Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Status and gaps in characterization of animal genetic resources

M. Tixier-Boichard
INRA, AgroParisTech, UMR GABI, Jouy-en-Josas, France

ABSTRACT: Characterization provides data on present and potential uses of animal genetic resources. FAO manages DAD-IS but the degree of completeness is below 50%. Besides the impressive progress in animal genomics, characterization must consider also phenotypes, production systems and all services provided by livestock to humans. Three gaps have been identified: (1) characterization of functional diversity, (2) data sharing, and (3) access to genetic resources. Standardizing phenotypic descriptors and landscape genomic analysis are promising approaches to fill the knowledge gap on functional diversity. Genetic resources must be taken on board in whole-genome studies. A single database cannot include all relevant information worldwide, but a web portal could be developed to provide a catalogue of datasets, described with metadata, on the model of the Global Biodiversity Information Facility (GBIF). Characterization is an iterative process that should aim to be more prospective than retrospective in order to support multiple objectives, including dynamic cryobanking.

Keywords: landscape genomics; data sharing; environment

Introduction

Genetic resources of livestock species cover the whole range of domestic populations used in different production systems, within and across countries. Production systems have been and are still changing, and so are genetic resources. Thus, genetic resources are both a heritage and a resource for future. Wild relatives should also be considered as a resource for some species such as pig, chicken and most fishes. In this frame, why do we need characterization ? Characterization is expected to provide data on present and potential uses of animal genetic resources, which will help to make decisions regarding their conservation and management. Typically, conservation policies ask for recommendations of what to preserve. From a theoretical viewpoint, the general recommendation is to preserve the capacity of populations to remain genetically viable with sufficient diversity to maintain the ability to adapt to new conditions. However, specialized genotypes are currently dominating the market, and those populations that do not perform well nowadays are at the risk to be lost. But what would we lose exactly ? How can we characterize the potential usefulness of a population? A major challenge for characterization is to anticipate our future needs. It is now well recognized that major environmental and social changes will take place by 2050, with the growth of the human population, the need to adapt to climate change, and/or to attenuate it by acting on production systems (Hoffmann, 2010). Impressive progress has been made in recent years in animal genomics, but characterization must consider also animal phenotypes and performance, production systems, natural habitats and the range of services that livestock populations are providing to humans. This paper will briefly review the current status of the characterization of genetic resources and underline the recent trends before addressing the gaps and make recommendations for an integrated and prospective approach of characterization.

Current status

An international framework set up by FAO. Characterization of genetic resources is organized at the country level and national coordinators are responsible for providing data to the global information system managed by FAO, Domestic Animal Diversity Information System (DAD-IS ; http://dad.fao.org). FAO develops and updates guidelines regarding characterization and management of genetic resources, but data quality remains the responsibility of the data provider. In 2007, the first State of the World (SoW) project was a major endeavor to collect data on breeds and make them available worldwide (http://www.fao.org/docrep/010/a1250e/a1250e00.htm). The SoW includes all steps from inventory and characterization to monitoring and conservation policies. The second SoW is underway.

FAO statistics are based upon the breed concept for which a broadened definition has been proposed : « *either a sub-specific group of domestic livestock with definable and identifiable external characteristics that enable it to be separated by visual appraisal from other similarly defined groups within the same species or a group for which geographical and/or cultural separation from phenotypically similar groups has led to acceptance of its separate identity* » (FAO, 1999). Actually, genetic resources of a livestock species include an array of populations: traditional populations, standardized breeds, selected breeds and derived lines (Tixier-Boichard et al., 2008). These populations differ by their selection history and, generally, by their morphological characteristics. Knowing better the selection status of a population would help to assess its potential usefulness for different production systems, and would avoid irrelevant comparisons.

Regarding molecular characterization, FAO has published sets of markers and guidelines for sampling and genotyping of livestock species but DAD-IS does not cover genotyping data. Many studies have been realized with molecular markers, which did not always follow FAO recommendations, and were generally not connected with each other for a given species. Only a few genotyping studies have been conducted at a worldwide scale, for instance in chickens (Granevitze et al., 2009). Regarding phenotypic characterization, FAO has identified two main approaches: an exploratory one, where the objective is to investigate the existence of a distinct breed in a given area, and a confirmatory one, where the objective is to validate breed identity and provide systematic description of the breeds.

