E

联合国粮食及农业组织 | Food and Agriculture Organization of the United Nations | Organisation des Nations Unies pour l'alimentation et l'agriculture | Продовольственная и сельскохозяйственная организация Объединенных Наций | Organización de las Naciones Unidas para la Alimentación y la Agricultura | منظمة الأغذية والزراعة للأمم المتحدة

COMMISSION ON GENETIC RESOURCES FOR FOOD AND AGRICULTURE

Item 10.2 of the Provisional Agenda

Eighteenth Regular Session

27 September – 1 October 2021

DRAFT PRACTICAL GUIDE ON GENOMIC CHARACTERIZATION OF ANIMAL GENETIC RESOURCES

TABLE OF CONTENTS

BACKGROUND

1. The Commission on Genetic Resources for Food and Agriculture (Commission), at its Thirteenth Regular Session,[1] endorsed the *FAO guidelines – Molecular genetic characterization of animal genetic resources*,[2] which were published in 2011. The guidelines: (i) outline the rationale for undertaking characterization of animal genetic resources, including molecular genetic characterization; (ii) describe the strategic choices to be made in planning molecular genetic characterization studies; (iii) provide explanations and recommendations on the steps to take while performing such studies, including animal sampling, genotyping and data analysis, highlighting potential pitfalls; and (iv) encourage standardization of data and integration of national studies into international analyses. The guidelines highlighted the need for their periodic updating and further refinement as experience with their use in the field is accumulated and as technologies for molecular genetic characterization advance.

2. At its Seventeenth Regular Session, the Commission requested FAO to continue developing and updating guidelines to facilitate the application of new scientific discoveries related to the identification, characterization and conservation of animal genetic resources.[3] It further requested FAO to strengthen partnerships with stakeholders and donors to continue technical and policy support for country implementation of the Global Plan of Action for Animal Genetic Resources.[4]

3. Biotechnologies for the sustainable use and conservation of genetic resources for food and agriculture have advanced substantially in recent years.[5] Genomics – the study of genes and their functions, and related characterization techniques[6] – is a field in which technological advancement has been particularly rapid and consequential. Enhancements in procedures for whole-genome sequencing and in high-throughput genotyping have led to greatly decreased costs per unit of genetic information that can be obtained from a single assay. Since the publication of the previous guidelines, thousands of individual animals have undergone whole-genome sequencing and hundreds of thousands of animals have been genotyped with DNA chips, each assaying thousands of single nucleotide polymorphism markers. Application of this information has increased knowledge of breed development and accelerated response to selection. Knowledge of the biology underlying phenotypes has also increased as a result. To help countries to benefit from these technological advancements in applications to improve the management of animal genetic resources, FAO has developed a new practical guide on genomic characterization, which is given in the annex of this document. The document is intended to update and complement the *FAO guidelines – Molecular genetic characterization of animal genetic resources*.

4. The draft practical guide has been prepared in cooperation with the International Society of Animal Genetics (ISAG) – FAO Advisory Group on Animal Genetic Diversity (Advisory Group). ISAG is an academic and scientific organization that focuses on basic and applied research on molecular genetics in domesticated animals.[7] FAO and ISAG have a long history of collaboration. Members of the Advisory Group have in the past served as editors and/or contributors to the previous guidelines, as well as the *Secondary guidelines: Measurement of domestic animal diversity,*[8] published in 1993, and the *Secondary Guidelines – Measurement of Domestic Animal Diversity (MoDAD): Recommended Microsatellite Markers*,[9] which were published in 2004. For the current draft practical guide, several Advisory Group members served as editors and authors of the individual

[1] CGRFA-13/11/Report, paragraph 79.
[2] FAO. 2011. *Molecular genetic characterization of animal genetic resources*. FAO Animal Production and Health Guidelines. No. 9. Rome. (also available at http://www.fao.org/3/i2413e/i2413e00.pdf).
[3] CGRFA-17/19/Report, paragraph 84.
[4] CGRFA-17/19/Report, paragraph 86.
[5] CGRFA-18/21/6; CGRFA-18/21/6/Inf.1.
[6] https://www.who.int/genomics/geneticsVSgenomics/en/
[7] https://www.isag.us/
[8] FAO. 1993. *Secondary guidelines: Measurement of domestic animal diversity (MoDAD)*. Rome
[9] FAO. 2004. *Secondary guidelines for Development of National Farm Animal Genetic Resources Management Plans: Measurement of domestic animal diversity (MoDAD): Recommended Microsatellite Markers*. Rome. (also available at http://www.fao.org/3/aq569e/aq569e.pdf).

sections. In addition, other Advisory Group members and international scientists were invited to serve as authors and/or reviewers of the document.

5. The draft practical guide was developed by using the previous guidelines as a template. Information that is still valid was retained, while in other instances the material was updated. New material was included to explain the relevant recently developed technologies. These developments in biotechnologies have been complemented by advances in methods to analyse and interpret the new information gathered, so the practical guide also contains new material to address the application of genomic methods to the study of animal genetic resources.

6. In addition to advancements in biotechnologies, important changes in the policy landscape since the publication of the previous guidelines have also occurred. In particular, the *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity*[10] is of relevance for genomic characterization of animal genetic resources, given the importance of international cooperation for understanding national animal genetic resources in a wider context and the frequent movement of biological samples across borders while undertaking associated research. The practical guide explains the relevance of the protocol and shares advice on steps for researchers to take to help ensure compliance with applicable national laws on access and benefit-sharing.

7. A key aim of the practical guide is to inform readers on how to undertake a genomic characterization study on animal genetic resources from start to finish. The document explains different approaches for genomic analysis and considers different study objectives. The sections of the practical guide address the following topics: (i) rationale for genomic characterization of animal genetic resources; (ii) the basics of carrying out molecular genetic studies; (iii) genomic tools and methods; (iv) genomic applications for assessment of genomic variation within and between populations; and (v) overall conclusions and recommendations. The practical guide also includes several appendices that provide detailed information and examples related to specific issues of importance. For genomic characterization of animal genetic resources.

