Molecular tools for exploring genetic diversity

1 Introduction

Recent advances in the field of genomic technology have constituted a major innovation in livestock production. The increasing availability of molecular tools is deeply affecting the ways in which livestock species are studied and managed. This section provides an overview of recent developments related to molecular tools and their use, focusing particularly on the period since the first report on The State of the World’s Animal Genetic Resources for Food and Agriculture (first SoW-AnGR) (FAO, 2007) was prepared.

The first SoW-AnGR noted that the main roles of molecular technologies in the characterization of AnGR include:

- assessing functional and neutral genetic variability within and between populations, including investigation of their history (domestication, expansion or reduction of the population size, migrations, introgression episodes, etc.);
- assessing the current state of a population in terms of risks related to inbreeding and genetic drift, using estimators such as effective population size; and
- genetic characterization of traits (e.g., physical appearance, productivity, disease resistance and other adaptability traits) specific to given populations.

The report highlighted the following three ongoing developments in molecular biology as being particularly relevant to AnGR management:

- the establishment of whole genome sequences for various livestock species;
- the development of technologies for measuring polymorphisms at loci spread across the entire genome; and
- the development of technologies for measuring gene transcription and expression on a large scale.

Since the first SoW-AnGR was prepared, the list of species whose genomes have been sequenced has continued to grow. It now includes chicken (2004), sheep (2010), cattle (2009), horse (2009), pig (2012), rabbit (2009), turkey (2009) and goat (2013). The costs of genotyping and sequencing have declined sharply during this period (Figure 4B1). High-density SNP arrays, allowing the simultaneous assay of several tens of thousands to several hundreds of thousands of SNPs, are available for use in livestock species at a cost of US$100 or less if a relatively large number of individuals are sequenced. Genomes

**FIGURE 4B1**
Change in cost per genome sequenced in humans

$100,000,000$ $10,000,000$ $1,000,000$ $100,000$ $10,000$ $1,000$
2001 2003 2005 2007 2009 2011 2013

Note: Costs expressed on a logarithmic scale.
**From DNA to phenotype**

The genome of livestock species is organized in pairs of chromosomes, each inherited from one of the parents and chemically made up of large molecules of DNA (deoxyribonucleic acid). Each gene in an individual, therefore, has two copies, known as alleles, one on each chromosome of a pair (with the exception of non-paired sex chromosomes). Chromosomes comprise genes and intergenic regions. The former encode proteins and other products. The latter, which represent the majority of the genome, are believed to play various regulatory roles (The ENCODE Project Consortium, 2012). Genes typically consist of coding sequences (exons) separated by introns and accompanied by regulatory elements. Like intergenic regions, introns carry no protein-coding information, but sometimes play a role in the regulation of gene expression.

Nucleic acids are strands of smaller molecules called nucleotides. DNA has four types of nucleotide (adenine, cytosine, guanine and thymine). The instruction encoded in a gene depends on the sequence in which these nucleotides are ordered. This sequence information is put into action in two steps. The first is the 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 molecule. In the second step, these molecules (transcripts) are edited and eventually translated into proteins (particular three-nucleotide sequences correspond to particular amino acids, the molecules that constitute a protein). Gene expression is highly tissue-specific and time-dependent. Not all genes are translated into proteins; some express their function as non-coding RNA molecules that play important roles in protein synthesis (transfer RNA and ribosomal RNA) and various regulatory processes (microRNA and long non-coding RNA, two types of regulatory RNA that differ in terms of the number of nucleotides they consist of – approximately 20 and more than 200 nucleotides, respectively).

Phenotypic differences between individuals, populations and species are a consequence both of environmental effects (including epigenetic mechanisms – see Box 485) and of variations in DNA sequences. These variations may be caused by point mutations leading to the substitution of single nucleotides (single nucleotide polymorphisms – SNPs), insertions, deletions, duplications, copy number variations or inversions of DNA fragments. If SNPs are in exons, different alleles may lead either to the same amino acid (synonymous SNPs) or to a different amino acid (non-synonymous SNPs) being included in the resulting protein. DNA variations can be classified as “functional” or “neutral”.

In the case of functional variation, changes in the sequence of nucleotides in the DNA molecule induce changes in the phenotypic function of the organism. In the case of neutral variation, a change in the DNA sequence does not give rise to any change in function.

Because genes or SNPs that lie near each other on a chromosome (i.e. are physically “linked” to each other) tend to be inherited together, a neutral variant can be associated with a functional one. In addition, this interdependence between linked genes means that the various combinations of their alleles are not distributed randomly (a phenomenon termed “linkage disequilibrium”). Linkage and linkage disequilibrium allow the use of one site of polymorphism as a “genetic marker” for polymorphism in a nearby region.