To support characterization, DAD-IS is provides templates for describing a breed with historical data, qualitative information (specific features) morphological features, mean values of performance traits, management conditions and population parameters (total size, number of breeding animals). In spite of the efforts to update DAD-IS and promote a network organization through regional focal points, the degree of coverage of the different fields is very variable : the degree of data completeness calculated as the ratio [filled fields/all possible fields] ranges from 34% to 50% across continents. The ratio varies even more between countries, and geographic scale should be considered to calculate a weighted completeness. Population reports provide average performance, but little or no information is given on performance variability. The most complete records deal with demographic parameters. Thus, DAD-IS provides a global view on existing breeds but its use by research and by breeders is likely to be very limited since the information remains very general.

At the country level. Data collection on livestock populations is organized according to practices and priorities of breeders, with a support from national policy which greatly varies across countries. An important issue when designing an information system (IS) is to know the users of this system in order to meet their needs. In practice, characterization of genetic resources and access to these data is set up differently according to the current use of a population:

- breeds selected for intensive production systems as well as experimental lines are deeply phenotyped, their genealogy is known and a lot of genomic information is being gathered ; the International Consortium for Animal Recording (http://www.icar.org) provides standards for phenotyping production traits, but this is done only for cattle, sheep and goats ; the access to the national databases for commercial populations is generally restricted to the breeders who provided data and to those who are going to use them for genetic evaluation ; furthermore, for commercial lines, such as poultry, data on performance traits are kept entirely confidential ;
- experimental lines are also deeply phenotyped, with specific traits being defined according to the scientific objective, data are described in scientific publications but are not generally made available as raw data, although IS are usually set up by research intitutions ;
- local populations, either recognized with a breed name (most often the case in Europe , Asia) or not (few breed names in Africa) are not systematically described ; generally single-population studies are published, recorded traits are a subset of the traits defined for selected populations or experimental line and are not so much relevant for the systems in which local breeds are embedded. A simple survey of the Web of Science database shows that the number of publications regarding local animal breeds has increased from 310 between 2002-2006 to 496 between 2010-2014, but a very small proportion of them includes genetic data (15/310 and 21/496, respectively). Although this is a very rough estimation, it may suggest that results on local breeds are not always published in peer-reviewed journals.

FAO has already identified the information gap on local breeds and the Global Plan of Action includes a program to support data collection for such populations in countries lacking a national organization. A strong recommendation is to describe the production environments at the same time as the phenotypes and the genotypes of breeds.

Recent trends

Genomics and landscape genomics. Lenstra et al (2012) have reviewed the different types of molecular markers currently available for diversity studies, and the information that can be inferred from them. Our knowledge of within-breed diversity, relationships between breeds, and breeds' history has very much progressed thanks to the extensive use of molecular markers since 10 years. Mitochondrial DNA is still widely used to analyze the domestication process by gathering samples from a wide geographic scale (as in Naderi et al, 2008, for goats, or Miao et al., 2013 for chickens). Microsatellites have been widely used because of their high level of polymorphism and large-scale studies have been published, some taking advantage from European projects (SanCristobal et al., 2006 for pigs ; Granevitze et al., 2009, for chickens), but many analysis remained limited to some breeds or some countries, which calls for the need of doing meta-analysis. Meta-analysis of microsatellite data have been very difficult to conduct because of technical differences between laboratories which complicate the merging of genotypes. Presson et al (2006) developed a Bayesian method for merging human microsatellite data which could be useful for animal populations (MicroMerge software). Even without this approach, many studies reached the general conclusion that truly indigenous populations still existed in developing countries, with a high level of genetic variability but a minimal phenotypic description.

Most livestock species now have a whole-genome sequence assembled and made available through international repositories. This technological and numeric revolution has changed the way of doing selection and has also impacted the characterization of genetic resources. The first consequence has been the production of millions of SNP (single nucleotide polymorphism) markers. SNP are becoming the markers of choice because of their abundance and standardized recording. Marker sets ranging from 30k to 60K, and more rarely above 100K, are currently available in several species. Such a high density of markers improves the accuracy of genetic analysis, not only to reconstruct population history but also to search for adaptive diversity. Thus, studies aimed at the characterization of genetic diversity in a large set of breeds (as in Gautier et al., 2010, for French cattle breeds) are now converging with studies aimed at the detection of selection signatures (as in Wilkinson et al., 2013, for pigs, or in Vaysse et al., 2011, for dogs). However, the choice of SNPs is not neutral: typically, most SNP chips are designed for selected populations and do not represent a random sample of available SNPs within a domestic species. This represents a selection bias that may limit hypothesis testing in population genetics for local breeds which have not undergone the same history as the selected populations for which the SNP chip was designed. The importance of this bias on the assessment of within and between breeds variability has, however, not been quantified.