[10] https://www.cbd.int/abs

ANNEX

Draft practical guide on genomic characterization of animal genetic resources

Table of contents

Foreword

Acknowledgements

User guidance

Abbreviations and acronyms

AB	ascertainment bias
ABC	approximate Bayesian computation
ABS	access and benefit sharing
aDNA	ancient DNA
AFS	allele frequency spectrum
AIM	ancestry informative marker
AMOVA	analysis of molecular variance
AnGR	animal genetic resources (for food and agriculture)
bp	base pair
CBD	Convention on Biological Diversity
CNV	copy-number variation/variant
DAD-IS	Domestic Animal Diversity Information System
DNA	deoxyribonucleic acid
D_R	Reynolds genetic distance
EAAP	European Federation of Animal Science
FA	factor analysis
Gb	gigabase
GBS	genotyping-by-sequencing
GWAS	genome-wide association study
HDP	high density panel of markers
H_e	expected heterozygosity
H_o	observed heterozygosity
HWE	Hardy–Weinberg equilibrium
IBD	identity by descent
IBS	identity by state
ILRI	International Livestock Research Institute
indels	insertions and deletions (of nucleotides)
ISAG	International Society for Animal Genetics
K	number of genetic groups in model-based clustering analyses
kb	kilobase
LD	linkage disequilibrium
LDP	low density panel of markers
LRS	long-read sequencing
MAF	minor allele frequency
MAT	mutually agreed terms
Mb	megabase
MCMC	Monte Carlo Markov chain
MDS	multidimensional scaling
MSMC	multiple sequentially Markov coalescent method
MSY	male-specific part of the Y-chromosome
MTA	material transfer agreement
mtDNA	mitochondrial DNA
N_b	number of breeding animals
N_e	effective population size
NJ	neighbour-joining
PAR	pseudoautosomal region
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
π	nucleotide diversity
PIC	prior informed consent
PSMC	pairwise sequentially Markovian coalescent method
QTL	quantitative trait locus

ROH	runs of homozygosity
SD	standard deviation
SNP	single nucleotide polymorphism
SPA	Strategic Priority Area (of the Global Plan of Action)
SPAG	spatial areas of genotype probability
SRS	short-read sequencing
SV	structural variation
θ	population mutation rate
WGS	whole-genome sequencing

SECTION 1

Introduction

INTRODUCTION

RATIONALE FOR CHARACTERIZATION OF ANIMAL GENETIC RESOURCES

Domestic animal diversity is an important component of global biodiversity. About 40 species of domestic animals and poultry contribute to meeting the needs of humankind, providing meat, fibre, milk, eggs, draught animal power, sport and recreation, skins, and manure, and are an essential component of many mixed farming systems. Within these species, around 8 800 breeds and strains (FAO, 2021) constitute the animal genetic resources (AnGR) that are of crucial significance for food and agriculture.

The present pattern of diversity of AnGR is the result of a long and complicated history that started with animal domestication. Depending upon the species, domestication occurred 10 000 to 1 000 years ago. Since then, domestic livestock have spread with human migration and trading to all inhabited continents. Local adaptation, artificial selection, mutation and genetic drift turned the genetic diversity captured with domestication into a vast array of differences in appearance, physiology and agricultural traits. During recent centuries this differentiation has been accentuated by the emergence and development of breeds – more or less isolated populations that were subject to systematic selection. This development has been most pronounced in the temperate zones where the demands of food supply led to a rationalization of agriculture. The last 50 years have seen the global spread of a few highly developed breeds, most of which originated in Europe. A well-known example is the high-yielding Holstein-Friesian breed of dairy cattle, which has become by far the most widely dispersed cattle breed in the world.

The global diffusion of these specialized breeds is endangering or even risking the extinction of many well-adapted local breeds/populations. This trend is occurring both moderately to highly intensive production environments and in marginal areas (Godber and Wall, 2014, Sponenberg *et al*., 2018), where local husbandry practices are being abandoned (Köhler-Rollefson *et al.*, 2009). Local breeds are usually much less productive than the highly developed international transboundary breeds when raised in optimal conditions but are adapted to the local climate (Mirkena *et al*., 2010) and do perform in a natural environment without intensive management. Indiscriminate crossbreeding and introduction or increased use of the specialized exotic breeds have been reported as the two most important causes of genetic erosion at global level (FAO, 2015).

This erosion of diversity of AnGR has become a major concern (Hodges, 2006, FAO 2007a, Bruford *et al*., 2015). The negative consequences of genetic erosion and inbreeding depression have been amply documented and may be manifested by loss of viability, fertility and disease resistance, and the frequent occurrence of recessive genetic diseases (FAO 2007b; Taberlet *et al.,* 2008; Howard *et al*., 2017). According to the report on the *Status and trends of animal genetic resources – 2020* (FAO, 2021), approximately 7 percent of reported livestock breeds have become extinct, and more than 25 percent are considered to be at risk of extinction. Moreover, the situation is presently unknown for more than 50 percent of breeds, most of which are reared in developing countries.

Several recent developments have strengthened the interest in local breeds. First, we now realize that their local adaptation will become even more essential in view of the ongoing and escalating climate change (Hoffmann *et al*., 2010). Second, there is a growing interest in the restoration of ecosystems (UN, 2020), including "rewilding" of abandoned agricultural areas, in which domestic ungulates are kept under feral or semi-feral conditions and occupy the niche of mega-herbivores (Carey, 2016). For this purpose, only animals that are adapted to extensive management are suitable. Third, the global COVID-19 pandemic has increased the awareness about zoonotic diseases, and the potential benefits (Liverani *et al*., 2013) of a high biodiversity of wildlife, livestock and production systems (Simianer and Reimer, 2021).

FAO has a history of working with countries and other stakeholders to improve the productivity of livestock and the livelihoods of their citizens while maintaining AnGR diversity (FAO, 1990a,b). Specific priorities for AnGR management are set out in the *Global Plan of Action for Animal Genetic Resources* (Global Plan of Action) (FAO, 2007a), the internationally agreed policy framework for management of AnGR.

One of the Strategic Priority Areas of the Global Plan of Action is the characterization, inventory and monitoring of trends in AnGR diversity. In brief, this Strategic Priority Area addresses the gathering of information to increase the knowledge about AnGR. Knowledge is necessary to properly assess the value of breeds and to guide decision making in livestock development and breeding programmes. Assessments of the capacity to manage AnGR have revealed that a lack of knowledge is a major constraint (for example, FAO, 2007b; 2015)

In these guidelines, the words "breed" and "population" will be used almost interchangeably. Technically, "population" is a more-general term and includes both well defined "breeds" and on groups of animals that have not yet been defined as a breed. The breed is the most common operational unit in the conservation of genetic resources. However, the biological breed concept to describe groups of animals having particular genetic characteristics is not always applicable. Many breeds originating from industrialized countries are well-defined and phenotypically distinct and have been largely isolated genetically throughout the course of their development. In contrast, a significant proportion of other breeds, especially those in Asia and Africa, correspond to local populations without structured breeding programmes, and therefore differ only gradually according to geographical separation. In addition, breeds with different names may have a recent common origin (Felius *et al*., 2011) and crossbreeding has been common since the invention of breeds in the 18th century.