---

can be sequenced for less than US$3 000 each with moderate coverage (e.g. “eight-fold” coverage – meaning that, on average, each position in the genome is sequenced eight times). Sequencing smaller fractions of genomes (restriction site associated DNA sequencing – RAD-Seq) can be used directly in the characterization of individual animals (this is termed “genotyping by sequencing”) (De Donato et al., 2013). Similarly, the development of tools capable of assaying a high density of transcripts and even direct transcriptome sequencing (also known as “RNA-seq” – short for RNA sequencing), has increased capacity to study gene expression and hence to unravel...
the complex physiological regulation of target traits (D’Alessandro and Zola, 2012).

2 Developments in the use of DNA markers

Progress in sequencing techniques and the opportunities offered by the development of high-density marker arrays have considerably improved the availability of DNA information over the last ten years, both in terms of the number of markers identified and in terms of the cost of genotyping.

Until recently, microsatellites remained one of the most popular types of marker in genetic characterization studies (Lenstra et al., 2012), used for example in projects such as “GlobalDiv”, which ran from 2007 to 2010 and combined microsatellite datasets from various diversity studies from different parts of the world (Ajmone-Marsan et al., 2010). Microsatellite data continue to be used, especially in developing countries (e.g. Abdullah et al., 2012; Azam et al., 2012) and in the context of conservation and priority setting at regional level (e.g. Medugorac et al., 2011; Ginja et al., 2013). However, they are increasingly being superseded by the use of SNP marker arrays. With the advent of next-generation sequencing, mitogenomics (analysis of the whole mitochondrial genome rather than a limited fragment of mitochondrial DNA) can be routinely used in livestock species, including less intensively studied species such as goats (Doro et al., 2014) and horses (Achilli et al., 2012). The recent generation of whole genome reference sequences for many livestock species has allowed “population genetics” to become “population genomics”. Population genomics uses large sets of SNPs to study specific variations across the genome and determine how they have been shaped by the history (e.g. changes in population size, selection, and cross-breeding) of livestock populations. SNPs can be assigned to various classes (neutral vs. genic, intron vs. exon or synonymous vs. non-synonymous), which provides opportunities for more detailed analysis of diversity. The past decade has also witnessed a revolution in sequencing technologies that has led to the development of various platforms for DNA and RNA sequencing, known collectively as next-generation sequencing technologies (see Metzker, 2010 and Davey et al., 2011 for reviews). These tools can rapidly (in a few days or weeks) provide sequence data in the form of short reads (sequenced DNA fragments between 100 and 400 base pairs long on average) that collectively cover the whole genome of a sample (or the transcriptome of a particular organ) several times. Identifying SNPs from this type of data is relatively easy, provided that a reference sequence has been established (Nielsen et al., 2011), which is the case for most livestock species. Methods have also been developed for SNP discovery in newly sequenced species (Norman et al., 2013) and these approaches may prove useful for less common livestock species.

High-density SNP panels are now widely used for genome-wide association studies (GWAS), genomic prediction and population genomic analyses. However, the preliminary phase, i.e. SNP discovery or SNP selection from databases, is critical. If data have not been obtained randomly, standard estimators of population genetic parameters should be applied with caution. Non-random selection may occur if SNP sets are derived for use on a given set of breeds but later used on other breeds or if SNP sets are filtered to meet certain criteria (e.g. a minimum allele frequency).

Many current tools are affected by both these factors, as they have been developed primarily using widely used international transboundary breeds and with the use of SNP-filtering criteria. Such protocols bias the distribution of allelic frequencies relative to what would be expected in a random sample. The resulting inaccuracy in estimation of genetic parameters is known as “ascertainment bias”. Bias caused by problems of this kind is probably present in most commercial and ready-to-use medium- and high-density SNP panels currently available for use in livestock species. Unbiased estimates of the absolute genetic diversity (i.e. the nucleotide diversity) of a population can, in theory, be
Box 4B2

Glossary: genetic markers

Genetic marker: a DNA sequence variation that is informative with respect to a specific location (locus) on a particular chromosome.

Microsatellites: segments of DNA characterized by a variable number of copies (typically 5 to 50) of sequence motifs of around two to five bases (referred to as a repeat unit). At any one locus (site in the genome), there are usually several different "alleles" in a population, each allele identifiable by the number of repeat units detected via polymerase chain reaction (PCR). Many microsatellites have a large degree of polymorphism. In many species, they were the first standard marker technology used to characterize diversity. However, due to their comparatively infrequent presence across the genome, inconsistent reproducibility across laboratories and genotyping platforms, and higher genotyping cost per locus, microsatellites are being replaced by other technologies.