The main challenge is thus to give sense and extract functional information from the current flow of genomic data. At the Interlaken Conference where the Global Plan of Action for animal genetic resources was adopted, Sere et al (2008) underlined the strong potential of landscape genomics to understand *« the match between livestock breeds, populations and genes, and the physical, biological and economic landscape.* ». Overlaying environmental information with genetic information in a study of village chickens in West-Africa revealed a genetic similarity between local populations found in regions sharing high levels of precipitation, from Cameroon to Côte d'Ivoire (Leroy et al., 2012). In the subset of data for Benin, Ghana and Côte d'Ivoire, three genetically differentiated areas were identified, matching with Major Farming Systems (namely Tree Crop, Cereal-Root Crop, and Root Crop) described by the FAO. Although the genetic data were still not dense, interesting trends could be observed, suggesting a match between genotypes and climatic or farming conditions. In the case of local breeds where selection pressure is mainly due to environmental conditions, scanning for SNP differentiation with a higher density of markers made possible to identify genomic regions under selection, as shown by Gautier et al (2009) in nine West African cattle populations studied with 36K SNPs. The scan identified regions harboring candidate genes related to immune response, nervous system, skin and coat, which suggested a relationship with climatic conditions and response to parasites such as *Trypanosoma*.

Initially developed for wild populations, landscape genetics/genomics (depending on the nature and amount of genetic information used) aims to understand the processes of gene flow and local adaptation by studying how geographical and environmental variables structure the genetic variation within and between populations. In the case of domestic populations, the environmental variables will also include the farming systems and will be affected by the gene flow more or less controlled by breeders. Multivariate methods make possible to handle heterogeneous data and are very well suited to landscape genetic analysis, such as spatial multivariate analysis and redundancy analysis, which can be used when the number of environmental variables is lower than the number of genetic variables (i.e. 10 environmental variables for 60K genotypes). Pariset et al. (2009) compared spatial analysis to Fst analysis in order to detect outlier loci among 27 SNPs, chosen in candidate genes likely to be associated to selection in European goats. Combining the two methods made it possible to identify three genes that were under selection pressure driven by the environment (for instance relative humidity in summer). Laloë et al. (2010) used spatial principal component (sPCA) analysis and spatial metric multidimensional scaling (sMDS) to analyze microsatellite data for European cattle (101 breeds) sheep (46 breeds) and goats (45 breeds). The goats appeared to be the most geographically structured species, followed by cattle. For all three species, the main genetic cline was from southeast to northwest. The observed structure results both from environmental effects and breeding practices and from drift. For example, a relatively strong north-south contrast was observed for sheep breeds that may reveal a climatic effect, but may also be due to the difference between English and Merino-types of breeds, which have been used frequently for crossbreeding (Laloë et al., 2010).

Finally, the most powerful way to use the genome sequence without any marker bias is to resequence the whole genome of a subset of individuals chosen to represent well-defined breeds. An efficient approach is to pool the samples from each breed, resequence the DNA pools and analyze the patterns of variation (SNPs, structural variants) identified after mapping on the reference genome. Two whole-genome studies have the explored domestication history of livestock species at an unprecedented resolution in chickens (Rubin et al., 2010) and in pigs (Groenen et al., 2012; Rubin et al., 2012) with some spectacular results such as the TSHR mutation found to be diagnostic of domestication in chickens.

Standards for environmental descriptors. Landscape genetics is expected to identify genomic regions controlling functional traits, provided that environmental data are connected to relevant information for breeders as well as for the whole society. Thus, describing the economic, social and environmental context in which the breeds are used is highly needed to fully benefit from landscape genomics approach.

To address this requirement, FAO organized several meetings and working groups in order to develop a standard set of production environment descriptors (PEDs) for use in DAD-IS and in phenotypic characterization studies (FAO, 2012). It must be noted that some descriptors, particularly the climatic ones, correspond to standards used in other recording systems, such as weather stations, so that available information may be connected as much as possible with the breed performance data. To facilitate such connections, DAD-IS incorporates global maps including climatic data, but also elevation, land cover, soil features. However, management practices are not easy to map, and cultural requirements are even more difficult to describe, so that these data often have to be collected through specific surveys set up for characterization, simultaneously to biological sampling for genome studies.

