47.04	0.72	-0.15	5	1.00	0.00	0.00	0.03	2	0.29	0.29	0.66	0.72	0.09	6.84	0.30	0.02	0.71	78	83.7	0.43
35	ZELAZNA	OAPLZEL	Poland	21.68	52.04	0.91	-0.24	6	0.97	0.03	0.00	0.03	2	0.71	0.71	0.65	0.73	0.11	6.84	0.29	0.02	0.71	77	84.8	0.42
36	KALARRITIKO	OAGRKAL	Greece	21.83	39.48	0.96	-0.20	3	1.00	0.00	0.00	0.03	2	0.23	0.23	0.62	0.72	0.11	8.23	0.25	0.02	0.68	72	80	0.43
37	KARAGOUNIKO	OAGRKAR	Greece	22.02	39.16	0.88	-0.16	4	0.97	0.03	0.00	0.03	2	0.57	0.57	0.70	0.75	0.08	7.97	0.30	0.02	0.69	74	77.2	0.42
38	TRANSYLVANIAN MERINO	OAROMER	Romania	22.42	47.30	0.88	-0.24	5	1.00	0.00	0.00	0.03	2	0.71	0.71	0.66	0.76	0.13	6.23	0.29	0.02	0.70	73	78.3	0.43
39	PILORITIKO	OAGRPIL	Greece	23.02	39.31	0.64	0.03	6	0.97	0.03	0.00	0.03	2	0.08	0.08	0.66	0.72	0.08	8.26	0.28	0.02	0.70	67	78.3	0.43
40	SKOPELOS	OAGRSKO	Greece	23.23	39.16	0.91	-0.14	8	0.90	0.10	0.00	0.10	2	0.76	0.76	0.70	0.73	0.06	6.29	0.29	0.02	0.72	75	81.5	0.42
41	TURCANA	OAROTUR	Romania	23.66	46.26	0.96	-0.21	8	0.90	0.10	0.00	0.10	2	0.71	0.71	0.71	0.74	0.05	8.39	0.29	0.02	0.70	72	82.4	0.42
42	KYMI	OAGRKIM	Greece	24.11	38.50	0.87	-0.14	8	2.00	0.00	0.00	0.00	2	0.29	0.29	0.70	0.71	0.07	7.39	0.29	0.02	0.71	73	82.6	0.40
43	SFAKIA	OAGRSFA	Greece	24.21	35.27	0.91	-0.10	4	0.90	0.10	0.00	0.03	1	0	0	0.71	0.74	0.05	8.06	0.28	0.02	0.70	72	79.3	0.43
44	TSAKA	OAROTSI	Romania	24.79	46.19	0.91	0.13	6	0.90	0.10	0.00	0.10	2	0.76	0.76	0.66	0.72	0.08	9.26	0.30	0.02	0.69	73	81.5	0.42
45	ANOGEIANO	OAGRANO	Greece	24.82	35.25	0.91	-0.04	6	0.90	0.10	0.00	0.03	2	0.71	0.71	0.66	0.73	0.10	8.39	0.28	0.02	0.69	74	82.4	0.42
46	LESVOS	OAGRLES	Greece	26.19	39.23	0.93	-0.21	9	0.90	0.10	0.00	0.10	2	0.71	0.71	0.71	0.74	0.05	8.06	0.30	0.02	0.70	75	81.4	0.41
47	BLACK KARAKUL	OAROKAR	Romania	26.40	47.40	0.79	-0.43	2	0.90	0.10	0.00	0.02	2	0.29	0.29	0.66	0.71	0.07	7.39	0.29	0.02	0.71	73	79.3	0.41
48	OSSIMI	OAEGOSS	Egypt	30.70	28.10	0.86	-0.32	0	0.66	0.17	0.17	0.02	2	0.48	0.43	0.70	0.74	0.05	7.39	0.28	0.02	0.71	73	78.3	0.43
49	AKKARAMAN	OATRAKK	Turkey	40.31	39.86	0.78	-0.34	16	0.66	0.17	0.17	0.01	1	1	1	0.66	0.72	0.06	6.42	0.31	0.02	0.72	72	79.1	0.42
50	MOR KARAMAN	OATRMOR	Turkey	42.71	39.91	0.86	-0.41	17	0.45	0.48	0.07	0.02	1	1	1	0.67	0.72	0.08	8.39	0.30	0.02	0.66	74	81.3	0.45

Appendix 11

APPENDIX 12

Molecular data set for goat breeds (1)

