Section C

Molecular markers – a tool for exploring genetic diversity

1 Introduction

DNA markers are useful in both basic (e.g. phylogenetic analysis and search for useful genes) and applied research (e.g. marker assisted selection, paternity testing and food traceability). This section focuses mainly on their application in characterization of AnGR diversity, and in the

Box 70
DNA, RNA and protein

DNA (deoxyribonucleic acid) is organized in pairs of chromosomes, each inherited from one of the parents. Each gene in an individual, therefore, has two copies, called alleles, one on each chromosome of a pair. In mammals, genes are scattered along chromosomes, separated by long, mainly repetitive, DNA sequences. Genes are formed by coding sequences (exons) separated by introns. The latter carry no protein-coding information, but sometimes play a role in the regulation of gene expression. The instruction encoded by genes is put into action through two processes. The first is transcription (copy) of genetic information into another type of nucleic acid, RNA (ribonucleic acid). Both exons and introns are transcribed into a primary messenger RNA (mRNA) molecule. This molecule is then edited, a process which involves removing the introns, joining the exons together, and adding unique features to each end of the mRNA. A mature mRNA molecule is, thereby, created, which is then transported to structures known as ribosomes located in the cell cytoplasm. Ribosomes are made of ribosomal RNA (rRNA) and proteins, and provide sites for the second process – translation of the genetic information, previously copied to the mRNA, into a polypeptide (an entire protein or one of the chains of a protein complex). The mRNA molecule is read or translated three nucleotides (a codon) at a time. Complementarity between the mRNA codon and the anti-codon of a transfer RNA (tRNA) molecule which carries the corresponding amino acid to the ribosome ensures that the newly formed polypeptide contains the specific sequence of amino acids required.

Not all genes are translated into proteins; some express their function as RNA molecules (such as the rRNA and tRNA involved in translation). Recently, new roles of RNA in the process of mRNA editing and in the regulation of gene expression have been discovered (Storz *et al.*, 2005; Aravin and Tuschl, 2005; Wienholds and Plasterk 2005). Indeed, non-coding RNAs appear to be key players in various regulatory processes (Bertone *et al.*, 2004; Clop *et al.*, 2006). Thus, three types of molecules are available for investigating genetic characteristics at cellular, tissue and whole organism levels: the DNA which contains the encoded instruction; the RNA which transfers the instructions to the cell "factory"; and the proteins which are built according to the instructions, and make functioning cells and organisms.

search for functional variants of relevant genes. It is important to note that RNA and proteins also contain key information, and therefore deserve parallel study; their role in the search for functional variants is also explored below.

Diversity among organisms is a result of variations in DNA sequences and of environmental effects. Genetic variation is substantial, and each individual of a species, with the exception of monozygotic twins, possesses a unique DNA sequence. DNA variations are mutations resulting from substitution of single nucleotides (single nucleotide polymorphisms – SNPs), insertion or deletion of DNA fragments of various lengths (from a single to several thousand nucleotides), or duplication or inversion of DNA fragments. DNA variations are classified as "neutral" when they cause no change in metabolic or phenotypic traits, and hence are not subjected to positive, negative, or balancing selection; otherwise, they are referred to as "functional". Mutations in key nucleotides of a coding sequence may change the amino acid composition of a protein, and lead to new functional variants. Such variants may have an increased or decreased metabolic efficiency compared to the original "wild type", may lose their functionality completely, or even gain a novel function. Mutations in regulatory regions may affect levels and patterns of gene expression; for example, turning genes on/off or under/over-expressing proteins in specific tissues at different development or physiological stages.

Although analysis of single types of biomolecules has proven extremely useful in understanding biological phenomena, the parallel large-scale investigation of DNA, RNA and proteins opens up new perspectives in the interpretation and modelling of the complexity of living organisms. New scientific disciplines with the suffix "–omics" are coming into existence. In these fields, recent advances in the preparation, identification and sequencing of DNA, RNA and proteins, and in large-scale data storage and analysis, are bringing about a revolution in our understanding. A global, integrated view of an entire set of biological molecules involved in complex biological processes

Box 71
The new "–omics" scientific disciplines

Genomics charts genes and the genetic variations among individuals and groups. It provides an insight into the translation of genetic information to metabolic functions and phenotypic traits. It unveils biological processes and their interactions with environmental factors. Genomics involves the combination of a set of high-throughput technologies, such as proteomics and metabolomics, with the bioinformatic techniques that enable the processing, analysis and integration of large amounts of data.

Box 72
Recent developments in molecular biology

Current revolutionary developments in molecular biological research relevant to livestock breeding and genetic diversity conservation include:

1. establishment of the entire genome sequence of the most important livestock species;
2. development of technology to measure polymorphisms at loci spread all over the genome (e.g. methods to detect SNPs); and
3. development of microarray technology to measure gene transcription at a large scale.

Information obtained through the sequencing of the entire genome (achieved for chickens and almost complete for pigs and cattle), integrated with SNP technology, will speed up the search for genes. Quantitative trait loci (QTL) mapping to identify chromosome regions influencing a target trait, the presence of candidate genes located in the same region, and investigation of their patterns of expression (e.g. by microarray and proteomic analyses) and their function across species, will come together to identify key genes and to unravel the complexity of physiological regulation for target traits.

See below for further discussion of these developments.

is emerging. Structural genomics, transcriptomics and proteomics are followed by metabolomics, and interactomics among others, and at a still higher level of complexity, systems biology (Hood *et al.*, 2004; Box 71).

Investigation of biological complexity is a new frontier which requires high-throughput molecular technology, high computer speed and memory, new approaches to data analysis, and integration of interdisciplinary expertise (Box 72).

2 The roles of molecular technologies in characterization

Information on genetic diversity is essential in optimizing both conservation and utilization strategies for AnGR. As resources for conservation are limited, prioritization is often necessary. New molecular tools hold the promise of allowing the identification of genes involved in a number of traits, including adaptive traits, and polymorphisms causing functional genetic variation (QTN – Quantitative Trait Nucleotides). However, we do not have sufficient knowledge to prioritize conservation choices on the basis of functional molecular diversity, and alternative measures are still needed. Phenotypic characterization provides a crude estimate of the average of the functional variants of genes carried by a given individual or population. However, the majority of phenotypes of the majority of livestock species are not recorded.

First role. In the absence of reliable phenotype and QTN data, or to complement the existing data, the most rapid and cost-effective measures of genetic diversity are obtained from the assay of polymorphisms using anonymous molecular genetic markers. Anonymous markers are likely to provide indirect information on functional genes for important traits, assuming that unique populations that have had a particular evolutionary history at the neutral markers (e.g. because of ancient isolation or independent domestication) are likely to carry unique variants of functional variations. Molecular techniques have also proved useful in the investigation of the origin and domestication of livestock species, and their subsequent migrations, as well as providing information on evolutionary relationships (phylogenetic trees), and identifying geographical areas of admixture among populations of different genetic origins. Subchapter 3.1 presents an outline of molecular techniques for the assessment of genetic diversity within and between breeds.

Second role. Effective population size (N_e) is an index that estimates the effective number of animals in a population that reproduce and contribute genes to the next generation. N_e is closely linked to the level of inbreeding and genetic drift in a population, and therefore is a critical indicator for assessing the degree of endangerment of populations (see Sections A and F). Traditional approaches to obtaining reliable estimates of N_e for breeding populations are based on pedigree data or censuses. The necessary data on variability of reproductive success and generation intervals are often not reliably available for populations in developing countries. Molecular approaches may, therefore, be a promising alternative (see subchapter 3.2 for further details).

Third role. A top priority in the management of AnGR is the conservation of breeds that have unique traits. Among these, the ability to live and produce in challenging conditions, and to resist infectious diseases are of major importance, particularly for developing countries. Complex traits, such as adaptation and disease resistance, are not visible or easily measurable. They can be investigated in experiments in which the animals are submitted to the specific environmental conditions or are infected with the relevant agent. However, such experiments are difficult and expensive to perform, and raise concerns about animal welfare. This is the reason why researchers are extremely interested in identifying genes controlling complex traits. Such genes can be sought by a number of different approaches.

Tools being developed to target functional variation are described in subchapter 3.3.