The scientific community has also identified the need to standardize description of phenotypes and environmental conditions. As many databases exist around institutions, as well as around countries, different vocabularies are generally used. These databases could be connected if a ‘dictionary’ was available to describe the meaning, provide a correspondence and connections between different vocabularies. In order to set up such a dictionary, concepts underlying phenotypes and production systems need to be formalized, as well as relationships between these concepts, this corresponds to the definition of an ontology. Gene ontology is a great challenge now gathering a wide research community, and transferring this approach to phenotype ontology is also a challenge. Model species have developed specific tools to search for phenotypes as efficiently as searching for genes, as can be seen for mouse (http://www.informatics.jax.org) and for rat (http://rgd.mcw.edu). Regarding livestock, recorded phenotypes are generally specific of a breed type or a production system (i.e. the relevant age to measure body weight is quite different for broiler or a layer) which limits large-scale comparisons. The possibility to define some common “anchor” traits to record should be investigated. A working group has been set up at INRA to develop an ontology for animal traits (ATOL) and for environmental conditions (EOL) which can be downloaded from http://www.atol-ontology.com/index.php/en/. This initiative started in 2009 and produced a documented web site in 2012. It already has some international partners in USA and is open to interested colleagues. Obviously, such a project needs regularly updating and a lot of consultation and dialogue with researchers. It has not been developed to be applied to the characterization of genetic resources, and is likely to operate at a much finer scale than what is currently achieved in breed surveys. It is, however, a strong indication that the need for standardizing descriptors is making its way among scientists in agronomic research, who also have connections with national structures in charge of the characterization of genetic resources.

Filling the gaps

Characterization of genetic resources would be most useful if it could be more prospective than retrospective. The data currently available on genetic resources are retrospective and cover a relatively short period of time, it is generally restricted to simple statistics, they cannot really be used to make any prediction of future usefulness of a population. This is particularly true for local breeds, which are generally considered as being adapted to a harsh environment, but the mechanisms underlying this adaptation are poorly known, and sometimes confounded with low production level. Local breeds are endangered because they do not fit well the current economic context, although they may well be the resource for important changes that livestock breeding will face by 2050. Which characterization do we need to be more prospective and how can we use it ? Three main gaps may be identified at the level of (1) characterization of functional diversity (2) data access (3) access to genetic resources.

Functional diversity The major knowledge gap involves adaptive traits and sustainability, for selected as well as unselected populations. Several approaches can be explored to fill this gap:

(i) setting the characterization of genetic resources in the conceptual framework of ecosystem services which takes into account production, but also environmental impacts (positive and negative, local and distant) as well as cultural values; in this framework, trade-offs between services are of utmost importance, as could be trade-offs between productive and adaptive traits in livestock; although the concept of ecosystem services is still evolving and open to debate (Lele et al., 2013), it is starting to be applied to value ecosystem services in complex ecosystems incorporating livestock, as shown by Silvestri et al (2013) in Kenya ; thus, the need to quantify various services and their interactions is opening a new research field which is relevant for the characterization of genetic resources in various production systems; the frame of ecosystem services is strongly connected with sustainability, as shown by Broom et al. (2013) who evaluated sylvopastoral systems but this frame should not be limited to grazing and extensive systems such as pastoralism and should involve all production systems and the genetic resources embedded in them ;

(ii) combining whole-genome data with deep phenotyping, this requires high-throughput recording which is costly and likely to be set up only for selected populations where production traits and production objectives are rather well defined; however, results are likely to identify genes or regulatory pathways that will be relevant for the advanced characterization of other animal populations ;

(iii) applying landscape genomics to the joint analysis of whole-genome data and environmental variation, this requires reliable and standardised environmental descriptors and can be applied for selected or unselected populations as long as geographical information is available at the individual level ; its major advantage is that it does not need any prior hypothesis on the mechanisms involved ;

(iv) identifying biomarkers, i.e. fast-and-easy measurements (genotypes or metabolites) correlated with phenotypes of interest in the frame of pilot projects incorporating a range of populations representative of the genetic diversity of a species; such biomarkers would be added in the “minimal set” of traits recommended for characterization. Biomarkers of choice should be as non-invasive as possible, and make possible longitudinal studies to be monitored at different ages or different steps of the production cycle; examples would be blood metabolites, DNA markers, or image analysis for morphological or biophysical parameters such as body temperature.