It is often argued that breeds, by being associated with a group of people, is a social construct rather than a biological concept ("A breed is a breed if enough people say it is", K. Hammond as cited by Oldenbroek and Van der Waaij, 2015). On the other hand, recent developments in biotechnologies have greatly increased the power of molecular analyses to discern the genetic make-up of organisms and identify differences and similarities among populations. Molecular-genetic analysis can now complement phenotypic characterization and indigenous knowledge to help identify the breeds or groups of breeds that have retained their uniqueness, and to gather information to help guide their future management. Genomic characterization is a useful tool regardless of the precision with which a "breed" is defined.

MOLECULAR CHARACTERIZATION – HISTORY AND PROSPECTS

Scientists began use molecular data for livestock in the early 1990s and since then these data have become continually more relevant for the characterization of genetic diversity (Groeneveld *et al.*, 2010; Bruford *et al*., 2015). In 1993, an FAO working group proposed a global programme for characterization of AnGR, including molecular genetic characterization, and formulated the *Secondary guidelines: measurement of domestic animal diversity (MoDAD)* (FAO, 1993) with recommendations for the molecular analysis of domestic animal diversity on a global scale.

The FAO MoDAD report succeeded in creating awareness of the need to monitor AnGR diversity. In addition, the proposal of the program helped motivate many nationally funded research projects and larger regional and international projects supported by organizations such as the European Commission, the Nordic Council of Ministers, the International Atomic Energy Agency (IAEA), the International Livestock Research Institute (ILRI) and the World Bank.

Scientists in many countries have undertaken studies to characterize locally available breeds, while large-scale international efforts on breed characterization have built comprehensive molecular datasets for most livestock species. The study of genetic diversity of livestock at the molecular level has developed into an active area of research that frequently contributes to the capacity building of young scientists and receives considerable attention in scientific press and at the conferences of organizations such as the International Society for Animal Genetics (ISAG) and the European Federation of Animal Science (EAAP).

The first MoDAD guidelines were followed-up in 2011 by the FAO report *Molecular genetic characterization of animal genetic resources* (FAO 2011). At that time, the state-of-the-art technology was still largely based on the use of microsatellites as a genetic marker, which after 1990 had revolutionized the science of molecular genetics. However, this category of markers has two major disadvantages for studying genetic diversity. First, allele calling is difficult to reproduce across

laboratories. Second, in spite of the efforts of FAO and ISAG towards standardization, many laboratories continued to use private panels of genetic markers, which precluded the joining of datasets and seriously decreased the impact of such studies.

Starting about 10 to 15 years ago, these markers have for almost all research applications become outdated by the availability for most livestock species of whole-genome sequences. The genomic sequences allowed the identification of millions of single nucleotide polymorphisms (SNPs) and other types of genetic variaiants such as insertions and deletions (indels), structural variation (SV), copy-number variation (CNV). The invention of "bead arrays" allowed the simultaneous genotyping of 10 000 to 1 000 000 SNPs (Nicolazzi *et al*., 2015). Subsequently, whole-genome sequencing (WGS) (Eusebi *et al*., 2020) became more and more affordable. Section 3 will describe these approaches in more detail.

These developments have generated an ongoing "tsunami" of genotype and sequence datasets and have been accompanied by the development of a multitude of new statistical analyses and programs for data management and analysis (Biscarini *et al*., 2018). Another consequential development has been the substantial progress in the analysis of ancient DNA (aDNA) (McHugo *et al*., 2019). All these developments have contributed to the realization of most of the original MoDAD objectives, albeit with different markers than envisaged originally:

- The wild ancestral species of most livestock species were already identified based on mitochondrial DNA (mtDNA). More detailed studies, often using aDNA, have yielded more details on the interaction of several livestock species and their wild ancestors during and after domestication (MacHugh *et al*., 2017).
- The differentiation of breeds as well as their homogeneity can now be assessed unambiguously. This is in itself relevant, but also guides downstream analyses that assume breeds to be homogeneous populations.
- The genetic constitution of breeds can now be more accurately assessed via quantitative measures of diversity, admixture or subdivision, inbreeding, introgression and assortative mating.
- The evolutionary history of species and populations can be reconstructed on the basis of the phylogenetic relationships of breeds. The resulting data are commonly visualized by a "tree-like" (hierarchical) topology, but may also indicate gene flow between breeds with different histories.

For the most important domestic species we have obtained a fairly comprehensive global view of diversity – that is, also including Asian, African and South American breeds as well as those from high-income countries - by integrating national or regional datasets of SNP genotypes or whole genome sequences, even if these datasets originated from different laboratories (in contrast to studies based on microsatellites).

Other positive developments are the realization of more effective cryoconservation programs (Paiva *et al*., 2016; De Oliveira *et al*., 2019) and the globalization of the diversity studies with active participation of institutes from all inhabited continents. The genomic characterization of breeds also led to the insight that livestock breeds have never been static phenomena, but rather are the result of a continuously dynamic history that has involved selection pressures that change over time and has often included crossbreeding (Felius *et al*., 2015). As a consequence, attempts to establish priorities for conservation of breeds on the basis of neutral genetic markers (Lenstra *et al*., 2012) have been de-emphasized.

Instead, attention has shifted toward the adaptive variation, SNPs, indels and CNVs that possibly contribute to environmental adaptation or other relevant traits. At least for monogenic traits, most of the adaptive variation resides in the structural genes (Nicholas and Hobbs, 2014), but there are interesting examples of intergenic mutations (Zappala *et al*., 2016; Aldersey *et al*., 2020). In addition, several methods have been developed for detection of a local perturbance of the diversity across the genome that may suggest the effect of selection, the so-called selection signatures (Maynard Smith and Haigh, 1974; Randhawa *et al*., 2016; Friedrich and Wiener, 2019). In all these studies, the limiting factor is the validation of the signals found by genome-wide analysis or by studies of gene expression

in terms of direct causative effects. Such genome-wide association studies (GWAS) have also been facilitated by the availability of new genomic tools. Genomics are continually creating new opportunities for increased response to selection and improved management of genetic diversity (Bruford *et al*., 2015; Oldenbroek and Van der Waaij, 2015).

From a fundamental-scientific view, it will also be most interesting to link phenotypes to non-coding RNA molecules (Weikard *et al*., 2016 and to epigenetic phenomena Giuffra *et al*. (2019). Because of the multitude of genes, traits and breeds, this promises to remain a large and rewarding area of research.