Single-nucleotide polymorphism (SNP): a DNA sequence variation that results from a change in the nucleotide at a single location in the genome. SNPs usually have only two alleles. They may represent either neutral or functional genetic diversity and generally occur throughout the genome. In most species, SNPs occur, on average, once in every 100 to 300 positions in the DNA sequence. For most major livestock species, commercial arrays are available that allow substantial numbers of SNPs (from a few hundreds to over a million) to be genotyped in a single reaction at a low cost per marker. SNP arrays are now routinely used as more informative alternatives to microsatellite panels in genetic diversity studies.

Markers of sex-specific inheritance: certain parts of the genome have sex-specific inheritance. Mitochondrial DNA is passed from the mother to the offspring. The Y-chromosome in mammals is inherited from father to son, while the W-chromosome in birds is inherited from mother to daughter. This class of markers can include both SNPs and other sequence variations and has been instrumental in identifying wild ancestors, localizing domestication centres and reconstructing colonization and trading routes.

Sequence variation: with the emergence of whole-genome sequencing, the entire variation present in the DNA sequence is now available as a potential source of marker information. This variation comprises SNPs and insertions and deletions (InDels) (loss or gain of one or more nucleotides relative to the species reference sequence), duplications, copy number variations (CNVs) (variation in the number of copies of sections of the DNA) or inversions of DNA fragments. Sequencing can be performed on a whole genome basis or only for specific parts of the genome (e.g. the exome or genome fractions used for genotyping by sequencing). Whole-genome sequences will be the ultimate source of genetic diversity information, as they harbour the motifs responsible for genetic differences between breeds. However, efficient management, use and storage of this information will require, in addition to sequencing capacities, substantial development of resources in bioinformatics (methods, tools and hardware).

obtained only via whole genome sequencing. Statistical approaches that explicitly account for the methods used in SNP discovery and sample preparation have been developed for use when undertaking various kinds of population genetics analyses with SNPs (Nielsen et al., 2011; Kofler et al., 2011). Large-scale projects have also started to harvest genome-wide information for use in characterizing livestock populations at national or international scale, including studies on cattle (Gauthier et al., 2010), sheep (Kjes et al., 2012), horses (McCue et al. 2012; Orlando et al., 2013), pigs (Groenen et al., 2010), chickens (Weigend et al., 2015) and goats (Dong et al., 2013).

It is important to note that although cost per SNP is low relative to microsatellites (and decreases with the number of SNPs analysed) the costs of high-density assays – currently (2015)
US$50 to US$200 and depending heavily on the number of arrays purchased – are nonetheless prohibitive for many applications. Costs continue to decline, however, and financially realistic options are likely to eventually become available for most situations. This being said, even if lower cost genotyping assays become available, the bioinformatic infrastructure in most developing countries will still require further development. Both the sheer amount of raw data and the complexity of analytical models are several orders of magnitude larger than those associated with microsatellite-based analyses. This is true for work with SNP array data, but even more so for work with sequence data.

Further studies are in the process of identifying millions of SNPs and haplotypes (specific allelic combinations for a given set of loci) and also other sequence variants such as insertion-deletion polymorphisms (Indels) and copy number variants (CNV) (see Box 4B2 for explanations of these terms). Novel sequencing technologies are continuously evolving, accompanied by a drop in cost per sequenced genome (see Figure 4B1). Allele frequency differences and diversity measures derived from them can be obtained inexpensively by sequencing pooled DNA from multiple individuals from a population (e.g. Qanbari et al., 2012). Sooner or later, sequence-based approaches will become the standard methodology for generating data for use in livestock diversity studies.

Marker information will become even more useful when linked to biological background information available in specialized databases. Information about marked genes and their functions is available in the Ensembl database1 (among others) for many livestock species. Information on quantitative trait loci (QTL) is collected in the AnimalQTL database2 and genomic pathway information is available through KEGG.3 In human genetics, the Encode project4 is systematically annotating functional elements in the genome, and similar initiatives are emerging in other species, including livestock (Andersson et al., 2015). On this basis it can, for example, be judged whether observed between-breed diversity in a given genomic region is purely neutral and has been generated by genetic drift or is of functional relevance and may have been caused by selection. Making systematic use of such information will allow a shift from a purely statistical assessment of genetic diversity to a more informative functional approach.