3 Overview of molecular techniques

This section describes the most important molecular techniques currently being utilized and developed for the assessment of genetic diversity, and for targeting functional variation. Box 73 describes how DNA and RNA are extracted from biological material and prepared for analysis. The attributes of commonly used molecular markers are outlined in Box 74, and sampling (a very important aspect of molecular studies) is discussed in Box 75.

Protein polymorphisms were the first markers used for genetic studies in livestock. However, the number of polymorphic loci that can be assayed, and the level of polymorphisms observed at the loci are often low, which greatly limits their application in genetic diversity studies. With the development of new technologies, DNA polymorphisms have become the markers of choice for molecular-based surveys of genetic variation (Box 74).

3.1 Techniques using DNA markers to assess genetic diversity

Nuclear DNA markers

A number of markers are now available to detect polymorphisms in nuclear DNA. In genetic diversity studies, the most frequently used markers are microsatellites.

Microsatellites

Currently, microsatellites (Box 74) are the most popular markers in livestock genetic characterization studies (Sunnucks, 2001). Their high mutation rate and codominant nature permit the estimation of within and between-breed genetic diversity, and genetic admixture among breeds even if they are closely related.

Box 73
Extraction and multiplication of DNA and RNA

The first step in DNA, RNA and protein analysis is extraction and purification from biological specimens. Several protocols and commercial kits are available. The strategies applied depend on the source material and the target molecule. For example, DNA extraction from whole blood or white cells is relatively easy, while its extraction from processed food is rather difficult. RNA extraction from pancreatic tissue is difficult because of very rapid post-mortem degradation in this organ. Purity of DNA, RNA and proteins is often a key neglected factor in obtaining reliable results.

After isolating DNA (or RNA) from cells, the next step is to obtain thousands or millions of copies of a particular gene or piece of DNA. DNA fragment multiplication can be delegated to micro-organisms, typically *E. coli*, or accomplished *in vitro* using a polymerase chain reaction (PCR). This technique, which won the Nobel Prize for its inventor, Cary Mullis, exponentially amplifies any DNA segment of known sequence. The key component in a PCR reaction is the DNA polymerase isolated from *Thermus aquaticus*, a micro-organism adapted to live and multiply at very high temperature. This thermostable *Taq*- (after *Thermus aquaticus*) polymerase permits chain replication in cycles and produces a geometric growth in the number of copies of the target DNA. A PCR cycle includes three steps: i) DNA denaturation at 90–95 °C to separate the DNA into two single strands to serve as a template; ii) annealing of a pair of short single-strand oligonucleotides (primers) complementary to the target regions flanking the fragment of interest, at 45–65 °C; iii) extension or elongation of newly synthesized DNA strands led by primers and facilitated by the *Taq*-polymerase, at 72 °C. This cycle can be repeated, normally 25 to 45 times, to enable amplification of enough amplicons (a fragment of a gene or DNA synthesized using PCR) to be detected.

Box 74
Commonly used DNA markers

Restriction fragment length polymorphisms (RFLPs) are identified using restriction enzymes that cleave the DNA only at precise "restriction sites" (e.g. EcoRI cleaves at the site defined by the palindrome sequence GAATTC). At present, the most frequent use of RFLPs is downstream of PCR (PCR–RFLP), to detect alleles that differ in sequence at a given restriction site. A gene fragment is first amplified using PCR, and then exposed to a specific restriction enzyme that cleaves only one of the allelic forms. The digested amplicons are generally resolved by electrophoresis.

Microsatellites or SSR (Simple Sequence Repeats) or STR (Simple Tandem Repeats) consist of a stretch of DNA a few nucleotides long – 2 to 6 base pairs (bp) – repeated several times in tandem (e.g. CACACACACACACACA). They are spread over a eukaryote genome. Microsatellites are of relatively small size, and can, therefore, be easily amplified using PCR from DNA extracted from a variety of sources including blood, hair, skin or even faeces. Polymorphisms can be visualized on a sequencing gel, and the availability of automatic DNA sequencers allows high-throughput analysis of a large number of samples (Goldstein and Schlötterer, 1999; Jarne and Lagoda, 1996). Microsatellites are hypervariable; they often show tens of alleles at a locus that differ from each other in the numbers of the repeats. They are still the markers of choice for diversity studies as well as for parentage analysis and Quantitative Trait Loci (QTL) mapping, although this might be challenged in the near future with the development of cheap methods for the assay of SNPs. FAO has published recommendations for sets of microsatellite loci to be used for diversity studies for major livestock species, which were developed by the ISAG–FAO Advisory Group on Animal Genetic Diversity (see DAD-IS library http://www.fao.org/dad-is/).

Minisatellites share the same characteristics as microsatellites, but the repeats are ten to a few hundreds bp long. Micro and minisatellites are also known as VNTRs (Variable Number of Tandem Repeats) polymorphisms.

Amplified fragment length polymorphisms (AFLPs) are a DNA fingerprinting technique which detects DNA restriction fragments by means of PCR amplification.

STS (Sequence Tagged Site) are DNA sequences that occur only once in a genome, in a known position. They needn't be polymorphic and are used to build physical maps.

SNPs are variations at single nucleotides which do not change the overall length of the DNA sequence in the region. SNPs occur throughout the genome. They are highly abundant and are present at one SNP in every 1000 bp in the human genome (Sachinandam *et al.*, 2001). Most SNPs are located in non-coding regions, and have no direct impact on the phenotype of an individual. However, some introduce mutations in expressed sequences or regions influencing gene expression (promoters, enhancers), and may induce changes in protein structure or regulation. These SNPs have the potential to detect functional genetic variation.

Some controversy has surrounded the choice of a mutation model – infinite allele or step-wise mutation model (Goldstein et al, 1995) – for microsatellite data analysis. However, simulation studies have shown that the infinite allele mutation model is generally valid for assessment of within-species diversity (Takezaki and Nei, 1996).

The mean number of alleles (MNA) per population, and observed and expected heterozygosity (Ho and He), are the most common parameters for assessing within-breed diversity. The simplest parameters for assessing diversity among breeds are the genetic differentiation or fixation indices. Several estimators have been

Box 75
Sampling genetic material

Sample collection is the first and the most important step in any diversity study. Ideally, samples should be unrelated and representative of the populations under investigation. Generally, the sampling of 30 to 50 well-chosen individuals per breed is considered sufficient to provide a first clue as to breed distinctiveness and within-breed diversity, if a sufficient number of independent markers is assayed (e.g. 20–30 microsatellites; Nei and Roychoudhury, 1974; Nei, 1978). However, the actual numbers required may vary from case to case, and may be even lower in the case of a highly inbred local population, and higher in a widely spread population divided into different ecotypes.

The choice of unrelated samples is quite straightforward in a well-defined breed, where it can be based on the herd book or pedigree record. Conversely, it can be rather difficult in a semi-feral population for which no written record is available. In this case, the use of a geographic criterion is highly recommendable, i.e. to collect a single or very few (unrelated) animals per flock from a number of flocks spread over a wide geographic area. The record of geographical coordinates, and photo-documentation of sampling sites, animals and flocks is extremely valuable – to check for cross-breeding in the case of unexpected outliers, or for identifying interesting geographic patterns of genetic diversity. A well-chosen set of samples is a long-lasting valuable resource, which can be used to produce meaningful results even with poor technology. Conversely, a biased sample will produce results that are distorted or difficult to understand even if the most advanced molecular tools are applied.

proposed (e.g. FST and GST), the most widely used being FST (Weir and Basten, 1990), which measure the degree of genetic differentiation of subpopulations through calculation of the standardized variances in allele frequencies among populations. Statistical significance can be calculated for the FST values between pairs of populations (Weir and Cockerham, 1984) to test the null hypothesis of a lack of genetic differentiation between populations and, therefore, the partitioning of genetic diversity (e.g. Mburu *et al.*, 2003). Hierarchical analysis of molecular variance (AMOVA) (Excoffier *et al.*, 1992) can be performed to assess the distribution of diversity within and among groups of breeds.