REFERENCES

Aldersey, J.E., Sonstegard, T.S., Williams, J.L. & Bottema, C.D.K. 2020. Understanding the effects of the bovine POLLED variants. *Animal Genetics*, 51: 166–176. https://doi.org/10.1111/age.12915.

Biscarini, F., Cozzi, P., & Orozco-ter Wengel, P. 2018. Lessons learnt on the analysis of large sequence data in animal genomics. *Animal Genetics*, 49: 147-158. http://doi.org/10.1111/age.12655.

Bruford, M. W., Ginja, C., Hoffmann, I., Joost, S., Orozco-terWengel, P., Alberto, F.J., Amaral, A. J. *et al.* 2015. Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. *Frontiers in Genetics*, 6: 314. http://doi.org/http://doi.org/10.3389/fgene.2015.00314.

Carey, J. 2016. Rewilding. *Proceedings of the National Academy of Sciences of the United States of America*,, 113: 806-808. http://doi.org/10.1073/pnas.1522151112.

De Oliveira Silva, R., Ahmadi, B. V., Hiemstra, S. J., & Moran, D. 2019. Optimizing ex situ genetic resource collections for European livestock conservation. *Journal of Animal Breeding and Genetics*. 136: 662-73. http://doi.org/http://doi.org/10.1111/jbg.12368.

Eusebi, P. G., Martinez, A., & Cortes, O. 2020. Genomic tools for effective conservation of livestock breed diversity. *Diversity*, 12: 8. http://doi.org/http://doi.org/10.3390/d12010008.

FAO. 1990a. *Animal genetic resources. A decade of progress, 1980–1990*, by J. Hodges. Animal Production and Health Paper. Rome.

FAO. 1990b. *Manual on establishment and operation of animal gene banks*, by J. Hodges. Animal Production and Health Paper. Rome.

FAO. 2007a. *Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration*. Rome (available at http://www.fao.org/docrep/010/a1404e/a1404e00.htm).

FAO. 2007b. *The State of the World's Animal Genetic Resources for Food and Agriculture*, edited by D. & B. Rischkowsky. Rome.

FAO 2011. *Molecular Genetic Characterization of Animal Genetic Resources.* (available at http://www.fao.org/docrep/014/i2413e/i2413e00.pdf)

FAO. 2015. *The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture*, edited by B.D. Scherf & D. Pilling. FAO Commission on Genetic Resources for Food and Agriculture Assessments. Rome (available at http://www.fao.org/3/a-i4787e/index.html).

FAO. 2021. *Status and trends of animal genetic resources – 2020*. FAO. Rome.

Felius, M., Koolmees, P.A., Theunissen, B., Lenstra, J.A. & European Cattle Genetic Diversity Consortium. 2011. On the breeds of cattle—Historic and current classifications. *Diversity*, 3: 660-692. http://dx.doi.org/10.3390/d3040660

Felius, M., Theunissen, B., & Lenstra, J.A. 2015. Conservation of cattle genetic resources: the role of breeds. *The Journal of Agricultural Science*, 153: 152-162. http://doi.org/10.1017/S0021859614000124.

Friedrich, J., & Wiener, P. 2020. Selection signatures for high-altitude adaptation in ruminants. *Animal Genetics*, 51: 157-162. http://doi.org/10.1111/age.12900.

Giuffra, E., & Tuggle, C.K. 2019. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annual Review of Animal Biosciences, 7(1): 65-88. http://doi.org/10.1146/annurev-animal-020518-114913.

Hodges, J. 2006. Conservation of genes and culture: historical and contemporary issues. *Poultry Science*, 85: 200–209.

Godber, O. F., & Wall, R. 2014. Livestock and food security: vulnerability to population growth and climate change. *Global Change Biology*, 20: 3092-3102. http://doi.org/10.1111/gcb.12589.

Hoffmann, I. 2010. Climate change and the characterization, breeding and conservation of animal genetic resources. *Animal Genetics*, 41: 32-46. https://doi.org/10.1111/j.1365-2052.2010.02043.x

Howard, J. T., Pryce, J. E., Baes, C., & Maltecca, C. 2017. Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. J*ournal of Dairy Science*, 100: 6009-6024. http://doi.org/10.3168/jds.2017-12787.

Köhler-Rollefson, I., Rathore, H.S. & Mathias, E. 2009. Local breeds, livelihoods and livestock keepers' rights in South Asia. *Tropical Animal Health and Production*, 41: 1061–1070. https://doi.org/10.1007/s11250-008-9271-x.

Liverani, M., Waage, J., Barnett, T., Pfeiffer, D. U., Rushton, J., Rudge, J. W., Loevinsohn, M.M., Scoones, I., Smith, R.D., Cooper, B.S., White, L.J., Goh, S., Horby, P., Wren,1 B., Gundogdu, O.,Woods & Coker, R. J. 2013. Understanding and managing zoonotic risk in the new livestock industries. *Environmental Health Perspectives*, 121: 873-877. http://doi.org/10.1289/ehp.1206001.

MacHugh, D. E., Larson, G., & Orlando, L. 2017. Taming the past: Ancient DNA and the study of animal domestication. *Annual Review of Animal Biosciences*, 5: 329-351. http://doi.org/10.1146/annurev-animal-022516-022747.

Maynard Smith, J. & Haigh, J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23: 23–25.

McHugo, G. P., Dover, M. J., & MacHugh, D. E. 2019. Unlocking the origins and biology of domestic animals using ancient DNA and paleogenomics. *BMC Biology*, 17: 98. http://doi.org/10.1186/s12915-019-0724-7.

Mirkena, T., Duguma, G., Haile, A., Tibbo, M., Okeyo, A. M., Wurzinger, M., & Sölkner, J. 2010. Genetics of adaptation in domestic farm animals: A review. *Livestock Science*, 132: 1-12. http://doi.org/10.1016/j.livsci.2010.05.003.

Nicholas, F.W. & Hobbs, M. 2014. Mutation discovery for Mendelian traits in non-laboratory animals: a review of achievements up to 2012. *Animal Genetics*, 45:157-70. http://doi.org/10.1111/age.12103.

Nicolazzi, E. L., Biffani, S., Biscarini, F., Orozco ter Wengel, P., Caprera, A., Nazzicari, N., & Stella, A. 2015. Software solutions for the livestock genomics SNP array revolution. *Animal Genetics*, 46(4): 343-353. http://doi.org/10.1111/age.12295.

Oldenbroek, K. & Van der Waaij, L. 2015. *Textbook Animal Breeding and Genetics for BSc students.* Wageningen UR (available at https://wiki.groenkennisnet.nl/display/TAB/).