ID	breed	breed code	country	longitude	latitude	alppep1	alppep2	mtdna_nbhaplotypes	mtdna_haplogA_freq	mtdna_haplogC_freq	mtdna_haplogD_freq	mtdna_nucl_div	ych_nbhaplotypes	ych_haplogA_freq	ych_haplogB_freq	ych_haplogC_freq
1	BRAVA	CHFOBRA	Portugal	-8.01	41.56	0.82	0.18	8	1.00	0.00	0.00	0.02	2	0.00	0.50	0.50
2	VERATA	CHSPVRT	Spain	-5.64	40.11	0.70	-0.17	7	1.00	0.00	0.00	0.02	2	0.00	0.00	1.00
3	PAYOYA	CHSPPYY	Spain	-5.39	36.84	0.75	-0.16	8	1.00	0.00	0.00	0.02	3	0.13	0.00	0.88
4	FLORIDA	CHSPFLR	Spain	-5.16	37.82	0.81	0.03	8	0.88	0.13	0.00	0.03	3	0.22	0.56	0.22
5	MALAGUENA	CHSPMLG	Spain	-4.36	36.87	0.92	-0.14	8	0.88	0.13	0.00	0.04	2	0.22	0.00	0.78
6	CABRA DEL GUADARRAMA	CHSPGDR	Spain	-4.11	40.55	0.83	0.23	6	1.00	0.00	0.00	0.02	1	0.00	0.00	1.00
7	PYRENEAN	CHFRPYR	France	1.02	43.34	0.89	0.36	6	1.00	0.00	0.00	0.02	2	0.00	0.63	0.38
8	ROVE	CHFRROV	France	4.99	44.21	0.86	0.13	7	1.00	0.00	0.00	0.02	2	0.60	0.40	0.00
9	FRENCH ALPINE	CHFRALP	France	5.64	45.19	0.79	0.30	7	1.00	0.00	0.00	0.02	2	1.00	0.00	0.00
10	VALDOSTANA	CHITVAL	Italy	7.43	45.71	0.74	0.34	8	1.00	0.00	0.00	0.02	2	1.00	0.00	0.00
11	ST. GALLEN BOOTED GOAT	CHCHSGB	Switzerland	7.93	46.88	0.79	0.06	5	0.83	0.17	0.00	0.04	2	0.75	0.25	0.00
12	SWISS ALPINE	CHCHALP	Switzerland	8.22	46.72	0.81	0.28	15	1.00	0.00	0.00	0.02	2	0.75	0.25	0.00
13	VALAIS BLACK NECK	CHCHVBN	Switzerland	8.36	46.76	0.67	0.02	8	1.00	0.00	0.00	0.02	2	0.25	0.75	0.00
14	GRISONS STRIPED	CHCHGRS	Switzerland	8.41	46.77	0.75	0.44	7	1.00	0.00	0.00	0.02	1	0.11	0.89	0.00
15	PEACOCK GOAT	CHCHPCG	Switzerland	8.53	47.07	0.80	0.26	8	1.00	0.00	0.00	0.02	2	0.18	0.82	0.00
16	GERMAN ALPINE	CHDEBDE	Germany	9.08	49.87	0.88	0.22	8	1.00	0.00	0.00	0.02	2	0.44	0.56	0.00
17	SARDA	CHITSAR	Italy-Sard	9.16	39.55	0.85	0.04	8	1.00	0.00	0.00	0.03	2	0.63	0.00	0.38
18	CORSICAN	CHFRCOR	France	9.21	42.14	0.84	0.27	6	0.86	0.14	0.00	0.03	2	0.00	0.67	0.33
19	OROBICA	CHITORO	Italy	9.56	45.88	0.50	0.57	6	1.00	0.00	0.00	0.02	2	0.11	0.89	0.00
20	CAMOSCIATA DELLE ALPI	CHITCAM	Italy	9.82	46.01	0.74	0.51	5	1.00	0.00	0.00	0.03	1	1.00	0.00	0.00