Microsatellite data are also commonly used to assess genetic relationships between populations and individuals through the estimation of genetic distances (e.g. Beja-Pereira *et al.*, 2003; Ibeagha-Awemu *et al.*, 2004; Joshi *et al.*, 2004; Sodhi *et al.*, 2005; Tapio *et al.*, 2005). The most commonly used measure of genetic distances is Nei's standard genetic distance (DS) (Nei, 1972). However, for closely related populations, where genetic drift is the main factor of genetic differentiation, as is often the case in livestock breeds, particularly in the developing world, the modified Cavalli-Sforza distance (DA) is recommended (Nei *et al.*, 1983). Genetic relationship between breeds is often visualized through the reconstruction of a phylogeny, most often using the neighbour-joining (N-J) method (Saitou and Nei, 1987). However, a major drawback of phylogenetic tree reconstruction is that the evolution of lineages is assumed to be non-reticulate, i.e. lineages can diverge, but can never result from crosses between lineages. This assumption will rarely hold for livestock, where new breeds often originate from cross-breeding between two or more ancestral breeds. The visualization of the evolution of breeds provided by phylogenetic reconstruction must, therefore, be interpreted cautiously.

Multivariate analysis, and more recently Bayesian clustering approaches, have been suggested for admixture analysis of microsatellite data from different populations (Pritchard *et al.*, 2000). Probably the most comprehensive study of this type in livestock is a continent-wide study of African cattle (Hanotte *et al.*, 2002), which reveals the genetic signatures of the origins,

secondary movements, and differentiation of African cattle pastoralism.

Molecular genetic data, in conjunction with, and complemented by, other sources such as archaeological evidence and written records, provide useful information on the origins and subsequent movements and developments of genetic diversity in livestock species. Mapping the origin of current genetic diversity potentially allows inferences to be made about where functional genetic variation might be found within a species for which only limited data on phenotypic variation exist.

Combined analysis of microsatellite data obtained in separate studies is highly desirable, but has rarely been possible. This is because most population genetic studies using DNA markers are limited to small numbers of breeds, often from a single country (Baumung *et al.*, 2004). Often, different subsets of the FAO-recommended markers are used, and no standard samples are genotyped across projects. The application of different microsatellite genotyping systems causes variation between studies in the estimated size of alleles at the same loci. To promote the use of common markers, FAO is now proposing an updated, ranked list of microsatellite loci for the major livestock species[3]. FAO recommends the use of the markers in the order of ranking, to maximize the number of markers overlapping among independent investigations. For some species, DNA from standard animals is available. For example, aliquots of sheep and goat standard DNA used in the European Union (EU) Econogene project have been distributed to other large-scale projects in Asia and Africa, and can be requested through the Econogene Website (http://www.econogene.eu).

There are only a few examples of large-scale analyses of the genetic diversity of livestock species. Hillel *et al.* (2003) and SanCristobal *et al.* (2006a) investigated, respectively, chicken and pig diversity throughout Europe; Hanotte *et al.* (2002) obtained data on cattle at the scale of almost the entire African continent; Tapio *et al.* (2005) assessed sheep diversity at a large regional scale in northern European countries; and Cañon *et al.* (2006) studied goat diversity in Europe and the Near and Middle East. However, for most species, a comprehensive review is still lacking. Ongoing close coordination between large-scale projects promises the delivery of a global estimate of genetic diversity in the near future for some species such as sheep and goats. In the meantime, new methods of data analysis are being developed to permit the meta-analysis of datasets that have only a few breeds and no, or only a few, markers in common (Freeman *et al.*, 2006). This global perspective on livestock diversity will be extremely valuable to reconstruct the origin and history of domestic animal populations and, indirectly, of human populations. It will also highlight regional and local hotspots of genetic diversity which may be targeted by conservation efforts.

SNPs

SNPs (Box 74) are used as an alternative to microsatellites in genetic diversity studies. Several technologies are available to detect and type SNP markers (see Syvänen, 2001, for a review). Being biallelic markers, SNPs have rather low information content, and larger numbers have to be used to reach the level of information obtained from a standard panel of 30 microsatellite loci. However, ever-evolving molecular technologies are increasing automation and decreasing the cost of SNP typing. This is likely, in the near future, to permit the parallel analysis of a large number of markers at a lower cost. With this perspective, large-scale projects are ongoing in several livestock species to identify millions (e.g. Wong *et al.*, 2004) and validate several thousands of SNPs, and identify haplotype blocks in the genome. Like sequence information, SNPs permit a direct comparison and joint analysis of different experiments.

[3] Lists and guidelines can be found in the DAD-IS library at http://www.fao.org/dad-is.

PART **4**

SNPs seem to be appealing markers to apply in the future for genetic diversity studies because they can easily be used in assessing either functional or neutral variation. However, the preliminary phase of SNP discovery or SNP selection from databases is critical. SNPs can be generated through various experimental protocols, such as sequencing, single-stranded conformational polymorphism (SSCP) or denaturing high-performance liquid chromatography (DHPLC), or *in silico*, aligning and comparing multiple sequences of the same region from public genome and expressed sequence (EST) databases. When data have not been obtained randomly, standard estimators of population genetic parameters cannot be applied. A frequent example is when SNPs initially identified in a small sample (panel) of individuals are then typed in a larger sample of chromosomes. By preferentially sampling SNPs at intermediate frequencies, such a protocol will bias the distribution of allelic frequencies compared to the expectation for a random sample. SNPs do hold promise for future application in population genetic analyses; however, statistical methods that can explicitly take into account each method of SNP discovery have to be developed (Nielsen and Signorovitch, 2003; Clark *et al*., 2005).

AFLPs

AFLPs are dominant biallelic markers (Vos *et al*., 1995). Variations at many loci can be arrayed simultaneously to detect single nucleotide variations of unknown genomic regions, in which a given mutation may be frequently present in undetermined functional genes. However, a disadvantage is that they show a dominant mode of inheritance; this reduces their power in population genetic analyses of within-breed diversity and inbreeding. Nevertheless, AFLP profiles are highly informative in assessing the relationship between breeds (Ajmone-Marsan *et al*., 2002; Negrini *et al*., 2006; De Marchi *et al*., 2006; SanCristobal *et al*., 2006b) and related species (Buntjer *et al*., 2002).

Mitochondrial DNA markers

Mitochondrial DNA (mtDNA) polymorphisms have been extensively used in phylogenetic and genetic diversity analyses. The haploid mtDNA, carried by the mitochondria in the cell cytoplasm, has a maternal mode of inheritance (individuals inherit the mtDNA from their dams and not from their sires) and a high mutation rate; it does not recombine. These characteristics enable biologists to reconstruct evolutionary relationships between and within species by assessing the patterns of mutations in mtDNA. MtDNA markers may also provide a rapid way of detecting hybridization between livestock species or subspecies (e.g. Nijman *et al*., 2003).

The polymorphisms in the sequence of the hypervariable region of the D-loop or control region of mtDNA have contributed greatly to the identification of the wild progenitors of domestic species, the establishment of geographic patterns of genetic diversity, and the understanding of livestock domestication (see Bruford *et al*., 2003, for a review). For example, the Middle Eastern origin of modern European cattle was recently demonstrated by Troy *et al*. (2001). The study identified four maternal lineages in *Bos taurus* and also demonstrated the loss of bovine genetic variability during the human Neolithic migration out of the Fertile Crescent. In the same way, multiple maternal origins with three mtDNA lineages were highlighted in goats (Luikart *et al*., 2001), with Asia and the Fertile Crescent as possible centres of origin. Recently, a third mtDNA lineage was discovered in native Chinese sheep (Guo *et al*., 2005), a fourth in native Chinese goats (Chen *et al*., 2005), and a fifth in Chinese cattle (Lai *et al*., 2006). In Asian chickens, nine different mtDNA clades have been found (Liu *et al*., 2006), suggesting multiple origins in South and Southeast Asia. All these results indicate that our current knowledge of livestock domestication and genetic diversity remains far from complete. For further discussion of the origins of domestic livestock species see Part 1 – Section A.