Paiva, S. R., McManus, C. M., & Blackburn, H. 2016. Conservation of animal genetic resources – A new tact. *Livestock Science*, 193. http://doi.org/10.1016/j.livsci.2016.09.010.

Randhawa, I. A. S., Khatkar, M. S., Thomson, P. C., & Raadsma, H. W. 2016. A meta-assembly of selection signatures in cattle. *PLOS ONE*, 11: e0153013. http://doi.org/10.1371/journal.pone.0153013.

Simianer, H. & Reimer, C. 2021. COVID-19: a "black swan" and what animal breeding can learn from it, *Animal Frontiers*, 11: 57–59. https://doi.org/10.1093/af/vfaa046

Sponenberg, D. P., Beranger, J., Martin, A. M., & Couch, C. R. 2018. Conservation of rare and local breeds of livestock. *Revue Scientifique et Technique de l'OIE*, 37: 259-267. http://doi.org/10.20506/rst.37.1.2756

Taberlet, P., Valentini, A., Rezaei, H.R., Naderi, S., Pompanon, F., Negrini, R. & Ajmone-Marsan, P. 2008. Are cattle, sheep, and goats endangered species? *Molecular Ecology*, 17: 275–84.

Weikard, R., Demasius, W., & Kuehn, C. 2017. Mining long noncoding RNA in livestock. *Animal Genetics*, 48: 3-18. http://doi.org/10.1111/age.12493

UN. 2020. *United Nations Decade on Ecosystem Restoration.* (available at https://www.decadeonrestoration.org/ [accessed December 2020].

Zappala, Z., & Montgomery, S. B. 2016. Non-coding loss-of-function variation in human genomes. *Human Heredity*, 81: 78-87. http://doi.org/10.1159/000447453

SECTION 2

The basics of genomic diversity studies

THE BASICS

HOW TO CARRY OUT GENOMIC DIVERSITY STUDIES

This section provides a general overview of the steps to be undertaken when conducting genomic studies of animal genetic resources (AnGR). The process of genomic characterization can usually be broken into a series of activities that are distinct in terms of time and space, but coordination among the steps is critical. Although the genotyping technologies have changed substantially since the previous guidelines (FAO, 2011b), the basic steps for any genetic diversity study have remained largely the same.

PREREQUISITES

Consider the national context for the management of animal genetic resources

Genetic characterization is often undertaken as an academic research activity, which has value on its own, but the potential impact of the study will be greatly increased if undertaken in coordination with national framework for the management of AnGR (FAO, 2011a). Nearly all countries have identified an organization to serve as its National Focal Point for Management of AnGR and have nominated an individual to serve as a National Coordinator (FAO, 2021a). Contacting the National Coordinator prior to the study and keeping him or her informed as the study progresses is recommended for several reasons. Characterization is a key action in national implementation of the Global Plan of Action for Genetic Resources (FAO, 2007) and countries are regularly requested to report their activities to the FAO (e.g. FAO, 2021b). Many countries also have national strategies and action plans for management of AnGR (FAO, 2009) and characterization studies should be consistent with such plans. The National Coordinator should also be aware of the stakeholders that will be interested in the results and their application in management of the characterized populations. To maximize the efficiency of your data collection efforts, phenotypic characterization (FAO, 2012) of the populations and their production system should be undertaken in concert with sampling of biological material for genomic characterization.

Know your breeds

Characterization is itself an information-gathering activity, but efforts should be made to obtain as much background knowledge as possible about the target populations while planning your study. Sources of information will include the comprehensive breed encyclopedia of Porter *et al.,*(2016), scientific literature, national technical reports, popular press articles, breed databases such as the Domestic Animal Diversity Information System (DAD-IS; FAO, 2021a) and the Breeds of Livestock (Oklahoma State University, 2015) and websites of breed organizations. This information will help to formulate objectives, plan sampling strategies and develop hypotheses to be tested.

Written information should be complemented with input from local experts, livestock keepers and representatives from breed associations that are familiar with the herdbooks and breed history. These stakeholders may be able to share anecdotal information about breeds and provide contacts for sampling of biological material and measurements of phenotypic traits.

Define clear objectives

Genomic analyses can be used to address a wide range of objectives and can be applied in many ways (see Section 4). These objectives are important for deciding on the number and type of animals to be sampled, the optimal genetic markers to be used and complementary information to be collected during sampling. A common objective is to elucidate the relationships of a group of local breeds and their relationships to other breeds in the same geographical area, especially if these breeds are ancestral to one of more of the local breeds. For example, Kim *et al*., (2020) analysed the ancestry indigenous African cattle and compared these cattle with European and Asian breeds. They found that the seemingly disctinct “breeds” shared a common history of gene flow. Many livestock breeds have already been subject to genomic characterization and the data from those studies is often in the public domain. Therefore, comparison of targeted breeds with other breeds may not require sampling and

genotyping of the complementary breeds. However, such analyses require a preliminary assessment to ensure that genotyping approaches were similar and shared common markers.

Design the sampling

As noted earlier, genomic characterization will ideally be done in concert with phenotypic characterization and evaluation of the production environment. This will especially be the case when the study represents the first time that a population is characterized, and the objective is to capture the range of variability. To accomplish this, the sampling plan should consider the structure of the production environment, geographic locations and familial relationships (as much as possible). The following recommendations will help ensure a genetically diverse sample:

- preferably, sample in the areas that are close to the site of the development of the breed(s);
- depending on the objectives of the study, relevant populations in other local regions or countries (transboundary breeds) may be sampled as well;
- if applicable, sample animals that represent all different subtypes or subpopulations from different agroclimatic zones (which should be recorded in the complementary data, see below);
- typically, to improve representativeness avoid sampling of animals that are more closely related than the population in general, no more than 10 percent of any one herd or village population should be sampled and at most five animals be sampled from any single herd; and
- avoid sampling animals with common grandparents.

When planning for collection of samples from animals in different agroclimatic zones or any large area, a systemic approach is recommended. Two possible options are grid-sampling and sampling along linear transects. With grid-sampling, a "grid" of squares is formed by drawing sets of equally spaced North-South and East-West running lines over a map of the sampling area. Equal numbers of animals are then sampled from each square. With linear transects, straight lines are drawn across the sampling area and animals are sampled (as much as possible) from the farms or herds that fall along (or close to) the lines.