3 Characterization of within-population diversity

Classical estimators of genetic variability (heterozygosity, F-statistics, etc.) are still commonly used. However, some are not adapted for use with biallelic markers (e.g. number of alleles per locus, which is invariably equal to two for biallelic markers). As the use of SNPs has become more common, so has the use of individual and population genomic indicators of diversity and similarity, such as coancestry and inbreeding (Meuwissen and Goddard, 2001; Keller et al., 2011; Saura et al., 2013; Curik et al., 2014). Some of these indicators make it possible to test whether inbreeding effects are more or less important in specific genomic locations, or whether inbreeding comes from a more or less distant common ancestor (e.g. Ferenčaković et al., 2013). Estimators of genetic variability can also be used in conservation decision-making as a means of optimizing the choice of breeding animals so as to minimize the loss of genetic variability (Oldenbroek, 2007).

In parallel, several methods of estimating present and past effective population sizes have been developed or improved, based either on the correlation between allele frequencies (linkage disequilibrium) or on runs of homozygosity (Sved, 1971; Hill, 1981; Hayes et al., 2003; Waples, 2006; Li and Durbin, 2011; Hillestad et al., 2014). These approaches have been increasingly applied in livestock, including cattle (de Roos et al., 2008;
Flury et al., 2010), sheep (Klias et al., 2012), pigs (Umari and Tapio, 2011), chickens (Qanbari et al., 2010) and horses (Corbin et al., 2010). It should be noted, however, that the widely used approach suggested by Sved (1971) has some methodological shortcomings (Sved, 2008) and is especially sensitive to non-random samples of SNPs (Corbin et al., 2012; Ober et al., 2013).

At the time the first SoW-AnGR was prepared, it was generally considered that because of the limited number of markers available it was more efficient to use genealogical information than molecular information in conservation decision-making (Fernandez et al., 2005). This appears no longer to be the case. Commercial SNP arrays are now affordable and provide estimates of genetic relationships that account for the random segregation and recombination of chromosomes that occur during inheritance from parents to offspring. Because marker-based information provides better estimates of genetic relationships than pedigree data, inclusion of genomic data is likely to increase the efficiency of conservation schemes (Hasler et al., 2011; Toro et al., 2014).

### 4 Characterization of between-population diversity

Relationships between populations have long been assessed through the estimation of genetic distances, which are often used to construct phylogenetic trees to visually infer genetic relationships. However, a major drawback of reconstructing phylogenetic trees is that the evolution of lineages is assumed to be non-reticulate, i.e. it is assumed that while lineages may diverge, they never result from crosses between lineages. There is therefore a tendency for these methods to be replaced by alternative graphical networks or other approaches such as Bayesian clustering methods or multivariate analysis (Bertorelle et al., 2004). One of the most popular model-based Bayesian approaches in current use is the model-based clustering method developed by Pritchard et al. (2000) (STRUCTURE software), although alternatives are available (e.g. Alexander et al., 2009). The approach uses Monte Carlo Markov chain simulation to assign individuals to a chosen number of clusters (populations), inferring genetic origins without a priori knowledge or assumptions. This is a particularly important consideration in livestock populations, where unsuspected admixture may have occurred. The approach is, however, not without limits. For example, inferred clusters may not always be ancestral, but rather related to highly inbred populations (“inbreeding bias”) or to populations over-represented in the dataset (“sampling bias”) (Lenstra et al., 2012). Multivariate analysis approaches are interesting alternatives to model-based approaches, as they are generally assumption-free methods and are specifically designed for summarizing large and complex datasets into a small number of synthetic variables (Jombart et al., 2010). These various approaches are usually applied to microsatellite or SNP marker information. They have been extensively used in livestock studies, either independently or (because of the complementary information they may provide) in parallel (Muchadeyi et al., 2007; Leroy et al., 2009; Gautier et al., 2010; Klias et al., 2012). Methods have been developed over the last few years that use dense haplotype data to unravel fine-scale population structure (Lawson et al., 2012) or apply advanced admixture analysis in order to infer the presence and historical timing of admixture events among human populations (Patterson et al., 2012; Pickrell and Pritchard, 2012; Hellenthal et al., 2014).

Recently, a growing number of methods for combining genomic information with information from other sources, often related to the environments where animals are raised, have been developed (Pariset et al., 2012). Landscape genomics is an approach that aims to use various methods (e.g. estimation of molecular distance, Bayesian and multivariate analyses) to identify environmental factors that shape genetic variability. For example, a study on Vietnamese goats showed that social organization and husbandry practices were as important as geographical distance in shaping genetic structure (Berthouly et al., 2009). The increasing density of markers genotyped may also allow these approaches to be used to identify chromosomal
regions and genes likely to be subject to positive selection linked to the environment. Finally, knowledge of the history of livestock populations has greatly increased in the last ten years, based on the development of new methods and the increasing availability of large sets of markers (see Part 1 Section A). For instance, a recent study on horse breeds (Wallner et al., 2013) showed that the diversity of the paternally inherited Y-chromosome was very low in comparison to that of maternally inherited mitochondrial DNA haplotypes, a finding consistent with the disproportionate use of a limited number of popular stallions over recent centuries. Genome-wide panels of markers also make it possible to discriminate areas of the genome whose variation has increased or decreased though history in relation to specific gene function.