Ideally, in the future, one could consider that whole-genome association studies across breeds and countries will have identified reliable associations between markers and phenotypes, or markers and services.

Eventually, causal mutations will be identified in a growing number of cases. At that stage, a reduced number of informative markers might be set up and used across populations for a predictive approach of adaptation and performance in a given environment. Another outcome of whole-genome studies across populations would be to identify mechanisms (re)generating genetic diversity in domestic populations, such as CNVs for instance, or epigenetic heritable marks. Intensive selection has been applied for about 50 years, which is very short compared to the time elapsed since domestication. Has this process affected genome dynamics or not ? It does not seem to be the case for CNVs, inasmuch as it has been shown that most of them were shared between wild and domestic pigs (Paudel et al., 2013).

Data sharing. To achieve reliable associations between genetic markers and phenotypes or services, extensive data sharing will be needed. This is a major gap since data ownership is generally not public and data sharing will encounter issues of confidentiality and economic competition. Data sharing requires a careful balance between risks and benefits for data providers and data users, which will generally require a collaborative agreement. It is thus not realistic to think that a single database could include all relevant information for the characterization of genetic resources. A more realistic strategy would rather be to rely on the concept of metadata, which can be used to describe what has been measured on which populations, when, how and by whom, and what are the conditions to access to these data. Metadata make it also possible to resolve the issue of heterogeneous data, as long as this heterogeneity can be described. Describing metadata needs an ontology, and we have seen that ontologies are being developed for livestock. Once datasets are described by metadata, the most useful way to share data is to set up a web portal with a catalogue of datasets, but not a catalogue of data which remain the property of the initial data collector. Of course, the data provider has the possibility to put data in open access, which does not mean there is no cost attached to it. As an example, the Global Biodiversity Information Facility (GBIF; http://www.gbif.org/) gathers 14,728 datasets from 590 data providers, dealing with occurrences of wild species worldwide; the central facility is collecting metadata through national nodes, which looks very much like DAD-IS and national coordinators. In the case of GBIF, the descriptors were standardized by the community of natural sciences and museums of natural history. GBIF is thus a data infrastructure which is stimulating re-use of data for research but also for managers of conservation programs. Domestic breeds represent a within-species component of biodiversity, with a range of scientific and economic objectives, and it would be worthwhile to transpose the concept of a web portal pointing to the numerous relevant datasets for animal genetic resources.

Access to genetic resources. Data on genetic resources are obtained from animals and can be connected to biological samples derived from them, either reproductive or genomic. Characterization is a major step for cryobanking in order to facilitate future use of cryobank samples for research, but also for facing future challenges by the livestock sector. Cryobanks offer the appropriate support to couple reproductive and genomic resources: molecular characterization done on genomic samples will improve the documentation of the reproductive samples, and reciprocally reproductive samples can be used to produce animals of a given genotype in order to collect additional phenotypic data with new tools or in new environments. This concept of a dynamic repository coupling genomic and reproductive samples is developing worldwide and will be illustrated in this conference. It is supported in France by the CRB-Anim national infrastructure (http://www.crb-anim.fr/crb-anim_eng/) which complies with the OECD definition of a Biological Resource Center.

In 2010, the Nagoya protocol for Access to genetic resources and Benefit Sharing (ABS) was adopted by the Convention of Biological Diversity. Implementation of ABS is likely to impact the day-to-day operations of cryobanks that will be involved in the traceability and documentation of samples. Minimal descriptors for material entering cryobanks are not currently defined: ideally, they should include not only the breed information but be geo-referenced and provide also a description of the production system, in order to facilitate future characterization of the preserved resources. The Nagoya protocol may be expected to trigger more extensive characterization of genetic resources since such data should provide a fair basis for negotiating benefit sharing between provider and user.