21	BIONDA DELL'ADAMELLO	CHITBIO	Italy	10.30	45.97	0.83	0.57	7	1.00	0.00	0.00	0.02	2	0.33	0.67	0.00
22	THURINGIAN FOREST GOAT	CHDETWZ	Germany	10.31	50.84	0.74	0.22	8	1.00	0.00	0.00	0.02	2	0.67	0.33	0.00
23	PINZGAUER	CHAUPIZ	Österreich	12.77	47.35	0.86	0.35	6	1.00	0.00	0.00	0.02	1	0.00	1.00	0.00
24	TAUERNSCHECKEN	CHAUTAS	Österreich	13.11	47.27	0.45	0.22	4	1.00	0.00	0.00	0.02	1	0.00	1.00	0.00
25	GIRGENTANA	CHITGIR	Italy-Sicily	13.87	37.51	0.74	0.17	10	1.00	0.00	0.00	0.02	1	1.00	0.00	0.00
26	GRIGIA MOLISANA	CHITGMO	Italy	14.40	41.52	0.90	0.29	6	1.00	0.00	0.00	0.02	1	1.00	0.00	0.00
27	ARGENTATA DELL'ETNA	CHITARG	Italy-Sicily	15.00	37.99	0.87	0.02	8	1.00	0.00	0.00	0.02	3	1.00	0.00	0.00
28	POLISH FAWN COLOURED GOAT	CHPLBUK	Poland	18.20	52.47	0.81	0.47	5	1.00	0.00	0.00	0.02	2	1.00	0.00	0.00
29	DUKATI	CHALDUK	Albania	19.50	40.29	0.83	-0.22	7	1.00	0.00	0.00	0.02	3	0.14	0.71	0.14
30	HUNGARIAN NATIVE	CHHUNAT	Hungary	19.90	47.02	0.84	0.02	-9999	1.00	0.00	0.00	-9999	3	0.11	0.44	0.44
31	MUZHAKE	CHALMUZ	Albania	20.21	40.09	0.86	-0.24	6	1.00	0.00	0.00	0.02	3	0.17	0.33	0.50
32	HASI	CHALHAS	Albania	20.45	42.20	0.71	-0.45	8	1.00	0.00	0.00	0.01	2	0.00	0.00	1.00
33	CAPORE	CHALCAP	Albania	20.62	40.88	0.84	-0.28	7	1.00	0.00	0.00	0.02	2	0.67	0.00	0.33
34	CARPATHIAN	CHROCAR	Romania	23.24	46.53	0.89	-0.13	7	1.00	0.00	0.00	0.02	1	0.00	1.00	0.00
35	SKOPELOS	CHGRSKO	Greece	23.31	39.33	0.81	-0.39	8	1.00	0.00	0.00	0.01	3	0.25	0.50	0.25
36	GREEK GOAT	CHGRGRG	Greece	24.04	38.86	0.87	-0.27	14	0.86	0.00	0.14	0.03	3	0.00	0.50	0.50
37	ANGORA	CHTKANG	Turkey	32.15	39.63	0.62	-0.68	6	1.00	0.00	0.00	0.02	1	1.00	0.00	0.00
38	BALADI	CHJOBAL	Jordan	35.64	31.32	0.61	-0.66	6	1.00	0.00	0.00	0.14	-9999	-9999	-9999	-9999
39	HAIR	CHTKHAI	Turkey	38.42	38.74	-9999	-9999	-9999	1.00	0.00	0.00	-9999	2	0.67	0.00	0.33
40	ABAZA	CHTKABA	Turkey	42.90	40.95	0.52	-0.77	8	1.00	0.00	0.00	0.01	1	0.00	0.00	1.00
41	GURCU	CHTKGUR	Turkey	43.24	40.38	0.48	-0.78	8	0.88	0.00	0.13	0.02	1	0.00	0.00	1.00