5 Molecular tools for targeting functional variation

Recent advances in genomics have clearly improved our capacity to characterize functional variation in livestock species. Detection and mapping of QTLs, i.e. markers physically linked to a genomic variant that underlies variation in a quantitative trait, have benefited from increased genome coverage, as well as from the development of new methodologies. In particular, the use of sequence data may allow causative polymorphism to be targeted directly instead of via QTLs. The molecular background of various breed-specific traits has been the subject of numerous investigations (Table 4B1) (see also Box 4B3). The molecular analysis of adaptive variation has also improved knowledge of the possible adverse effects of selection on the health and productivity of animals. For example:

- Several gene variants are pathogenic or confer sterility in homozygous animals. For instance, myostatin deficiency caused by mutations in the MSTN (myostatin) gene hinders the delivery of calves (Belling et al., 2005).
- Gene variants exhibiting clear antagonism between milk yield and fertility (increasing the former while decreasing the latter) have been identified in cattle (Kadri et al., 2014).
- The use of only a few top sires promotes inbreeding and thus increased homozygosity. This effect, which inevitably increases the proportion of offspring that have recessive genetic defects, can be assessed using neutral genetic markers (Lenstra et al., 2012). Several pathogenic mutations in livestock species, most of which are recessive, have been identified. They surpass in number the gene variants known to be involved in economic traits (Nicholas et al., 2012).
- Significant deficiency or complete absence of individuals homozygous for a given haplotype may indicate the presence of a recessive genetic defect causing early embryonic mortality. This concept has been successfully used in the identification of possible causes of reduced fertility in various cattle populations (Fritz et al., 2013).

International consortia have provided large amounts of data on SNPs and other variants. For example, the “1 000 bull genomes project” (Daetwyler et al., 2014) identified 28.3 million variants, related, inter alia, to coat colour, embryonic loss and production traits. However, it is still difficult to obtain genome sequences for a large number of animals at an affordable price. Methods have therefore been developed that can be used to “impute” or infer the genotypes of individual animals for which information is sparse (e.g. obtained using low- or medium-density SNP chips) from information on a subset of individuals that have been sequenced (e.g. those studied in the above-mentioned 1 000 bull genomes project) (Jansen et al., 2013).

Although the study of animal genetic diversity has typically concentrated on direct differences in genomes, the impact of genetic diversity on the expression of genes may be relevant, especially as interest grows in functional genetic diversity relative to neutral genetic diversity. Since the mid-1990s, the widespread use of DNA microarrays and serial analysis of gene expression (SAGE), both of which provide a snapshot of actively expressed
Box 483
How genetic tools helped to solve the mystery of the origin of the Booroola gene

Exceptionally high litter size in an Australian Merino flock kept at the Booroola Estate in Cooma, New South Wales, attracted the attention of scientists from the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO). Initial analysis of ewes’ pedigrees and performance records led to a hypothesis regarding the segregation of a major gene affecting this trait and increasing litter size in carriers. This hypothesis was first substantiated by analysis of litter-size segregation in families (Piper and Bindon, 1982). The origin of the high-fecundity gene in the low-prolific Merino population, however, remained a mystery until Professor Helen Newton Turner found evidence that ancestors of the Booroola flock could have had some admixture of Indian Bengal sheep brought to Australia from Calcutta in the 1790s (Turner, 1983). The hypothetical major gene increasing litter size was named Fecility Booroola (FecB).

The first genetic markers linked to the FecB locus were discovered by a New Zealand team led by Professor Grant Montgomery (Montgomery, 1993). Further research led to the conclusion that the Booroola gene is located on the sixth chromosome. The first molecular test, devised to enable the introgression of the FecB mutation into the Romney breed, was based on the polymorphism of three microsatellite sequences (Lord et al., 1998).

The real breakthrough with respect to the physiological basis for increased fecundity happened in 2001, when teams from AgResearch (New Zealand), INRA (France) and Edinburgh University (United Kingdom) independently discovered that carriers of the Booroola gene have a mutation in the bone morphogenetic protein receptor IB gene (BMPR-IB). The Booroola gene (FecB) is a dominant autosomal gene with an additive effect on ovulation rate.