3.2 Using markers to estimate effective population size

Hill (1981) suggested using gametic phase disequilibrium of DNA polymorphisms to estimate effective population size (N_e). This estimation can be based on genotypes for linked markers (microsatellites or SNPs). The expected correlation of allele frequencies at linked loci is a function of N_e and the recombination rate. N_e can, therefore, be estimated from the observed disequilibrium. Hayes *et al*. (2003) suggested a similar approach based on chromosome segment homozygosity, which, in addition, has the potential to estimate N_e for earlier generations, and therefore allows a judgement of whether an existing population was of increasing or decreasing size in the past. The study demonstrated, with example data sets, that the Holstein-Friesian cattle breed underwent a substantial reduction of N_e in the past, while the effective population size of the human population is increasing, which is in agreement with both census and pedigree studies.

3.3 Molecular tools for targeting functional variation

Approaches based on map position: quantitative trait loci (QTL) mapping

Genetic markers behave as Mendelian traits; in other words, they follow the laws of segregation and independent assortment first described by Mendel. Two genes that are located on the same chromosome are physically linked and tend to be inherited together. During meiosis, recombination between homologous chromosomes may break this linkage. The frequency of recombination between two genes located on the same chromosome depends of the distance between them. Recombination rate between markers is, therefore, an indication of their degree of linkage: the lower the recombination rate, the closer the markers. The construction of genetic maps exploits this characteristic to infer the likely order of markers and the distance between them.

Mapping exercises are generally accomplished following the co-segregation of polymorphic markers in structured experimental populations (e.g. F2 or backcross) or existing populations

Box 76
QTL mapping

If a QTL for a target trait exists, the plus- and minus-variant allele of the unknown responsible gene (Q and q) will co-segregate with the alleles at a nearby M1 marker (M1 and m1) that we are able to genotype in the laboratory. Let us hypothesize that M1 co-segregates with Q and m1 with q, that is M1 and Q are nearby on a same chromosome and m1 and q on the homologous chromosome (M1Q and m1q).

Let us also assume that an F2 population derived by the mating of heterozygous F1 individuals is genotyped. Following the genotyping, F2 progenies are grouped on the base of their marker genotype (M1M1 and m1m1; M2M2 and m2m2; ... MnMn and mnmn), and afterwards the average phenotype of the groups is compared. If no QTL is linked to a given marker (e.g. M2), then no significant difference will be detected between the average phenotypic value of the M2M2 and m2m2 progenies for the target trait. Conversely, when progenies are grouped by their genotype at the marker M1, then the group M1M1 will mostly be QQ at the QTL, and the group m1m1 will mostly be qq. In this case, a significant difference is observed between progeny averages, and therefore the presence of a QTL is detected. In species, such as poultry and pigs, where lines and breeds are commonly interbred commercially, this exercise can be accomplished in experimental populations (F2, BC) while in ruminants two (daughter design – DD) or three (grand-daughter design – GDD) generation pedigrees are generally used. In DD the segregation of markers heterozygous in a sire (generation I) is followed in the daughters (generation II) on which phenotypic data are collected. In GDD, the segregation of markers heterozygous in a grand-sire (generation I) are followed in his half-sib sons (generation II), whose phenotype is inferred from those of the grand-daughters (generation III).

under selection programmes (families of full siblings or half siblings). Medium to high density genetic maps of a few hundred to a few thousand markers are available for most livestock species.

To identify a QTL for a given trait, a family segregating for the trait is genotyped with a set of mapped molecular markers evenly spread over the genome (Box 76). A number of statistical methods exist to infer the presence of a significant QTL at a given marker interval, but all rely on the fact that families possess a high level of linkage disequilibrium, i.e. large segments of chromosomes are transmitted without recombination from parents to progeny.

The result of a QTL mapping experiment is the identification of a chromosome region, often spanning half of a chromosome, in which a significant effect is detected for the target trait. Modern research is actively using mapping to identify QTL influencing adaptive traits. Examples of such traits include, in chickens, increased resistance to *Salmonella* colonization and excretion (Tilquin *et al.*, 2005), and susceptibility to develop pulmonary hypertension syndrome (Rabie *et al.*, 2005); and in cattle, trypanotolerance (Hanotte *et al.*, 2002).

The QTL mapping phase is generally followed by the refinement of the map position of the QTL (QTL fine mapping). To accomplish this task, additional markers, and above all additional recombination events in the target area, are analysed. A clever approach has recently been designed and applied to the fine mapping of a chromosome region on BTA14 carrying a significant QTL for milk fat percentage and other traits (Farnir *et al.*, 2002). This approach exploits historical recombination in past generations to restrict the map position to a relatively small 3.8 cM (centimorgan) region, a size that has permitted the positional cloning of the gene (DGAT1) (Grisart *et al.*, 2002).

Following fine mapping, the genes determining the performance trait can be sought among the genes that are located in the regions identified. Candidate genes may be sought in the same species (e.g. when a rich EST map is available or when the genome is fully sequenced) or in orthologous regions of a model organism for which complete genome information is available.

Occasionally, key information on gene function arrives from an unexpected source. This was the case with the myostatin gene, the function of which was first discovered in mice and then found to be located in cattle in the chromosomal region where the double-muscling gene had previously been mapped (McPherron and Lee, 1997).

It is clear that identifying the responsible gene (quantitative trait genes – QTG) and the functional mutation (QTN) of a complex trait is still a substantial task, and several approaches are needed to decrease the number of positional candidate genes. Information on gene function is fundamental in this respect. However, we are still ignorant about the possible function(s) of the majority of genes identified by genome and cDNA (complementary DNA) sequencing. This is why the investigation of patterns of gene expression may provide useful information, in combination with the positional approach previously described, to identify candidate genes for complex traits. This combined approach is referred to as genetical genomics (Haley and de Koning, 2006). New advances in the investigation of patterns of gene expression are described in the next section.

Alternative approaches are presently being investigated to detect adaptive genes using genetic markers (Box 77). They are now at the experimental stage, and only further research will permit an evaluation of their efficacy.

The ultimate goal of QTL mapping is to identify the QTG, and eventually the QTN. Although only a few examples exist to date in livestock, these are the kind of mutations that could have a direct impact on marker assisted breeding and on conservation decision-making. Conservation models considering functional traits and mutation need to be developed, as an increasing number of QTG and QTN will be uncovered in the near future.

Investigating patterns of gene expression

In the past, the expression of specific traits, such as adaptation and resistance, could only be measured at the phenotypic level. Nowadays, the

Box 77
The population genomics approach

An alternative approach to the identification of genome regions carrying relevant genes has recently been proposed. It consists of the detection of "selection signatures" via a "population genomics" approach (Black *et al.*, 2001; Luikart *et al.*, 2003). Three main principles of the population genomics approach to QTL mapping are that:

1. neutral loci across the genome will be similarly affected by genetic drift, demography, and evolutionary history of populations;
2. loci under selection will often behave differently and, therefore, reveal "outlier" patterns of variation, loss of diversity (increase of diversity if the loci were under a balanced selection), linkage disequilibrium, and increased/decreased Gst/Fst indices; and
3. through hitchhiking effects, selection will also influence linked markers, allowing the detection of a "selection signature" (outlier effects), which can often be detected by genotyping a large number of markers along a chromosome and identifying clusters of outliers. This approach utilizes phenotypic data at the breed level (or subpopulations within a breed), rather than at the individual level, and thereby nicely complements classical QTL mapping approaches within pedigrees.

The population genomics approach can also identify genes subjected to strong selection pressure and eventually fixed within breeds, and in particular, genes involved in adaptation to extreme environments, disease resistance, etc. Many of these traits, which are of great importance to the sustainability of animal breeding, are difficult or impossible to investigate by classic QTL mapping or association study approaches. The potential of population genomics has recently been investigated from a theoretical point of view (Beaumont and Balding, 2004; Bamshad and Wooding, 2003), and through experimental work with different types of markers in natural populations (AFLPs: Campbell and Bernatchez, 2004; microsatellites: Kayser *et al.*, 2003; SNPs: Akey *et al.*, 2002). The approach has recently been applied within the Econogene project (http://lasig.epfl.ch/projets/econogene). In preliminary analyses, three SNPs in MYH1 (myosin 1), MEG3 (callypige), and CTSB (cathepsin B) genes in sheep have shown significant outlier behaviour (Pariset *et al.*, 2006).