Choose the genetic marker technology

The type of genetic marker technology used will influence the types of inferences that can be drawn from a genetic characterization study, and thus must be chosen judicisously. As a basic rule, the most advanced technology that is available for the species to be studied should be chosen, because this method will generally be the most informative. However, practical factors must be taken into consideration to address various trade-offs. The three technologies that are most frequently used for genetic characterization (in order of increasing sophistication) are (i) microsatellites, (ii) single nucleotide polymorphisms (SNP), and (iii) whole genome sequencing (WGS). These technologies were briefly introduced in Section 1.

Because SNP and WGS are more suited to assaying the entire genomcs of organisms are thus much more appropriate for "genomic characterization", these technologies are emphasized in this document. Nevertheless, microsatellites have a long history for genetic characterization and are still being utilized in situations where practical considerations take centre stage (e.g. Yadav, Arona and Jain, 2017; Madilindi *et al.*, 2020) and thus merit some discussion (Box 1)

BOX 1

Microsatellites: glory and decline

Microsatellites are a type of genetic marker discovered in the 1980s.They consist of repetitive units of 2 to 6 base pairs and are thus also known as "STRs" for short tandem repeat or "SSR" for short sequence repeat. Because of their abundance and fast PCR-based scoring, they revolutionized the genetic localization of heritable traits. In the 1990s, hundreds of microsatellites were discovered and

published for the most common livestock species. Since 2000 they have been used for many genetic diversity studies.

They can be credited with a considerable scientific progress and have also been relevant for the management of livestock genetic resources:

- rapid tests of paternity and identity (still being done);
- genomic localization of loci corresponding to monogenic traits;
- quantitative estimates of the relative diversity of a breed via the expected heterozygosity (He);
- reasonably accurate indication of non-random mating via the FIS heterozygote-deficit parameter; and
- estimates of genetic distances and of differentiation between breeds, revealing common descent, crossbreeding and geographic clines.

These analyses have been performed for global, continental or national populations of most livestock populations (Groeneveld *et al*., 2010).

On the other hand, microsatellites have several limitations, which affect the analysis of molecular diversity.

- With typically no more than one marker per chromosome in most characterization studies, microsatellites do not attain genomic coverage.
- Scoring alleles cannot be automated and is error prone.
- Estimates of genetic diversity depend on the panel of microsatellite markers, so comparisons are valid only within a dataset. Panels recommended by FAO (2011b) have not been universally adopted.
- Absolute allele sizes are not consistent across laboratories, even if the same equipment is used. Consequently, results of different labs with the same marker panel can only be combined by sharing reference samples or by using allelic ladders as a size standard.
- Because of the low number of markers used in most studies, typically 15 to 30, genetic distances between individuals cannot be estimated reliably. Also due to the limited number of markers, localization of genetic traits is very imprecise.
- They are not suitable at all for the recently developed and statistically powerful modes of analysis.)

Since the publication of the previous FAO guidelines on molecular characterization of diversity (FAO, 2011b), microsatellites have been largely replaced by genome-wide bead arrays, whole-genome sequences and genotyping by sequencing (see Section 3). These more informative approaches generally yield superior results.

On the other hand, indirect benefits of microsatellites should not be overlooked. They raised high expectations for genetic characterization and thus catalyzed the formation of a network of international collaborations and consortia, which still are in place. This paved the way for a rapid and most productive implementation of genome-wide and whole genome sequencing.

Do microsatellites still have a place in genetic characterization? They remain appropriate for standardized routine paternity testing. They may also be the only practical option if local options for SNP genotyping or WGS are not available or unreasonably expensive. There are also special cases for their use when merging new information with data from genotypes of historic samples that are no longer available (although this requires retyping at least some of the historical samples as a reference). However, for the reasons outlined above, microsatellites cannot anymore be generally recommended for genome-wide characterization of genetic variation.

Regarding the choice between SNP and WGS, the latter yields much more information and thus provides potentially more precision in susbsequent genomic analyses. WGS also provides information about other forms of polymorphism, such as insertions and deletions (indels) and structural variation (SV). However, WGS is much more expensive, and the greater amount of information produced requires substantially greater human and technical capacity for data storage, computing and analysis,

which may be out of reach in many circumstances, especially in developing countries. Nearly all of the analyses described in these guidelines can be undertaken with data from SNP arrays, so this will be the most appropriate technology for most of the users of these guidelines.

Numbers that count

For reliable estimation of allele frequencies of biallelic markers (such as SNP, at least 20 animals per breed should be typed (Hein, Schierup and Wiuf, 2003). Smaller sample size decreases the precision of estimates of allelic frequencies, a parameter that underlies many genetic characterization analyses. This imprecision may inflate estimates of genetic distances among other breeds, for example. For multiallelic markers at least 40 animals are required. A larger sample size is also recommended for subdivided breeds.

For specific objectives, smaller (larger) samples sizes may be sufficient (required). Based on a comprehensive simulation to evaluate the properties of different statistical methods, a reasonable power to identify selection signatures (Section 4) can be achieved with a smaller sample size (around 15 animals per population) when using a high-density SNP genotypes (e.g. >1 SNP per kb; Ma *et al.*, 2015). For genome-wide association studies (GWAS) more samples will likely be needed (Ball, 2013: see Sections 3 and 4). Collection of more samples than these minimum values will help to increase precision of inferences and provide a back-up in case of low-quality samples.

Know the rules

Genomic characterization involves the collection and analysis of , which is usually regarded as a genetic resource and may generate data of relevance for intellectual property claims. Therefore, it's recommended that a formal agreement for acquisition and exchange of the material is made with the provider of the material prior to the start of the study. The standard approach is for all involved parties to sign a material transfer agreement (MTA), which stipulates the terms of the material exchange, describes how the material will be used and handled after characterization and analysis. The provider may already have a standard MTA and provide access to the genetic resources only upon acceptance of the terms of the MTA. The MTA should also create legal certainty as to what can be done with the material (such as storage or distribution after the study), as well as with any data resulting from the study. Both parties must be informed about any limits to the rights that the provider can cede to the receiver of the material, in accordance with national regulations. In principle, the providers of the material are to maintain ownership throughout the duration of the study. The providers of the genetic material should be regularly informed about the progress of the study and results obtained. An example MTA is provided in Appendix 2.

Furthermore, international exchange of research material involves a specific set of rules. First, AnGR are considered sovereign to their country of origin. Both the providers of material and their collaborators must therefore adhere to the terms of each country's regulations implementing the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization (Nagoya Protocol) of the Convention on Biological Diversity (CBD), such as the signing of an access and benefit sharing (ABS) contract negotiated on mutually-agreed terms (MAT). More information about the Nagoya Protocol can be found in Box 2 and in other specialized documentation prepared by FAO (2019). Second, because biological material is being exchanged, collaborating scientists must be aware of any sanitary regulations involving international shipments and obtain the necessary permits.