Molecular data set for goat breeds (2)

ID	breed	breed code	country	longitude	latitude	microsat_ht	microsat_Fis	microsat_mean_nb_alleles	aflp_average_exp_ht	aflp_se_exp_ht	aflp_av_Jac_simindex	aflp_nb_polym_loci	aflp_prc_polym_loci	aflp_freq_recess_genotype
1	BRAVIA	CHPOBRA	Portugal	-8.01	41.56	0.55	0.08	2.75	0.21	0.02	0.66	59	58.4	0.65
2	VERATA	CHSPVRT	Spain	-5.64	40.11	0.57	0.06	2.70	0.20	0.02	0.65	60	59.4	0.64
3	PAYOYA	CHSPPYY	Spain	-5.39	36.84	0.59	0.09	2.95	0.20	0.02	0.69	59	58.4	0.63
4	FLORIDA	CHSPFLR	Spain	-5.16	37.82	0.66	0.02	3.75	0.19	0.02	0.68	62	61.4	0.64
5	MALAGUENA	CHSPMLG	Spain	-4.36	36.87	0.67	0.03	3.26	0.20	0.02	0.62	61	60.4	0.64
6	CABRA DEL GUADARRAMA	CHSPGDR	Spain	-4.11	40.55	0.58	0.08	3.01	0.20	0.02	0.67	66	65.3	0.65
7	PYRENEAN	CHFRPYR	France	1.02	43.34	0.45	0.25	2.77	0.20	0.02	0.65	54	53.5	0.64
8	ROVE	CHFRROV	France	4.99	44.21	0.55	0.09	2.84	0.19	0.02	0.64	58	57.4	0.65
9	FRENCH ALPINE	CHFRALP	France	5.64	45.19	-9999	-9999	-9999	0.19	0.02	0.65	61	60.4	0.65
10	VALDOSTANA	CHITVAL	Italy	7.43	45.71	0.60	0.02	2.78	0.19	0.02	0.72	55	54.5	0.62
11	ST. GALLEN BOOTED GOAT	CHCHSGB	Switzerland	7.93	46.88	0.55	0.02	1.82	0.21	0.02	0.65	56	55.4	0.65
12	SWISS ALPINE	CHCHALP	Switzerland	8.22	46.72	0.59	0.03	2.76	0.18	0.02	0.63	58	57.4	0.65
13	VALAIS BLACK NECK	CHCHVBN	Switzerland	8.36	46.76	0.54	0.01	1.75	0.17	0.02	0.69	54	53.5	0.66
14	GRISONS STRIPED	CHCHGRS	Switzerland	8.41	46.77	0.62	0.01	2.61	0.21	0.02	0.63	58	57.4	0.65
15	PEACOCK GOAT	CHCHPCG	Switzerland	8.53	47.00	0.54	0.03	1.93	0.19	0.02	0.65	56	55.4	0.66
16	GERMAN ALPINE	CHDEBDE	Germany	9.08	49.87	0.63	0.07	2.92	0.20	0.02	0.67	58	57.4	0.65
17	SARDA	CHITSAR	Italy-Sard	9.16	39.55	-9999	-9999	-9999	0.20	0.02	0.63	63	62.4	0.64
18	CORSICAN	CHFRCOR	France	9.21	42.14	0.58	0.05	2.75	0.20	0.02	0.63	60	59.4	0.66
19	OROBICA	CHITORO	Italy	9.56	45.88	0.53	0.05	2.55	0.20	0.02	0.65	55	54.5	0.64
20	CAMOSCIATA DELLE ALPI	CHITCAM	Italy	9.82	46.01	0.60	0.07	2.90	0.20	0.02	0.65	64	63.4	0.65
21	BIONDA DELL'ADAMELLO	CHITBIO	Italy	10.30	45.97	0.62	0.04	2.96	0.20	0.02	0.64	64	63.4	0.64
22	THURINGIAN FOREST GOAT	CHDETWZ	Germany	10.31	50.84	0.61	0.00	2.78	0.20	0.02	0.65	57	56.4	0.65
23	PINZGAUER	CHAUPIZ	Germany	12.77	47.35	0.66	-0.02	2.98	0.21	0.02	0.63	63	62.4	0.63
24	TAUERNSCHECKEN	CHAUTAS	Österreich	13.11	47.27	0.64	0.00	2.99	0.19	0.02	0.72	52	51.5	0.62
25	GIRGENTANA	CHITGIR	Italy-Sicily	13.87	37.51	0.55	0.07	2.48	0.18	0.02	0.71	49	48.5	0.64
26	GRIGIA MOLISANA	CHITGMO	Italy	14.40	41.52	0.63	0.05	3.81	0.20	0.02	0.63	56	55.4	0.65
27	ARGENTATA DELL'ETNA	CHITARG	Italy-Sicily	15.00	37.99	0.68	0.03	3.15	0.21	0.02	0.63	63	62.4	0.63
28	POLISH FAWN COLOURED GOAT	CHPLBUK	Poland	18.20	52.47	0.60	0.01	2.38	0.19	0.02	0.66	55	54.5	0.64
29	DUKATI	CHALDUK	Albania	19.50	40.29	0.69	0.08	3.14	0.23	0.02	0.60	64	63.4	0.66
30	HUNGARIAN NATIVE	CHHUNAT	Hungary	19.90	47.02	0.67	-0.07	3.68	0.21	0.02	0.65	58	57.4	0.65
31	MUZHAKE	CHALMUZ	Albania	20.21	40.09	0.68	0.04	2.52	0.21	0.02	0.62	64	63.4	0.65
32	HASI	CHALHAS	Albania	20.45	42.20	0.66	0.05	3.12	0.20	0.02	0.67	60	59.4	0.64
33	CAPORE	CHALCAP	Albania	20.62	40.88	0.66	0.06	2.98	0.19	0.02	0.66	65	64.4	0.63
34	CARPATHIAN	CHROCAR	Romania	23.24	46.53	0.71	0.04	3.32	0.23	0.02	0.62	69	68.3	0.61
35	SKOPELOS	CHGRSKO	Greece	23.31	39.33	0.61	0.03	2.82	0.19	0.02	0.66	55	54.5	0.65
36	GREEK GOAT	CHGRGRG	Greece	24.04	38.86	0.68	0.05	3.18	0.20	0.02	0.62	59	58.4	0.65
37	ANGORA	CHTKANG	Turkey	32.15	39.63	-9999	-9999	-9999	0.18	0.02	0.68	63	62.4	0.65
38	BALADI	CHJOBAL	Jordan	35.64	31.32	-9999	-9999	-9999	0.21	0.02	0.62	74	73.3	0.64
39	HAIR	CHTKHAI	Turkey	38.42	38.74	-9999	-9999	-9999	0.20	0.02	0.66	69	68.3	0.63
40	ABAZA	CHTKABA	Turkey	42.90	40.95	-9999	-9999	-9999	0.21	0.02	0.66	61	60.4	0.64
41	GURCU	CHTKGUR	Turkey	43.24	40.38	-9999	-9999	-9999	0.20	0.02	0.61	66	65.3	0.64

APPENDIX 13

. . . .
                                                .
                                                .
                                                .