Garole sheep of Bengal

Photo credit: Kanhaiya M. Chavan.

The discovery of the mutation and the development of the molecular test enabled the identification of the mutation in the Garole sheep of Bengal, a breed that is well known for its large litter sizes—thus supporting Professor Turner’s theory. At present, the BMPR-IB mutation has been found in a number of breeds that have high fecundity. The list includes Javanese Thin Tail sheep (Davis, 2009) and some Chinese breeds such as the Huyang, Small Tail Han (STH), Cele, Duolang and Chinese Merino (Hua and Yang, 2009). It seems that the original mutation took place in Mongolian Fat Tail sheep and was introgressed into Chinese breeds and later into the Indian Garole and Javanese breeds as a result of the movement of people and animals along the Silk Road.

So the mystery was solved thanks to the persistence of scientists and development of technology. Over time, the FecB mutation has been introgressed into about 40 breeds, all around the world (Walkden-Brown et al., 2008).

Provided by Elbieta Martynuk.

genes and transcripts in a biological sample, has facilitated high-throughput molecular studies of the transcriptome. Microarray experiments provide a cost-effective means of studying the transcriptome, and the bioinformatic and statistical analyses (referred to as “analysis pipelines”)
<table>
<thead>
<tr>
<th>Species/breed(s)</th>
<th>Phenotype</th>
<th>Gene or locus</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cattle</td>
<td>Dairy traits</td>
<td>Several candidate genes</td>
<td>Flori et al., 2009</td>
</tr>
<tr>
<td></td>
<td>Dairy and beef traits</td>
<td>Several candidate genes</td>
<td>Rothammer et al., 2013</td>
</tr>
<tr>
<td>Danish Red</td>
<td>High milk yield, low fertility</td>
<td>Deletion removing RNASE1G2&lt;sup&gt;*&lt;/sup&gt;</td>
<td>Kadri et al., 2014</td>
</tr>
<tr>
<td>Several</td>
<td>Milk protein content</td>
<td>ABCG2&lt;sup&gt;*&lt;/sup&gt;</td>
<td>Braunschweig, 2010</td>
</tr>
<tr>
<td>Several</td>
<td>Muscular hypertrophy</td>
<td>MSTN&lt;sup&gt;+&lt;/sup&gt; (different mutations)</td>
<td>Nicholas and Hobbs, 2012, O’Rourke et al., 2013</td>
</tr>
<tr>
<td>Holstein</td>
<td>Stature</td>
<td>PLAST-CYCHD7 Intergenic</td>
<td>Karin et al., 2011</td>
</tr>
<tr>
<td>Dexter</td>
<td>Short stature</td>
<td>AGA4&lt;sup&gt;*&lt;/sup&gt;</td>
<td>Cavanagh et al., 2007</td>
</tr>
<tr>
<td>Dutch Belted</td>
<td>Belted pattern</td>
<td>HEST (candidate gene)</td>
<td>Drogemüller et al., 2010</td>
</tr>
<tr>
<td>Swiss Brown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheep</td>
<td>Litter size</td>
<td>GDF5&lt;sup&gt;*&lt;/sup&gt; (FecG, different mutations)</td>
<td>Vasse et al., 2013</td>
</tr>
<tr>
<td>Several</td>
<td>Litter size</td>
<td>BMP15&lt;sup&gt;*&lt;/sup&gt; (FecX, different mutations)</td>
<td>Nicholas and Hobbs, 2012</td>
</tr>
<tr>
<td>Several</td>
<td>Litter size</td>
<td>BMPRTB (Bororoda, FecB)</td>
<td>Davis et al., 2006</td>
</tr>
<tr>
<td>Locaune</td>
<td>Litter size</td>
<td>BAGALNT2 (FecQ)</td>
<td>Drouillet et al., 2013</td>
</tr>
<tr>
<td>Texel and others</td>
<td>Muscular hypertrophy</td>
<td>MSTN&lt;sup&gt;+&lt;/sup&gt; (FecR)</td>
<td>Clopp et al., 2006</td>
</tr>
<tr>
<td>Donet</td>
<td>Muscular hypertrophy</td>
<td>CLPG&lt;sup&gt;*&lt;/sup&gt;</td>
<td>Braunschweig, 2010</td>
</tr>
<tr>
<td>Pig</td>
<td>Muscle growth</td>
<td>IGF2</td>
<td>Braunschweig, 2010</td>
</tr>
<tr>
<td>Horse</td>
<td>Type 1 muscle fibres</td>
<td>MSTN</td>
<td>Peterson et al., 2013</td>
</tr>
<tr>
<td></td>
<td>Endurance</td>
<td>GYS1&lt;sup&gt;*&lt;/sup&gt;</td>
<td>McCoy et al., 2014</td>
</tr>
<tr>
<td>Chicken</td>
<td>Naked neck</td>
<td>BMP12</td>
<td>Mou et al., 2011</td>
</tr>
<tr>
<td>Several</td>
<td>Frizzle feather</td>
<td>KRTPS</td>
<td>Ng et al., 2012</td>
</tr>
<tr>
<td>Several</td>
<td>Silky feather</td>
<td>PSSS2</td>
<td>Feng et al., 2014</td>
</tr>
<tr>
<td>Several</td>
<td>Comb shape</td>
<td>MNR2</td>
<td>Imland et al., 2012</td>
</tr>
</tbody>
</table>