Within the same project, a novel approach based on Spatial Analysis Method (SAM) has been designed to detect signatures of natural selection within the genome of domestic and wild animals (Joost, 2006). Preliminary results obtained with this method are in agreement with those obtained by the application of theoretical models in population genetics, such as those developed by Beaumont and Balding (2004). SAM goes a step further compared to classical approaches, as it is designed to identify environmental parameters associated with selected markers.

transcriptome (the ensemble of all transcripts in a cell or tissue), and the proteome (the ensemble of all proteins) can be directly investigated by high-throughput techniques, such as differential display (DD) (Liang and Pardee, 1992), cDNA-AFLP (Bachem *et al.*, 1996), serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995; 2000), mass spectrometry, and protein and DNA microarrays. These techniques represent a breakthrough in RNA and protein analysis, permitting the parallel analysis of virtually all genes expressed in a tissue at a given time. Thus, the techniques contribute to the decoding of the networks that are likely to underlie many complex traits.

-Omics technologies are often compared to turning on the light in front of a Michelangelo fresco rather than using a torch that permits a view only of parts of the whole. The overall view allows the meaning of the representation to be understood and its beauty to be appreciated. In reality, the power of these techniques is paralleled

at present by the difficulty and cost involved in applying them and in analyzing the data produced. The isolation of homogeneous cell samples is rather difficult, and is an important prerequisite in many gene expression profiling studies. The large number of parallel assays results in low cost per assay, but at a high cost per experiment. Equipment is expensive, and high technical skill is needed in all experimental phases. This is in addition to the general difficulty in analysing RNA compared to DNA. RNA is very sensitive to degradation, and particular care has to be taken while extracting it from tissues that have a very active metabolism. Indeed, sample conservation and manipulation is one of the keys to success in RNA analysis experiments. The application of nanotechnologies to the analysis of biological molecules is opening up very promising perspectives in solving these problems (Sauer *et al.*, 2005).

Data handling is a further problem. Molecular datasets such as gene expression profiles can be produced in a relatively short time. However, the standardization of data between laboratories is needed for consistent analysis of different biological datasets. Agreements on standardization, as well as the creation of interconnected databases, are essential for the efficient analysis of molecular networks.

Transcript profiling

This section briefly describes SAGE and microarray techniques. Descriptions of other techniques may be found in a number of recent reviews (e.g. Donson *et al.*, 2002). SAGE generates complete expression profiles of tissues or cell lines. It involves the construction of total mRNA libraries which enable a quantitative analysis of the whole transcripts expressed or inactivated at particular steps of a cellular activation. It is based on three principles: (i) a short sequence tag (9–14 bp) obtained from a defined region within each mRNA transcript contains sufficient information to uniquely identify one specific transcript; (ii) sequence tags can be linked together to form long DNA molecules (concatemers) which can be cloned and sequenced – sequencing of the concatemer clones results in the quick identification of numerous individual tags; (iii) the expression level of the transcript is quantified by the number of times a particular tag is observed.

Microarrays can be used to compare, in a single experiment, the mRNA expression levels of several thousands of genes between two biological systems, for example, between animals in a normal environment and animals in a challenging environment. Microarray technology can also provide an understanding of the temporal and spatial patterns of expression of genes in response to a vast range of factors to which the organism is exposed.

Very small volumes of DNA solution are printed on a slide made of a non-porous material such as glass, creating spots that range from 100 to 150 µm in diameter. Currently, about 50 000 complementary DNAs (cDNAs) can be robotically spotted onto a microscope slide. DNA microarrays contain several hundreds of known genes, and a few thousands of unknown genes. The microarray is spotted with cDNA fragments or with prefabricated oligonucleotides. The latter option has the advantage of a higher specificity and reproducibility, but can be designed only when the sequence is known. Microarray use is based on the principle of "hybridization", i.e. the exposure of two single-stranded DNA, or one DNA and one RNA, sequences to each other, followed by the measurement of the amount of double-stranded molecule formed. The expression of mRNA can be measured qualitatively and quantitatively. It indicates gene activity in a tissue, and is usually directly related to the protein production induced by this mRNA.

Gene expression profiling contributes to the understanding of biological mechanisms, and hence facilitates the identification of candidate genes. The pool of genes involved in the expression of trypanotolerance in cattle, for example, has been characterized by SAGE (Berthier *et al*., 2003), and by cDNA microarray analysis (Hill *et al*., 2005). The parallel investigation of the expression of

many genes may permit the identification of master genes responsible for phenotypic traits that remain undetected by differential expression analysis. These master genes may, for instance, possess different alleles all expressed at the same level, which promote the expression of downstream genes with different efficiency. In this case, the master gene can be sought either by exploiting current knowledge of metabolic pathways, or via an expression QTL (eQTL) approach (Lan *et al*., 2006). In this approach, the level of expression of the downstream genes is measured in a segregating population. The amount of transcript of each gene is treated as a phenotypic trait, and QTL that influence the gene expression can be sought using methodologies described above. It is worth noting that data analysis for the detection of QTL is still quite difficult to master. This is also true for transcript profiling techniques because of the many false signals that occur.

Protein profiling

The systematic study of protein structures, post-translational modifications, protein profiles, protein–protein, protein–nucleic acid, and protein–small molecule interactions, and the spatial and temporal expression of proteins in eukaryotic cells, are crucial to understanding complex biological phenomena. Proteins are essential to the structure of living cells and their functions.

The structure of a protein can be revealed by the diffraction of x-rays or by nuclear magnetic resonance spectroscopy. The first requires a large amount of crystalline protein, and this is often restrictive. In order to understand protein function and protein–protein interactions at the molecular level, it would be useful to determine the structure of all the proteins in a cell or organism. At present, however, this has not been achieved. Interestingly, the number of different protein variants arising from protein synthesis (alternative splicing and/or post-translational modifications) is significantly greater than the number of genes in a genome.

Mass spectrometry (an analytical technique for the determination of molecular mass) in combination with chromatographic or electrophoretic separation techniques, is currently the method of choice for identifying endogenous proteins in cells, characterizing post-translational modifications and determining protein abundance (Zhu *et al.*, 2003). Two-dimensional gel electrophoresis is unique with respect to the large number of proteins (>10 000) that can be separated and visualized in a single experiment. Protein spots are cut from the gel, followed by proteolytic digestion, and proteins are then identified using mass spectrometry (Aebersold and Mann, 2003). However, standardization and automation of two-dimensional gel electrophoresis has proved difficult, and the use of the resulting protein patterns as proteomic reference maps has only been successful in a few cases. A complementary technique, liquid chromatography, is easier to automate, and it can be directly coupled to mass spectrometry. Affinity-based proteomic methods that are based on microarrays are an alternative approach to protein profiling (Lueking *et al.*, 2003), and can also be used to detect protein–protein interactions. Such information is essential for algorithmic modelling of biological pathways. However, binding specificity remains a problem in the application of protein microarrays, because cross-reactivity cannot accurately be predicted. Alternative approaches exist for detecting protein–protein interactions such as the two hybrid system (Fields and Song, 1989). However, none of the currently used methods allow the quantitative detection of binding proteins, and it remains unclear to what extent the observed interactions are likely to represent the physiological protein–protein interactions.

Array-based methods have also been developed for detecting DNA–protein interaction *in vitro* and *in vivo* (see Sauer *et al.*, 2005, for a review), and identifying unknown proteins binding to gene regulatory sequences. DNA microarrays are employed effectively for screening nuclear extracts for DNA-binding complexes, whereas

protein microarrays are mainly used for identifying unknown DNA-binding proteins at proteome-wide level. In the future, these two techniques will reveal detailed insights into transcriptional regulatory networks.

Many methods of predicting the function of a protein are based on its homology to other proteins and its location inside the cell. Predictions of protein functions are rather complicated, and also require techniques to detect protein–protein interactions, and to detect the binding of proteins to other molecules, because proteins fulfil their functions in these binding processes.

4 The role of bioinformatics

Developing high-throughput technologies would be useless without the capacity to analyse the exponentially growing amount of biological data. These need to be stored in electronic databases (Box 78) associated with specific software designed to permit data update, interrogation and retrieval. Information must be easily accessible and interrogation-flexible, to allow the retrieval of information, that can be analysed to unravel metabolic pathways and the role of the proteins and genes involved.