History of logistic regression

The logistic function was invented in the 19th century mainly for the description of the growth of populations, and for the course of autocatalytic chemical reactions where a product itself acts as catalyst while the supply of raw substance is fixed (Cramer, 2002). At the beginning, the «ancestor» of the logistic function was a simple exponential model for population growth in a young country submitted to no constraint which Malthus used in 1789 to claim that such a population would increase in geometric progression (Cramer, 2002, and references therein). A belgian astronomer named Quetelet (1796-1874) knew that this kind of extrapolation based on exponential growth would lead to impossible values and after having first adjusted the function, he then asked Verhulst, his pupil, to carry on with this research. Verhulst introduced an extra term to the function to represent the increasing resistance to further growth. He published a paper in 1845 in the Proceedings of the Belgian Royal Academy in which he named «the logistic function», using examples based on population growth in France, Belgium, Essex and Russia. Interestingly, the logistic function was discovered a second time in 1920 by Raymond Pearl and Lowell J. Reed (Department of Biometry and Vital Statistics at Johns Hopkins University, Baltimore) in the context of a population growth study in the United States. Both were unaware of Verhulst's publications.

Logistic regression and logit link

In any regression, a key parameter is the conditional mean $E(Y|x)$[1], which is the expected value of the Y variable given the value of the independent x variable. In a linear regression, this quantity is expressed as $E(Y|x) = \beta_0 + \beta_1 x$. This expression implies that $E(Y|x)$ may take on any value as x ranges between $-\infty$ and $+\infty$.

In the case of a binary variable (presence versus absence), the result of observations is either a «success» or a «failure» (1 or 0, Bernoulli distribution) and $E(Y|x)$ is expressing a probability. The probability of success is $p = P(Y=1)$ and the probability of failure is $P(Y = 0) = 1-p$. Whatever the value of x, the expected value will range between 0 and 1.

It is not possible to use standard linear regression to calculate a function expressing a relationship between a binomial dependant variable and a quantitative independent variable. The first reason is that the predicted values will become greater than 1 and less than 0 when moving far enough on the x-axis (x ranges between $-\infty$ and $+\infty$) and such values are theoretically inadmissible. Then one assumption of regression is that the variance of Y is constant across values of x what cannot be

---

1. $E(Y|x)$ is read «the expected value of Y, given the value x».

the case with a binary variable[1]. Finally, the error is not normally distributed as Y only takes «0» and «1» values.

We need to find a function that relates the independent variable x to the rolling[2] mean of the bivariate dependant variable, $P(\hat{Y})$[3], and which limits predicted values in a range between 0 and 1.

The logistic function

Let us suppose that we only know a given number of wet days per year, an environmental parameter as independent variable. We want to predict if an AFLP marker is existing or not for a value of 8 wet days per year. We can talk about the probability for this marker to exist or not, or about the odds[4] of existing or not.

Let us consider a probability of 0.9 for the marker to exist for this value of 8 wet days per year :

$$\text{Odds} = \frac{1}{1-p} = \frac{0.9}{1-0.9} = 9$$

This is an odds of 9 to 1.

Now, the odds for the marker of not existing for this value of x is given by :

$$\text{Odds} = \frac{1}{1-p} = \frac{0.1}{1-0.1} = 0.11$$

It should be the opposite odds, what the value of 0.11 doesn't express. This is where the properties of the natural logarithm are used, to express this asymmetry. Indeed, $ln(9) = 2.19$ and $ln(0.11) = -2.19$.

It means that the *ln* odds for an AFLP marker to exist for a given value of x is exactly opposite to the *ln* odds of not existing. This is illustrated by Figure Appendix 13.1 on page XXXIII. Interesting observations are that :

- the natural logarithm is zero when x is 1;
- when x is larger than 1, the natural logarithm curves up slowly;
- when x is less than 1, the natural logarithm is less than 0, and decreases rapidly (vertical asymptote) as x approaches 0.

Consequences are that :

- if p = 0.5, the odds are 0.5 / 0.5 = 1, and $ln(1) = 0$;
- if p > 0.5, $ln(p/(1-p))$ is positive;
- if p < 0.5, $ln(p/(1-p))$ is negative.

---

1. When 50 percent of the markers are existing for a given value of x, then the variance is 0.25 (its maximum value). As we move toward extreme values, the variance decreases. When p = 0.10, the variance is 0.1 * 0.9 = 0.09. So as p approaches 1 or 0, the variance approaches 0.

2. In the sense of «progressing or spreading by stages or by occurrences in different places in succession, with continued or increasing effectiveness».

3. «^» = estimated.

4. The odds is not the same as a probability. It can be found by counting the number of goats with the examined marker for 8 wet days per year for example, and by dividing it by the total number of goats. This is called «cote» in french.