BOX 2

Nagoya Protocol: In brief

The Nagoya Protocol was developed to ensure the fair and equitable sharing of the benefits arising out of the utilization of genetic resources across countries, one of the three objectives of the Convention of Biological Diversity (CBD, 2011). This requires appropriate access to genetic resources and appropriate transfer of relevant technologies, considering all rights over those

resources and to technologies and appropriate funding. The Nagoya Protocol was adopted by the CBD in 2010 and has since been ratified by more than 125 member governments.

As confirmed by the Nagoya Protocol, countries, in the exercise of their sovereign rights over natural resources, may provide that access to genetic resources for their utilization shall be subject to their "prior informed consent" (PIC). Many countries have decided to not restrict access to AnGR. However, where domestic legislation or regulatory requirements require this, users need to obtain permission prior to accessing and using genetic resources for research and development, and share the benefits arising from this utilization of genetic resources on mutually agreed terms (MAT). The Nagoya Protocol, as well as many national ABS laws, do not go into any detail as to what constitutes "utilization" of a genetic resource and therefore requires PIC and MAT. Certain activities, such as taxonomic research which may or may not be further developed commercially, could require PIC and MAT in one country and not in another. It is therefore important to specify in detail the intended activities being undertaken and to clarify with the comptent authority of the country that provides the genetic resources if and what kind of permit is required.

In some countries, access to traditional knowledge associated with genetic resources that is held by indigenous peoples and local communities is also subject to their PIC or approval and involvement and MAT have to be established.

Countries that ratify the Nagoya Protocol are expected to ensure that genetic resources utilized within their jurisdiction have been accessed in accordance with PIC and that MAT have been established, as required by the ABS measures of the country from which the resources were obtained. It is therefore crucial to comply with the applicable ABS measures irrespective of where the genetic resources are actually used for research and development. This is because Parties to the Nagoya Protocol are under an obligation to provide that any research and development activity on genetic resources within their jurisdiction complies with the ABS measures under which the genetic resources have been accessed.

Through so-called "checkpoints" Parties to the Nagoya Protocol ensure that research and development on genetic resources within their territory are based on the PIC of the Party that provided these genetic resources, as applicable. Once PIC has been granted and the permit or its equivalent have been made available to the ABS Clearing House, they constitute an Internationally Recognized Certificate of Compliance.

Source: Hartmut Meyer

The ABS contract and MAT define the limits of utilization and the sharing of benefits. In general, obtaining these contracts involves a phase of negotiation that must be considered before one starts a genomic characterizations study. Box 3 provides key points to consider regarding this negotiation process.

BOX 3

Incorporation of compliance to national ABS frameworks into project planning and implementation

Compliance with ABS measures adds a level of administrative complexity to the execution of reseach and development projects involving AnGR. The ABS measures require specific planning but may help top avoid eventual disputes regarding the sharing of benefits arising from such projects that could otherwise occur in the course of such a project.

The following points regarding ABS compliance should be considered in preparation for genomic characterization studies involving international exchange of AnGR:

(i) Account in the project timeline for the time required to negotiate the ABS contract and obtain PIC, the national ABS permit and other applicable permits (typically 2 months to 2 years, depending on the country).
(ii) Be aware that it may not even be possible to negotiate the required contract.

(iii) Include in the project budget sufficient funds to cover whatever benefits have been negotiated with the providing country. Expected benefits are variable by country; including provider countries as project collaborators may be advantageous, inasmuch as some of the benefits will be in-kind, rather than monetary.

(iv) Consider carefully both present and potential future uses of the AnGR samples and prepare the wording of the contract accordingly. If you want to use the resources for a purpose other than that stated in the original agreement, a new agreement will be required.

(v) Keep all relevant documentation associated with the negotiation of the contract and granting of PIC and the ABS permit. These documents will be crucial for demonstrating compliance with the ABS laws of the providing country if the country in which you are using AnGR has adopted ABS compliance laws (as e.g. the EU and the UK).

(vi) Keep a record of all benefits (both monetary and non-monetary) obtained by both the provider and user of the genetic resource.

Source: Karen Marshall

IN THE FIELD

Collecting samples

For this most crucial step, the following considerations are relevant:

- Almost all cells or tissues may be used for DNA analysis: blood, semen, hide, bone, tissue (e.g. ear tissue), plucked hair (only the root cells contain nuclei, but cut hairs can be used for analysis of mitochondrial DNA) and feathers.
- High-quality DNA is most easily obtained from samples of peripheral blood, organs or other tissues. Most convenient are blood samples collected in an anti-coagulant (EDTA or natrium citrate). A protocol for blood collection is provided in Appendix 3.
- Collect enough material for present and future studies (considering ABS agreements for future studies). This process may either include the collection of multiple samples per animal or of a sufficiently large sample that can be aliquoted into separate portions in the laboratory. For most PCR-based applications, including high-density SNP arrays and whole genome sequencing, 5ml of blood is adequate. Note that poultry species have enucleated erythrocytes and, therefore, much less blood (~1 ml) is required.
- Blood samples can be transported at ambient temperatures, but in tropical regions samples should be processed within 36 hours.
- For longer storage, samples can be placed in a room-temperature preservative such as Queen's buffer (0.01 M Tris/HCl, 0.01 M NaCl, 0.01 M EDTA and 1 percent n-laurosylsarcosine, pH 8.0; Seutin *et al*., 1991).
- Tissue samples of 1 cm squared should be minced to 1 mm squared pieces and placed in Queen's buffer or 70 percent ethanol. Air-drying of ethanol-treated samples allows long-term storage and easy transport of samples. Alternatively, pieces of tissue may be dehydrated directly by placing them in vials on crystals of silica gel.
- Hair samples should be desiccated as soon as possible and stored dry.
- FTA® cards can be used for collection of genetic material with DNA to be amplified by PCR, but special protocols are required for some species to obtain double-stranded DNA. Moreover, the single-stranded DNA obtained with standard isolation protocols is not suitable for all other applications.
- Samples that are to be used for cloning and Southern blotting require double-stranded DNA of high molecular weight.
- Labelling of samples should be unambiguous and permanent. Whenever possible the official identification number of the animal should be recorded.
- Bank it: store all samples and document all relevant information unambiguously in such a way that it can be retrieved and understood, even by persons not involved in the sampling.

Collecting data

Recording the following information for each sample is essential:

- animal identification, preferably a herdbook registration number;
- date;
- location, preferably based on global positioning system (GPS) coordinates;
- name of collector;
- breed;
- sex of animal; and
- type of sample (blood, hair, etc.).