Bioinformatics is crucial to combine information from different sources and generate new knowledge from existing data. It also has the potential to simulate the structure, function and dynamics of molecular systems, and is therefore helpful in formulating hypotheses and driving experimental work.

5 Conclusions

Molecular characterization can play a role in uncovering the history, and estimating the diversity, distinctiveness and population structure of AnGR. It can also serve as an aid in the genetic management of small populations, to avoid excessive inbreeding. A number of investigations

Box 78
Databases of biological molecules

A number of databases exist which collect information on biological molecules:

DNA sequence databases:

- European Molecular Biology Lab (EMBL): http://www.ebi.ac.uk/embl/index.html
- GenBank: http://www.ncbi.nlm.nih.gov/
- DNA Data Bank of Japan (DDBJ): http://www.ddbj.nig.ac.jp

Protein databases:

- SWISS-PROT: http://www.expasy.ch/sprot/sprot-top.html
- Protein Information Resource (PIR): http://pir.georgetown.edu/pirwww/
- Protein Data Bank (PDB): http://www.rcsb.org/pdb/

Gene identification utility sites Bio-Portal

- GenomeWeb: http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-geneid.html
- BCM Search Launcher: http://searchlauncher.bcm.tmc.edu/
- MOLBIOL: http://www.molbiol.net/
- Pedro's BioMolecular Research tools: http://www.biophys.uni-duesseldorf.de/BioNet/Pedro/research_tools.html
- ExPASy Molecular Biology Server: http://www.expasy.ch/

Databases of particular interest for domestic animals:

http://locus.jouy.inra.fr/cgi-bin/bovmap/intro.pl
http://www.cgd.csiro.au/cgd.html
http://www.ri.bbsrc.ac.uk/cgi-bin/arkdb/browsers/
http://www.marc.usda.gov/genome/genome.html
http://www.ncbi.nlm.nih.gov/genome/guide/pig/
http://www.ensembl.org/index.html
http://www.tigr.org/
http://omia.angis.org.au/
http://www.livestockgenomics.csiro.au/ibiss/
http://www.thearkdb.org/
http://www.hgsc.bcm.tmc.edu/projects/bovine/

have described within and between-population diversity – some at quite a large scale. However, these studies are fragmented and difficult to compare and integrate. Moreover, a comprehensive worldwide survey of relevant species has not been carried out. As such, it is of strategic importance to develop methods for combining existing, partially overlapping datasets, and to ensure the provision of standard samples and markers for future use as worldwide references. A network of facilities collecting samples of autochthonous germplasm, to be made available to the scientific community under appropriate regulation, would facilitate the implementation of a global survey.

Marker technologies are evolving, and it is likely that microsatellites will increasingly be complemented by SNPs. These markers hold great promise because of their large numbers in the genome, and their suitability for automation in production and scoring. However, the efficiency of SNPs for the investigation of diversity in animal species remains to be thoroughly explored. The subject should be approached with sufficient critical detachment to avoid the production of biased results.

Methods of data analysis are also evolving. New methods allow the study of diversity without *a priori* assumptions regarding the structure of the populations under investigation; the exploration of diversity to identify adaptive genes (e.g. using population genomics, see Box 77); and the integration of information from different sources, including socio-economic and environmental parameters, for setting conservation priorities (see Section F). The adoption of a correct sampling strategy and the systematic collection of phenotypic and environmental data, remain key requirements for exploiting the full potential of new technologies and approaches.

In addition to neutral variation, research is actively seeking genes that influence key traits. Disease resistance, production efficiency, and product quality are among the traits having high priority. A number of strategies and new high-throughput –omics technologies are used to this end. The identification of QTN offers new opportunities and challenges for AnGR management. Information on adaptive diversity complements that on phenotypic and neutral genetic diversity, and can be integrated into AnGR management and conservation decision-tools. The identification of unique alleles or combinations of alleles for adaptive traits in specific populations may reinforce the justification for their conservation and targeted utilization. Gene assisted selection also has the potential to decrease the selection efficiency gap currently existing between large populations raised in industrial production systems, and small local populations, where population genetic evaluation systems and breeding schemes cannot be effectively applied. Marker and gene assisted selection may not, however, always represent the best solution. These options need to be evaluated and optimized on a case-by-case basis, taking into account short and long-term effects on population structure and rates of inbreeding, and cost and benefits in environmental and socio-economic terms – in particular impacts on people's livelihoods.

As in the case of other advanced technologies, it is highly desirable that benefits of scientific advances in the field of molecular characterization are shared across the globe, thereby contributing to an improved understanding, utilization and conservation of the world's AnGR for the good of present and future human generations.

PART **4**

Box 79
Glossary: molecular markers

For the purpose of this section the following definitions are used:

Candidate gene: any gene that could plausibly cause differences in the observable characteristics of an animal (e.g. in disease resistance, milk protein production or growth). The gene may be a candidate because it is located in a particular chromosome region suspected of being involved in the control of the trait, or its protein product may suggest that it could be involved in controlling the trait (e.g. milk protein genes in milk protein production).

DNA: the genetic information in a genome is encoded in deoxyribonucleic acid (DNA), which is stored in the nucleus of a cell. DNA has two strands structured in a double helix, which is made of a sugar (deoxiribose), phosphate, and four chemical bases – the nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T). An A on one strand always pairs with a T on the other through two hydrogen bonds, while a C always pairs with a G through three hydrogen bonds. The two strands are, therefore, complementary to each other.

Complementary DNA (cDNA): DNA sequences generated from the reverse transcription of mRNA sequences. This type of DNA includes exons and untranslated regions at the 5′ and 3′ ends of genes, but does not include intron DNA.

Genetic marker: a DNA polymorphism that can be easily detected by molecular or phenotypic analysis. The marker can be within a gene or in 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.

Haplotype: a contraction of the phrase "haploid genotype", is the genetic constitution of an individual chromosome. In the case of diploid organisms, the haplotype will contain one member of the pair of alleles for each site. It may refer to a set of markers (e.g. single nucleotide polymorphisms – SNPs) found to be statistically associated on a single chromosome. With this knowledge, it is thought that the identification of a few alleles of a haplotype block can unambiguously identify all other polymorphic sites in this region. Such information is very valuable for investigating the genetics behind complex traits.

Linkage: The association of genes and/or markers that lie near each other on a chromosome. Linked genes and markers tend to be inherited together.

Linkage disequilibrium (LD): is a term used in the study of population genetics for the non-random association of alleles at two or more loci, not necessarily on the same chromosome. It is not the same as linkage, which describes the association of two or more loci on a chromosome with limited recombination between them. LD describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies. Linkage disequilibrium is caused by fitness interactions between genes or by such non-adaptive processes as population structure, inbreeding, and stochastic effects. In population genetics, linkage disequilibrium is said to characterize the haplotype distribution at two or more loci.

Microarray technology: a new way of studying how large numbers of genes interact with each other and how a cell's regulatory networks control vast batteries of genes simultaneously. The method uses a robot to precisely apply tiny droplets containing functional DNA to glass slides. Researchers then attach fluorescent labels to mRNA or cDNA from the cell they are studying. The labelled probes are allowed to bind to cDNA strands on the slides. The slides are put into a scanning microscope that can measure the brightness of each fluorescent dot; brightness reveals how much of a specific mRNA is present, an indicator of how active it is.

Primer: a short (single strand) oligonucleotide sequence used in a polymerase chain reaction (PCR)

RNA: Ribonucleic acid is a single stranded nucleic acid consisting of three of the four bases present in DNA (A, C and G). T is, however, replaced by uracil (U).

References

Aebersold, R. & Mann, M. 2003. Mass spectrometry-based proteomics. *Nature*, 422 (6928): 198–207. Review.

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

Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Research,* 12(12): 1805–14.

Aravin, A. & Tuschl, T. 2005. Identification and characterization of small RNAs involved in RNA silencing. *Febs Letters,* 579(26): 5830–40.

Bachem, C.W.B., Van der Hoeven, R.S., De Bruijn, S.M., Vreugdenhil, D., Zabeau, M. & Visser, R.G.F. 1996. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analyses of gene expression during potato tuber development. *The Plant Journal*, 9: 745–753.