The natural logarithm *ln(x)* is the logarithm having base e, where *e = 2.718281...*
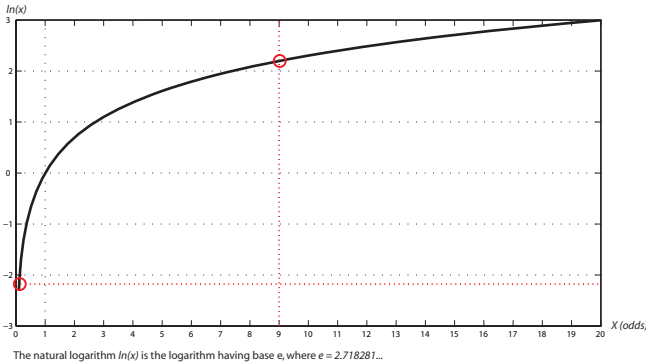
Fig Appendix 13.1. The function of the natural logarithm and the illustration of its properties for a given probability of 0.9 (odds = 9 for 1).

In logistic regression, the dependent variable is a *logit*[1], which is the natural logarithm of the odds, that is :

$$\ln(\text{odds}) = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

So, a logit is a natural logarithm of odds, and odds are a function of p, that is the probability to get a 1. In logistic regression, we find :

$$\text{logit}(p) = \beta_0 + \beta_1 X$$

But we would rather consider probabilities than odds. To get from logits to probabilities, it is necessary to take the natural logarithm out of both sides of the equation :

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

Then we have to convert odds to a simple probability :

$$p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

It is called the *logit transformation.* If the natural logarithm of odds are linearly related to x, then the relation between x and p is not linear, and has the form of the sigmoid curve shown in Figure Appendix 13.2 on page XXXIV.

---

1. The US physicist Joseph Berkson (1899-1982) introduced the probability model that used the logistic curve the 'logit' model, 'logit' standing for LOGistic uniT.
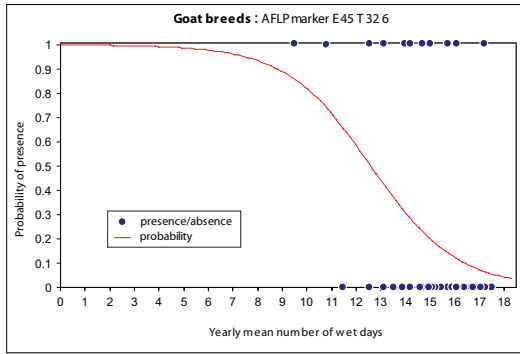
Fig Appendix 13.2. Probability of presence of an AFLP marker in goat breeds according to the yearly mean of wet days, and the corresponding sigmoid response curve. This example was calculated for goats bred in the Alps. There are 27 presences and 111 absences. Source : Econogene AFLP data.

APPENDIX 14

. . . .
 . . . .
                                            .
                                            .
                                            .

Rejection tables

In the tables displayed on the 4 next pages, each cell represents a model. A «0» is displayed when the concerned variable is not significantly participating in explaining the presence of the marker, and a «1» when it is significant, that is when the corresponding p-value is lower than the significance threshold of reference.

The synthetic results presented in chapter 7 for sheep microsatellites are based on 22 integral tables similar to the one displayed hereafter (11 significance thresholds for Wald and 11 for the G test). For sheep AFLP data, the synthetic results are based on 16 integral rejection tables (8 significance thresholds for Wald and 8 for the G test).

Appendix 14

Sheep - microsatellites - *first half of the microsatellite rejection table*

Wald test on 37 alleles (on a total of 744) x 118 environmental variables with a confidence threshold of 1.13E-17.

The first row shows the frequency of the alleles. On the second raw the ID is displayed, and on the third it is possible to read the name of the marker. On the left, the IDs and the names of the geoenvironmental variables are displayed (please refer to appendix 4 to read the entire name and description of the variables).

As 744 alleles were tested against 118 environmental variables, a choice of 37 alleles was made and is presented on this rejection table.