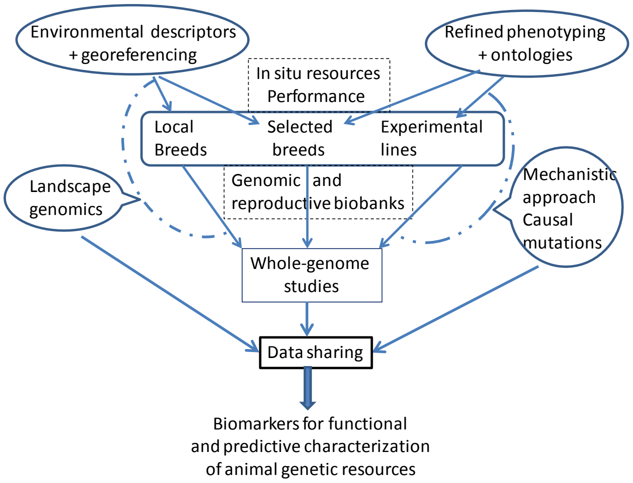

Figure 1. A tentative scheme for an integrated characterization of livestock genetic diversity

Conclusion

The options presented to fill the gaps underline the need to share data and methods, which should also decrease the effort and the cost of characterization for a given breed. Indeed, collecting data on a subset of breeds provide useful information to the characterization of a single breed in a similar environment. A more integrated characterization of livestock genetic diversity can be proposed (Figure 1) which should provide the data needed for a more prospective management of genetic resources. Elaborating scenarios is now considered as a method of choice to shed light on the main options and possible future of biodiversity on a global scale (Pereira et al., 2010), the same approach would be interesting to apply to within-species genetic diversity, and this requires reliable and accessible data on genotypes and phenotypes.

References

Broom, D. M., Galindo, F. A., and Murgueitio, E. (2013). Proc. Royal Soc. B-Biol. Sci. 280: 2013-2025.

FAO, (2012). Phenotypic characterization of animal genetic resources. FAO Animal Production and Health Guidelines N°11, Rome.

Gautier, M., Flori, L., Riebler, A. et al. (2009). BMC Genomics, 10: 550.

Gautier, M., Laloë, D., Moazami-Goudarzi, K. (2010). PLoS ONE, 5: e13038.

Granevitze, Z., Hillel, J., Feldman, M. et al. (2009). Anim. Genet. 40: 686-693.

Groenen, M.A.M., Archibald, A.L., Uenishi, H. et al. (2012). Nature, 491: 393-398.

Hoffmann, I. (2010). Anim. Genet. 41 (suppl. 1): 32-46

Laloë, D., Moazami-Goudarzi, K., Lenstra, J.A. et al. (2010). Diversity, 2:932-945.

Lele, S., Springate-Baginski, O., Lakerveld, R. et al. (2013). Conservation and Society, 11: 343-358.

Lenstra, J.A., Groeneveld, L.F., Eding, H. et al. (2012). Anim. Genet. 43: 483-502.

Leroy, G., Kayang, B.F., Youssao, I.A.K. et al. (2012). BMC Genet. 13:34.

Miao, Y-W., Peng, M-S., Wu, G-S. et al. (2013). Heredity, 110: 277-282.

Naderi S., Rezaei, H-R., Pompanon, F. et al. (2008). P.N.A.S. 105: 17659-17664.

Pariset, L., Joost, S., Ajmone Marsan, P. et al. (2009). BMC Genet. 10: 7.

Paudel, Y., Madsen, O., Megens, H-J. et al. (2013). BMC Genomics, 14: 449.

Pereira, H.M., Leadley, P.W., Proenca, V. et al. (2010). Science, 330:1496-1501.

Presson, A.P., Sobel, E., Lange, K. et al. (2006). J. Comput. Biol. 13: 1131-1147.

Rubin, C-J., Zody, M.C., Eriksson, J. et al. (2010). Nature, 464: 587-593.

Rubin, C-J., Megens, H-J., Barrio, A.M. et al. (2012). P.N.A.S. 27: 19529-19536.

SanCristobal, M., Chevalet, C., Haley, C.S. et al. (2006). Anim. Genet. 37: 189-198.

Sere, C., van der Zijpp, A., Persley, G. et al. (2008). AGRI, 42 : 3-24.

Silvestri, S., Zaibet, L., Said, M.Y. et al. (2013). Environmental Science & Policy, 31: 23-33.

Tixier-Boichard, M., Ayalew, W., Jianlin, H. et al. (2008). AGRI, 42: 29-44.

Vaysse, A., Ratnakumar, A., Derrien, T. et al. (2011). PLoS Genet. 7: e1002316.

Wilkinson, S., Lu, Z.H., Megens, H-J. et al. (2013). PLoS Genet. 9: e1003453.