Bamshad, M. & Wooding, S.P. 2003. Signatures of natural selection in the human genome. *Nature Reviews Genetics,* 4(2): 99–111. Review.

Baumung, R., Simianer, H. & Hoffmann, I. 2004. Genetic diversity studies in farm animals – a survey, *Journal of Animal Breeding and Genetics,* 121: 361–373.

Beaumont, M.A. & Balding, D.J. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13(4): 969–80.

Beja-Pereira, A., Alexandrino, P., Bessa, I., Carretero, Y., Dunner, S., Ferrand, N., Jordana, J., Laloe, D., Moazami-Goudarzi, K., Sanchez, A. & Cañon, J. 2003. Genetic characterization of southwestern European bovine breeds: a historical and biogeographical reassessment with a set of 16 microsatellites. *Journal of Heredity,* 94: 243–50.

Berthier, D., Quere, R., Thevenon, S., Belemsaga, D., Piquemal, D., Marti, J. & Maillard, J.C. 2003. Serial analysis of gene expression (SAGE) in bovine trypanotolerance: preliminary results. *Genetics Selection Evolution*, 35 (Suppl. 1): S35–47.

Bertone, P, Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. & Snyder, M. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306: 2242–2246.

Black, W.C., Baer, C.F., Antolin, M.F. & DuTeau, N.M. 2001. Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, 46: 441–469.

Bruford, M.W., Bradley, D.G. & Luikart, G. 2003. DNA markers reveal the complexity of livestock domestication. *Nature Reviews Genetics*, 4: 900–910.

Buntjer, J.B., Otsen, M., Nijman, I.J., Kuiper, M.T. & Lenstra, J.A. 2002. Phylogeny of bovine species based on AFLP fingerprinting. *Heredity*, 88: 46–51.

Campbell, D. & Bernatchez, L. 2004. Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology and Evolution*, 21(5): 945–56.

Cañon, J., Garcıa, D., Garcıa-Atance, M.A., Obexer-Ruff, G., Lenstra, J.A., Ajmone-Marsan, P., Dunner, S. & The ECONOGENE Consortium. 2006. Geographical partitioning of goat diversity in Europe and the Middle East. *Animal Genetics*, 37: 327–334.

Chen, S.Y., Su, Y.H., Wu, S.F., Sha, T. & Zhang, Y.P. 2005. Mitochondrial diversity and phylogeographic structure of Chinese domestic goats. *Molecular Phylogenetics and Evolution*, 37: 804–814.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15: 1496–1502.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B., Bouix, J., Caiment, F., Elsen, J.M., Eychenne, F., Larzul, C., Laville, E., Meish, F., Milenkovic, D., Tobin, J., Charlier, C. & Georges, M. 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genetics,* 38: 813–818.

De Marchi, M., Dalvit, C., Targhetta, C. & Cassandro, M. 2006. Assessing genetic diversity in indigenous Veneto chicken breeds using AFLP markers. *Animal Genetics*, 37: 101–105.

Donson, J., Fang, Y., Espiritu-Santo, G., Xing, W., Salazar, A., Miyamoto, S., Armendarez, V. & Volkmuth, W. 2002. Comprehensive gene expression analysis by transcript profiling. *Plant Molecular Biology*, 48: 75–97.

Excoffier, L., Smouse, P.E. & Quattro, J.M. 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131: 479–491.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M., Moisio, S., Simon, P., Wagenaar, D., Vilkki, J. & Georges, M. 2002. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics*, 161: 275–287.

Fields, S. & Song, O. 1989. A novel genetic system to detect protein–protein interactions. *Nature*, 340: 245–246.

Freeman, A.R., Bradley, D.G., Nagda, S., Gibson, J.P. & Hanotte, O. 2006. Combination of multiple microsatellite data sets to investigate genetic diversity and admixture of domestic cattle. *Animal Genetics*, 37: 1–9.

Goldstein, D.B., Linares, A.R., Cavalli-Sforza, L.L. & Feldman, M.W. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139: 463–471.

Goldstein, D.B. & Schlötterer, C. 1999. *Microsatellites: evolution and applications.* New York. Oxford University Press.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M. & Snell, R. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research*, 12: 222–231.

Guo, J., Du, L.X., Ma, Y.H., Guan, W.J., Li, H.B., Zhao, Q.J., Li, X. & Rao, S.Q. 2005. A novel maternal lineage revealed in sheep (*Ovis aries*). *Animal Genetics*, 36: 331–336.

Haley, C. & de Koning, D.J. 2006. Genetical genomics in livestock: potentials and pitfalls. *Animal Genetics*, 37(Suppl 1): 10–12.

Hanotte, O., Bradley, D.G., Ochieng, J.W., Verjee, Y. & Hill, E.W. 2002. African pastoralism: genetic imprints of origins and migrations. *Science*, 296: 336–339.

Hayes, B.J., Visscher, P.M., McPartlan, H.C. & Goddard, M.E. 2003. A novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, 13: 635–643.

Hill, E.W., O'Gorman, G.M., Agaba, M., Gibson, J.P., Hanotte, O., Kemp, S.J., Naessens, J., Coussens, P.M. & MacHugh, D.E. 2005. Understanding bovine trypanosomiasis and trypanotolerance: the promise of functional genomics. *Veterinary Immunology and Immunopathology*, 105: 247–258.

Hill, W.G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, 38: 209–216.

Hillel, J., Groenen, M.A., Tixier-Boichard, M., Korol, A.B., David, L., Kirzhner, V.M., Burke, T., Barre-Dirie, A., Crooijmans, R.P., Elo, K., Feldman, M.W., Freidlin, P.J., Maki-Tanila, A., Oortwijn, M., Thomson, P., Vignal, A., Wimmers, K. & Weigend, S. 2003. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. *Genetics Selection Evolution*, 35: 533–557.

Hood, L., Heath, J.R., Phelps, M.E. & Lin, B. 2004. Systems biology and new technologies enable predictive and preventative medicine. *Science,* 306: 640–643.

Ibeagha-Awemu, E.M., Jann, O.C., Weimann, C. & Erhardt, G. 2004. Genetic diversity, introgression and relationships among West/Central African cattle breeds. *Genetics Selection Evolution*, 36: 673–690.

Jarne, P. & Lagoda, P.J.L. 1996. Microsatellites, from molecules to populations and back. *Tree*, 11: 424–429.

Joshi, M.B., Rout, P.K., Mandal, A.K., Tyler-Smith, C., Singh, L. & Thangaraj, K. 2004. Phylogeography and origin of Indian domestic goats. *Molecular Biology and Evolution*, 21: 454–462.

Joost, S. 2006. *The geographical dimension of genetic diversity.* A GIScience contribution for the conservation of animal genetic resources. École Polytechnique Fédérale de Lausanne, Switzerland. (PhD thesis)

Kayser, M., Brauer, S. & Stoneking, M. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*, 20: 893–900.

Lai, S.J., Liu, Y.P., Liu, Y.X., Li, X.W. & Yao, Y.G. 2006. Genetic diversity and origin of Chinese cattle revealed by mtDNA D-loop sequence variation. *Molecular Phylogenetics and Evolution*, 38: 146–54.

Lan, L., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Ton-Keen Mui, E., Flowers, M.T., Schueler, K.L., Manly, K.F., Williams, R.W., Kendziorski, C. & Attie, A.D. 2006. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2: 51–61.

Liang, P. & Pardee, A.B. 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257: 967–997.

Liu, Y.P., Wu, G.S., Yao, Y.G., Miao, Y.W., Luikart, G., Baig, M., Beja-Pereira, A., Ding, Z.L., Palanichamy, M.G. & Zhang, Y.P. 2006. Multiple maternal origins of chickens: out of the Asian jungles. *Molecular Phylogenetics and Evolution*, 38: 12–19.

Lueking, A., Possling, A., Huber, O., Beveridge, A., Horn, M., Eickhoff, H., Schuchardt, J., Lehrach, H. & Cahill, D.J. 2003. A nonredundant human protein chip for antibody screening and serum profiling. *Molecular and Cellular Proteomics*, 2: 1342–1349.

Luikart, G., England, P.R., Tallmon, D., Jordan, S. & Taberlet, P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, 4: 981–994.