Second half of the microsatellite rejection table

| | | Frequency | 4 | 4 | 5 | 5 | 6 | 7 | 5 | 6 | 8 | 8 | 9 | 11 | 14 | 19 | 10 | 13 | 9 | 14 | 21 | 11 | 23 | 95 | 26 | 21 | 26 | 421 | 26 | 70 | 28 | 81 | 24 | 560 | 41 | 54 | 44 | 110 | 45 | 173 | 46 | 173 | 50 | 336 | 55 | 147 | 62 | 148 | 63 | 151 | 64 | 155 | 67 | 563 | 74 | 322 | 76 | 341 | 77 | 401 | 88 | 285 | 99 | 998 | |
		ID	OAFFCBB3_allele2_109	OARJMP29_allele1_159	OAFFCBB3_allele2_141	DYMS1_allele2_167	ILSTS28_allele1_143	ILSTS28_allele1_145	OAFFCB304_allele1_187	DYMS1_allele1_157	HUJ616_allele1_176	OAFFCBB3_allele1_131	MAF214_allele2_204	MAF214_allele1_184	DYMS1_allele1_173	OAFFCB304_allele2_185	OAFFCB304_allele1_155	OAFFCBB3_allele1_129	MAF214_allele1_184	HUJ616_allele1_120	MAF214_allele2_186	OAFFCB304_allele1_171	OAFFCB304_allele2_171	OAFFCB304_allele2_173	OARJMP29_allele1_170	HUJ616_allele1_116	OARJMP29_allele2_134	HUJ616_allele1_126	HUJ616_allele2_122	SRCRSP9_allele2_118	ILSTS28_allele1_157	DYMS1_allele1_181	ILSTS28_allele2_149	ILSTS28_allele2_163	ILSTS28_allele1_127	OAFFCBB3_allele1_103	OAFFCBB3_allele1_115	Total
Confidence Interval	61	prevaug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
99.9999999999	62	prevsep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	63	prevoct	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	64	prevnov	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	65	prevdec	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	66	prevyear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	67	tmpjan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	68	tmpfeb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	69	tmpmar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	70	tmpapr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	71	tmpmay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	72	tmpjun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	73	tmpjul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	74	tmpaug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	75	tmpsep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	76	tmpoct	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	77	tmpnov	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	78	tmpdec	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	79	tmpyear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	80	rdojan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	2
	81	rdofeb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	82	rdomar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
	83	rdoapr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	84	rdomay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2
	85	rdojun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	86	rdojul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	87	rdoaug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	88	rdosep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
	89	rdooct	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	3
	90	rdonov	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	91	rdodec	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	92	rdoyear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
	93	rehjan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	94	rehfeb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	95	rehmar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	96	rehapr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	97	rehmay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
	98	rehjun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	99	rehjul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	100	rehaug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	101	rehsep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	102	rehoct	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	103	rehnov	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	104	rehdec	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	105	rehyear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	106	sunjan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	107	sunfeb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	108	sunmar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	109	sunapr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	110	sunmay	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	111	sunjun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	112	sunjul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	113	sunaug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	114	sunsep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	115	sunoct	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	116	sunnov	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	117	sundec	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	118	sunyear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		#signifmodels	0	0	1	2	1	5	3	2	0	1	1	0	2	0	1	1	0	0	0	0	1	0	0	0	3	0	0	0	3	0	0	3	0	0	1	2	1
		34																																				

Rejection table for sheep - AFLPs

Wald test on 62 AFLP markers x 118 environmental variables with a confidence interval of 99.999 (1.37E-09)

First half of the AFLP rejection table

The first row shows the frequency of the AFLP marker. On the second raw the ID is displayed, and on the third it is possible to read the name of the marker.

On the left, the IDs and the names of the geoenvironmental variables are displayed (please refer to appendix 4 to read the entire name and description of the variables).

Second half of the AFLP rejection table

Appendix 14

CURRICULUM VITAE

Stéphane Joost

Date of birth	09.05.1968
Nationality	Swiss, Langnau i.E. (BE)
Address	Au Petit Clos, 1985 La Sage, Switzerland
Phone	(home) +41 27 283 12 57 (cell) +41 79 607 78 21
e-mail	stephane.joost@epfl.ch

Education

01/10/1995	Degree in Geography, and Computer Science, Faculty of Arts, University of Lausanne, Switzerland
	Certificate of Geology, Faculty of Sciences, University of Lausanne, Switzerland

Grants and awards

14/02/2005	Association of European GIS Laboratories Grant (AGILE) paper for the 8th AGILE conference in Estoril (Portugal)
28/06/2004	Research Days, ENAC Faculty award, EPFL
01/10/1995	Faculty award for the quality of the degree dissertation dedicated to the development of a GIS applied to the storage of residual hot water in the industry

Research experience

01/09/2001 - ...	Scientific collaborator and Ph.D. student at the Laboratory of Geographical Information Systems (GIS), Ecole Polytechnique Fédérale de Lausanne, Switzerland
01/09/2000 – 31/08/2001	Teaching and research assistant in thematic mapping and GIS at the Department of Geography, University of Geneva
01/09/1999 – 31/08/2000	Teaching and research assistant in transportation geography at the Institute of Geography, University of Lausanne
01/08/1995 – 31/08/1997	Teaching and research assistant at the Section of Computer and Mathematical Methods, University of Lausanne

Professional experience

01/09/1997 – 31/08/2001	Database development and web design at the Computer Centre, University of Lausanne
01/01/1996 – ...	Co-incorporator of MicroGIS Ltd, a company offering spatial analysis services and geographical information technologies developments
01/07/1993 – 31/12/1993	Hired by Etak Inc. (Menlo Park, California) for the development of part of a swiss road geodatabase designed for cars navigation systems

Language skills

French, italian, english, german

Publications

Peer-reviewed publications

Lassueur, T., Joost, S., Randin, C. Very high resolution digital elevation models: do they improve models of plant species distribution?, Ecological Modelling, accepted.