Note: Several mutations may already have played a role in more general adaptation to domestication (see Part 1, Section A, Table 1A2). * causative gene variant is pathogenic or confers sterility if homozygous; † recent gene mutation; ‡ ATP-binding cassette, sub-family G (WHITE), member 2.
Sources: Braunschweig, 2010; Nicholas and Hobbs, 2012.

used to transform raw microarray data into interpretable results are now well established (Ritchie et al., 2015). Since the time the first SoW-AnGR was prepared, the development of high-throughput sequencing in the field of transcriptomic analysis (RNA sequencing or RNA-seq) has led to radical changes (Mortazavi et al., 2008), primarily because RNA-seq approaches do not necessarily require prior knowledge of a genome sequence or annotation (identification of locations and functions of coding regions within a genome) and can therefore be used even in poorly characterized organisms. In addition, it enables a wide range of novel applications, including detection
of weakly expressed genes and alternative splicing isoforms (variations in the proteins translated from the same gene) (Wang et al., 2008; Pan et al., 2008), variable assembly of transcripts (Trapnell et al., 2010; Guttman et al., 2010; Robertson et al., 2010; Grabherr et al., 2011; Schulz et al., 2012) and allele-specific expression (Skelly et al., 2011).

Recent comparisons have indicated good overall agreement among results obtained using microarrays, quantitative PCR (polymerase chain reaction) and RNA-seq across different sequencing platforms (Zhao et al., 2014; Trapnell et al., 2013; Nookaew et al., 2012; Liu et al., 2011). However, although microarrays and RNA-seq are both used to characterize transcriptional activity, the experimental, bioinformatic and analytical steps associated with the two differ considerably (Oshlack et al., 2010). In particular, RNA-seq experiments generate much more data than alternative transcriptomic approaches and require more sophisticated analyses and therefore greater technical capacity in bioinformatics and biostatistics (e.g. Langmead and Salzberg, 2012; Grabherr et al., 2011; Oshlack and Wakefield, 2009; Zhou et al., 2014). The analytical processes of transcriptomics constitute a major area of research in bioinformatics and statistics.

In recent years, studies using RNA-seq to examine genetic variation in gene expression have been undertaken in cattle (Li et al., 2011), chickens (Endale Ahanda et al., 2014; Davis et al., 2015) and pigs (Corominas et al., 2013; Fischer et al., 2015). The objectives of these studies have included the identification of candidate genes influencing phenotypic differences and the study of differences in gene expression associated with specific SNPs.

The role of bioinformatics

The successful use of high-throughput technologies in the study of genetic diversity is largely contingent on the availability of support and expertise in bioinformatics and statistics. Increasingly large and complex datasets need to be understood, organized, quantified, and analysed. Developing and applying the methods and software tools needed to do this requires appropriate computing resources (including sufficient computational power and memory to store and manipulate large data files) and programming skills. For example, genome sequencing and RNA-seq studies often require the services of a dedicated bioinformatics team to pre-process the data, including raw-data quality control and sequence alignment or assembly, in addition to biostatisticians for eventual data analysis. Bioinformatic support is often also an integral part of the development, maintenance and interrogation of biological databases.

An increasing number of well-documented and open-access bioinformatics and statistical tools are available online. For example, the Bioconductor project⁵ is an open-source open-development software project that develops and provides widespread access to a diverse set of well-documented statistical and graphical analysis tools (written in the R programming language) for high-throughput genomic data. In addition, an increasing number of free and publicly accessible resources (e.g. the Galaxy project,⁶ an open web-based platform) are available to facilitate bioinformatic analyses without the need for extensive programming knowledge.