Luikart, G., Gielly, L., Excoffier, L., Vigne, J.D., Bouvet, J. & Taberlet, P. 2001. Multiple maternal origins and weak phylogeographic structure in domestic goats. *Proceedings of the National Academy of Science USA,* 98: 5927–5932.

Mburu, D.N., Ochieng, J.W., Kuria, S.G., Jianlin, H. & Kaufmann, B. 2003. Genetic diversity and relationships of indigenous Kenyan camel (Camelus dromedarius) populations: implications for their classification. *Animal Genetics*, 34(1): 26–32.

McPherron, A.C. & Lee, S.J. 1997. Double muscling in cattle due to mutations in the myostatin gene. *Proceedings of the National Academy of Science USA,* 94: 12457–12461.

Negrini, R., Milanesi, E., Bozzi, R., Pellecchia, M. & Ajmone-Marsan, P. 2006. Tuscany autochthonous cattle breeds: an original genetic resource investigated by AFLP markers. *Journal of Animal Breeding and Genetics*, 123: 10–16.

Nei, M. 1972. Genetic distance between populations. *The American Naturalist*, 106: 283–292.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics,* 89: 583–590.

Nei, M. & Roychoudhury, A.K. 1974. Sampling variances of heterozygosity and genetic distance. *Genetics*, 76: 379–390.

Nei, M., Tajima, F. & Tateno, Y. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *Journal of Molecular Evolution*, 19: 153–170.

Nielsen, R. & Signorovitch, J. 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology,* 63: 245–55.

Nijman, I.J., Otsen, M., Verkaar, E.L., de Ruijter, C. & Hanekamp, E. 2003. Hybridization of banteng (*Bos javanicus*) and zebu (*Bos indicus*) revealed by mitochondrial DNA, satellite DNA, AFLP and microsatellites. *Heredity*, 90: 10–16.

Pariset, L., Cappuccio, I., Joost, S., D'Andrea, M.S., Marletta, D., Ajmone Marsan, P., Valentini A. & ECONOGENE Consortium 2006. Characterization of single nucleotide polymorphisms in sheep and their variation as an evidence of selection. *Animal Genetics*, 37: 290–292.

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

Rabie, T.S., Crooijmans, R.P., Bovenhuis, H., Vereijken, A.L., Veenendaal, T., van der Poel, J.J., Van Arendonk, J.A., Pakdel, A. & Groenen, M.A. 2005. Genetic mapping of quantitative trait loci affecting susceptibility in chicken to develop pulmonary hypertension syndrome. *Animal Genetics*, 36: 468–476.

Saitou, N. & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4: 406–425.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S. & Altshuler, D.; International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409: 928–933.

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

SanCristobal, M., Chevalet, C., Peleman, J., Heuven, H., Brugmans, B., van Schriek, M., Joosten, R., Rattink, A.P., Harlizius, B., Groenen, M.A., Amigues, Y., Boscher, M.Y., Russell, G., Law, A., Davoli, R., Russo, V., Desautes, C., Alderson, L., Fimland, E., Bagga, M., Delgado, J.V., Vega-Pla, J.L., Martinez, A.M., Ramos, M., Glodek, P., Meyer, J.N., Gandini, G., Matassino, D., Siggens, K., Laval, G., Archibald, A., Milan, D., Hammond, K., Cardellino, R., Haley, C. & Plastow, G. 2006b. Genetic diversity in European pigs utilizing amplified fragment length polymorphism markers. *Animal Genetics*, 37: 232–238.

Sauer, S., Lange, B.M.H., Gobom, J., Nyarsik, L., Seitz, H. & Lehrach, H. 2005. Miniaturization in functional genomics and proteomics. *Nature Reviews Genetics*, 6: 465–476.

Sodhi, M., Mukesh, M., Mishra, B.P., Mitkari, K.R., Prakash, B. & Ahlawat, S.P. 2005. Evaluation of genetic differentiation in *Bos indicus* cattle breeds from Marathwada region of India using microsatellite polymorphism. *Animal Biotechnology*, 16: 127–137.

Storz, G., Altuvia, S. & Wassarman, K.M. 2005. An abundance of RNA regulators. *Annual Review of Biochemistry*, 74: 199–217.

Sunnucks, P. 2001. Efficient genetic markers for population biology. *Tree*, 15: 199–203.

Syvänen, A.C. 2001. Accessing genetic variation genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, 2: 930–941.

Takezaki, N. & Nei, M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, 144: 389–399.

Tapio, M., Tapio, I., Grislis, Z., Holm, L.E., Jeppsson, S., Kantanen, J., Miceikiene, I., Olsaker, I., Viinalass, H. & Eythorsdottir, E. 2005. Native breeds demonstrate high contributions to the molecular variation in northern European sheep. *Molecular Ecology*, 14: 3951–3963.

Tilquin, P., Barrow, P.A., Marly, J., Pitel, F., Plisson-Petit, F., Velge, P., Vignal, A., Baret, P.V., Bumstead, N. & Beaumont, C. 2005. A genome scan for quantitative trait loci affecting the *Salmonella* carrier-state in the chicken. *Genetics Selection Evolution*, 37: 539–61.

Troy, C.S., MacHugh, D., Bailey, J.F., Magee, D.A., Loftus, R.T., Cunningham, P., Chamberlain, A.T., Sykesk, B.C. & Bradley D.G. 2001. Genetic evidence for Near-Eastern origins of European cattle. *Nature*, 410: 1088–1091.

Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. 2000. Analyzing uncharted transcriptomes with SAGE. *Trends in Genetics*, 16: 423–425.

Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. 1995. Serial analysis of gene expression. *Science*, 270: 484–487.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J. & Kuiper, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, 23: 4407–1444.

Weir, B.S. & Basten, C.J. 1990. Sampling strategies for distances between DNA sequences. *Biometrics*, 46: 551–582.

Weir, B.S. & Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution*, 38: 1358–1370.

Wienholds, E. & Plasterk, R.H. 2005. MicroRNA function in animal development. *FEBS Letters*, 579: 5911–5922.

Wong, G.K., Liu, B., Wang, J., Zhang, Y., Yang, X., Zhang, Z., Meng, Q., Zhou, J., Li, D., Zhang, J., Ni, P., Li, S., Ran, L., Li, H., Zhang, J., Li, R., Li, S., Zheng, H., Lin, W., Li, G., Wang, X., Zhao, W., Li, J., Ye, C., Dai, M., Ruan, J., Zhou, Y., Li, Y., He, X., Zhang, Y., Wang, J., Huang, X., Tong, W., Chen, J., Ye, J., Chen, C., Wei, N., Li, G., Dong, L., Lan, F., Sun, Y., Zhang, Z., Yang, Z., Yu, Y., Huang, Y., He, D., Xi, Y., Wei, D., Qi, Q., Li, W., Shi, J., Wang, M., Xie, F., Wang, J., Zhang, X., Wang, P., Zhao, Y., Li, N., Yang, N., Dong, W., Hu, S., Zeng, C., Zheng, W., Hao, B., Hillier, L.W., Yang, S.P., Warren, W.C., Wilson, R.K., Brandstrom, M., Ellegren, H., Crooijmans, R.P., van der Poel, J.J., Bovenhuis, H., Groenen, M.A., Ovcharenko, I., Gordon, L., Stubbs, L., Lucas, S., Glavina, T., Aerts, A., Kaiser, P., Rothwell, L., Young, J.R., Rogers, S., Walker, B.A., van Hateren, A., Kaufman, J., Bumstead, N., Lamont, S.J., Zhou, H., Hocking, P.M., Morrice, D., de Koning, D.J., Law, A., Bartley, N., Burt, D.W., Hunt, H., Cheng, H.H., Gunnarsson, U., Wahlberg, P., Andersson, L., Kindlund, E., Tammi, M.T., Andersson, B., Webber, C., Ponting, C.P., Overton, I.M., Boardman, P.E., Tang, H., Hubbard, S.J., Wilson, S.A., Yu, J., Wang, J., Yang, H.; International Chicken Polymorphism Map Consortium. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, 432: 717–722.

Zhu, H., Bilgin, M. & Snyder, M. 2003. Proteomics. *Annual Review of Biochemistry*, 72: 783–812.