Pariset, L., Cappuccio, I., Joost, S., D'Andrea, M.S., Marletta, D., Ajmone Marsan, P., Valentini, A. and the Econogene Consortium, Characterization of single nucleotide polymorphisms (SNPs) in sheep and their variation as an evidence of selection, submitted to Animal Genetics.

Hussy C., Crivelli R., Joost S., Métral G. (2002) Foreigners in Switzerland, issues and territorial realities, in Revue de Géographie Alpine, 3:81-98, Grenoble.

Joost, S., Dessemontet, P. (2000) Geographische Grenzen der Postleitzahlen und Integration statistischer Daten, VPK/MPG, 9:557-559, Ed. Schweizerischer Verein für Vermessung und Kulturtechnik, Villmergen.

Peer-reviewed conferences

Joost, S., and the Econogene Consortium (2005) Combining biotechnologies and GIScience for livestock genetic resources conservation. Proceedings of the 8th AGILE Conference on Geographic Information Science, Estoril, pp.231-240.

Joost, S., and the Econogene Consortium (2005) Combining biotechnologies and GIScience to contribute to sheep and goat genetic resources conservation. Proceedings of the International Workshop on the role of biotechnology for the characterisation and conservation of crop, forestry, animal and fishery genetic resources. Fondazione per le Biotecnologie, Torino.

Ertz, O., Joost, S., Rappo, D. (2002) Towards Geoservices Portals, WGIS trends for Business Applications, IEEE Computer Society Press, Post-proceedings of the Second International Conference on Web Information Systems Engineering (WISE 2001, Kyoto), 2:102-106, Los Alamitos California.

Joost, S. (2001) Essai de clarification au sein de la hiérarchie des méthodes d'analyse spatiale appliquées aux activités commerciales, Carto 2001, Cartography, Professions and Perspectives, The Canadian Cartographic Association (CCA), Montreal.

Invited conferences

Joost, S. (2005) GIScience and Genetics : Exploiting the geographical dimension of molecular data, Second IPSo Congress on Proteomics and Genomics, Viterbo.

Joost, S., and the Econogene Consortium (2005) Exploratory spatial analysis applied to the investigation of large molecular data sets, Xth spanish conference of biometrics, Oviedo, Editiones de la Universidad de Oviedo.

Invited lecture

Joost, S. (2004) Spatial dimension of genetic information or how to put those boring frequencies on maps, International Summer School of Animal Genomics (ISSAG), Università degli studi della Tuscia, Tuscania.

Book

Pini, G., Joost, S., Widmer, G., Bridel, L. (2000) Interfaces de transport : interfaces de territoires?, Ed. Travaux et recherches de l'IGUL no.18, Conférence Universitaire de Suisse occidentale (CUSO), Lausanne.

Conferences with posters

Ajmone Marsan P., Negrini R., Joost S., Silveri L., Milanesi E., Carloz R., and Econogene Consortium (2004) A map of diversity of european and mediterranean sheep as revealed by AFLP markers, 29th ISAG meeting, Tokyo.

Joost S., Negrini R., Milanesi E., Caloz R., Ajmone Marsan P. and Econogene Consortium (2004) Relationships between ecological distance and genetic distance among goat breeds of Italy and Switzerland, EAAP Conference, Bled, Slovenia.

Negrini R., Joost S., Milanesi E., Bernardi J., Pellecchia M., Patrini M., Caloz R., Ajmone Marsan P. and Econogene Consortium (2004) Genetic diversity of european goats as measured by AFLP markers, EAAP Conference, Bled, Slovenia.

Joost, S., Assessment of environment-genome interaction in animals and plants : applications of Geographical Information Sciences in Population Genetics, Ph.D. students Days, Research Committee of the ENAC Faculty, EPFL, Lausanne, June 2004.

Parisod, C. & Joost, S. (2004) Spatial modelling of gene dispersal and postglacial recolonization in plants : alternative approaches for the evolutionary history of Biscutella laevigata (Brassicaceae) in western Alps, Annual population biology symposium, Department of Biology, University of Fribourg.

Joost, S. & Parisod, C. (2003) Spatial modelling of plant's gene dispersal and western Alps postglacial recolonization, 11th New Phytologist Symposium & Plant Canada 2003, Antigonish.

Parisod, C., Joost, S., Galland, N. et Caloz, R. (2003) Autopolyploidy and post-glacial recolonization of Biscutella laevigata (Brassicaceae): molecular analysis and spatial modeling of gene dispersal in the western Alps, poster session, The evolutionary legacy of the Ice Ages, Discussion meeting of the Royal Society, London.