It is highly desirable that when researchers gather large-scale genomic data for a given project they make them freely available to other researchers once the initial analyses have been completed. Increasingly, scientific journals and research-funding organizations request that data underlying publications or generated in publicly funded projects be deposited in open repositories. This kind of open-source policy will generate a large quantity and variety of reference data, across species and breeds, that can be used for increasingly comprehensive and informative diversity studies.

---

⁵ http://bioconductor.org
⁶ http://galaxyproject.org
Conclusions and research priorities

The world of genetics has been revolutionized over the last decade with the advent of massive parallel sequencing and high-throughput genotyping technologies. Other technologies and opportunities are on the way (see Box 484). These developments have opened many opportunities to utilize molecular techniques in the management of AnGR. However, while these technologies facilitate the sequencing of complete genomes or the genotyping of high-density SNP panels at moderate cost, they have not completely replaced traditional molecular markers such as microsatellites, mainly because of their still relatively high costs and the additional skills needed to analyze the enormous amount of data they produce. Low-cost alternatives, such as low-density SNP panels, that allow genetic variants scattered across the genome to be queried and can feasibly be used in small and medium-sized laboratories are in development, but remain to be implemented in practice.

Understanding of genetic diversity needs to be improved, even in the most widely used livestock species. For example, comprehensive assessments of genetic diversity using molecular genetic markers need to be extended to locally adapted breeds, particularly those with small population sizes. The value of the large quantities of data that currently exist in fragmented form needs to be maximized (e.g. by undertaking meta-analyses and by making as much data as possible publicly available for use by breeders, researchers and policy-makers). Improvements in sequencing and genotyping technologies have already provided standards that can be used as references for further genotyping and sequencing studies. Reference genomes, biological background information and population genotypic data are still not available for some species, but sequencing efforts currently underway in laboratories around the world will soon fill these gaps. For most populations and production systems, taking full advantage of the opportunities that advances in genomics have created for the study of genetic diversity will also require new and additional phenotypic data.

Understanding of the genetic basis of adaptive traits also needs to be improved. Potential approaches include the use of new technologies, such as genome sequencing and geographic

Box 484
What are the promises of the post-genomic era?

Over the last twenty years, the use of molecular tools has acquired paramount importance in animal breeding through the development of genetic tests, as well as the implementation of genomic selection in a growing number of species. The role of molecular tools is expected to continue expanding. Potential developments include:

- increased use of whole-genome sequencing for genomic selection, identification of new functional variants (allowing selection on new traits) and analysis of genetic diversity;
- the use of epigenetics (see Box 485) in the study of environment x genome interactions to provide insight into complex traits, especially those related to development;
- the use of meta-genomic studies that consider the gut microbiome to enable the optimization of the rumen microbial ecosystem for better feed-conversion efficiency; and
- combining genomics with other advanced biotechnologies, such as in vitro embryo transfer (selecting breeding candidates at the embryo stage) and genetic engineering (introducing genes of interest into the genome or even directly editing the genome through novel technologies such as the CRISPR/Cas system), which may bring about major changes in the way animals are raised and selected.

Some of these developments would, clearly, raise social and ethical concerns that would need to be addressed before putting them into practice.

Source: Gonzalez-Recio, 2012; Hayes et al., 2013.
The reality and promises of epigenetics for animal production

The term "epigenetics" refers to mechanisms that cause variation in gene activity and are based not on variation in the nucleotide sequence but on chemical modifications of DNA and chromatin, for instance, affecting the degree of chromatin compaction or the accessibility of regulatory sequences to transcription factors.

Epigenetic processes occur during cell, tissue and organismal differentiation processes such as gametogenesis, embryo genome activation, X-chromosome inactivation and genomic imprinting (the differential expression of certain genes depending on which parent they come from). Epigenetic modifications driven by environmental factors, such as diet, stress and disease, form a molecular memory (transient or permanent) that adapts the organism and programmes signalling and metabolic pathways appropriate for particular conditions. Epigenetic marks are epistemically and mitotically inherited and the epigenome represents an additional level in genotype-phenotype mapping.

As epigenetic variation explains part of the so-called missing heritability (i.e. the part of genetic variation not explained by individual genes), there may be interest in using epigenetics in livestock breeding, for instance through genomic imprinting and metabolic programming (stimulation of early development to influence later physiological outcome). Particularly in cross-breeding schemes, imprinting offers new opportunities to establish dam and sire lines with enhanced production performances and maternal abilities and to produce cross-breeds according to market requirements. The interaction between genetic (bred) and environmental factors such as diets and management practices is key to the use of epigenetics in animal breeding. Extensive knowledge of epigenetic mechanisms and gene regulation will, in future, offer new opportunities in livestock breeding in relation to environment-genome-epigenome interactions.

Provided by Klaus Wimmers.

References