Collection of the following information is strongly recommended and will be necessary for some types of analyses (e.g. phenotypes will be necessary for studies to determine genomic regions having influence on traits).

- Age, or date of birth;
- any relevant phenotype(s);
- basic pedigree information (i.e. parents, if known);
- digital photograph(s) of animal, showing any interesting morphological features and including a measuring stick to evaluate body measurements;
- size of herd; and
- notes about any recent change in geographic location of the animal.

An example of a sampling form is provided in Appendix 4.

In addition to the information recorded for each sample, for each breed a form like the example in Appendix 5 should be compiled on the basis of the information available. This form addresses breed origins, farming practices, basic production information, and features of the breed such as productivity, disease resistance or adaptation to local conditions. For further advice on collecting data on breeds' phenotypes and production environments, consult the *FAO Guidelines on Phenotypic Characterization of Animal Genetic Resources* (FAO, 2012).

IN THE LABORATORY

Technological advancement in genomics has been accompanied by the emergence of private companies providing custom genotyping or sequencing services. For many genomic applications, all laboratory procedures subsequent to sample collection can be outsourced, such that a traditional "wet lab" is not strictly needed. Outsourcing saves the investment in expensive and specialized equipment, which tends to become quickly obsolete. Nevertheless, several issues need to be considered even if the experiments are outsourced.

Extracting DNA

Extraction of DNA is a key step for in-house genotyping and for analyses outsourced to companies that require DNA instead of raw samples. Extracted DNA is also subject to less-stringent sanitary regulations than are biological samples when shipped across borders.

Several reliable protocols for DNA extraction are available. Older protocols are based on Proteinase K/SDS lysis of cells, organic extraction and alcohol precipitation. Salt precipitation avoids organic solvents, but the long-term stability of the DNA samples is problematic. Convenient commercial kits based on the specific binding of DNA to resins are available for several kinds of tissues and usually perform well. Whenever a new DNA extraction procedure is used, a practice run on test samples should be performed to gain experience and ensure that it works properly before being applied to the field samples.

DNA assay

Section 3 discusses protocols for genotyping and sequencing in detail. In general, these protocols are highly automated and straightforward, but the indicated procedures need to be followed precisely. The following are general suggestions that will be applicable in many situations:

- Genotyping methods that are sensitive to variation between laboratories, such as microsatellite typing, should include at least one reference sample shared by the laboratories involved.
- The inclusion of a duplicate sample should be considered to permit evaluation of accuracy.
- The quality of the resulting data must be checked critically, even for outsourced analyses.

DATA ANALYSIS

Sections 3 and 4 address specific issues of data analysis according to the type of data (i.e. individual loci or sequence data) and study objective. For the many scientific objectives, software is freely available and greatly simplifies the tasks associated with data analysis (see Appendix 6). The easy access to software does not, however, relieve the researcher from the responsibility for understanding the genetic and statistical principles underlying the analyses. Each scientist must understand whether and why a specific approach is or is not suitable for his or her data and objective. Instead of copy-pasting computer output into manuscripts, the researcher should evaluate the results critically. For many analyses, multiple options for software are available and repeating an analysis using other options or another program can be informative. When software requires the input of parameters, the effect of parameter changes should be evaluated, especially if true values of the parameters are unknown. Analysis of simulated data is another option for validation of an algorithms. Results of the data analyses should be interpreted in the context of existing biological, genetic and historical knowledge.

PUBLISH IT

Let the world know

Publish your findings in a scientific journal. Explain how the results are relevant in a greater context for the management of AnGR. Open-access journals are recommended because of their wide diffusion and free accessibility.

Share the credit

Properly acknowledge contributors of samples and/or data. In general, include all colleagues who supplied samples and/or data in the preparation of any scientific publications so that they are fully aware of the key results and can be recognized as co-authors.

Share the data

After publication, deposit your data in a public database and/or comply with requests to make datasets available as already requested by many journals.

For each of the three tasks listed above, it’s important to ensure that all activities are addressed in the MTA.

TRANSLATE THE RESULTS

Genomic characterization studies are excellent capacity building activities and contribute to academic knowledge, but they should be more used to improve the management of the AnGR involved. The country’s National Coordinator for the Management of Animal Genetic Resources should be informed about the study and should be provided with data that can be uploaded to the Domestic Animal Diversity Information System (DAD-IS; FAO, 2021). Researchers should communicate their results and conclusions to the main stakeholders of the breeds evaluated and, if request, provide recommendations. The following phenomena can be detected through genomic analyses and may be relevant for breed management:

- **Original diversity.** Populationslocated in or having originated near their species' centre of domestication are presumed to have the greatest genetic diversity. The level of diversity can be measured by the amount of heterozygosity, the number of unique alleles and haplotypes, or the nucleotide diversity.
- **Unique origin or history.** In many instances, populations from the same region have relatively high genetic similarity, even if they have distinct phenotypic characteristics, suggesting that they share a common origin or history. Breeds with a distinct origin are assumed to carry unique alleles and thus be of greater value for conservation.
- **Pedigree errors.** Genetic tools now allow for errors in recording of parentage to be identified and remedied; this is especially useful in the case of unknown paternity.
- **Crossbreeding or hybridization.** Breeds that have been subject to recent crossbreeding are often considered to have decreased conservation value relative to "pure" breeds because of lower distinctiveness, even if they still maintain a substantial portion of unique genetic diversity. This is especially the case when the crossbreeding or hybridization involved one or more international transboundary breeds. On the other hand, systematic selection may have created a stable and valuable combination of traits from two breeds, typically a high productivity from one breed and strong adaptation to a tropical climate from another breed. Well-known examples are the Girolando dairy cow and the Dorper sheep.
- **Consanguinity.** Conservation activities should avoid inclusion of related animals as much as possible. For example, genebanks should avoid selecting related donors and *in vivo* programmes should select sires that are as unrelated as possible. Genomics can be used to determine the genetic relatedness of potential donor animals.
- **Inbreeding depression.** High homozygosity within a breed often leads to decreased fitness and survival. Genomics can be used to diagnose excessive homozygosity and plan matings to decrease the incidence of inbreeding depression. In the worst-case scenario, strategic crossing with another breed may be required to introduce new genetic variation and genomics may be used to identify the most suitable breed.
- **Distinctive phenotypes.** Breeds are often defined by one or more distinctive genetic traits, which may justify their conservation.
- **Genetic defects.** Many genetic defects are determined by mutations of one or a few genes. Identification of the genomic locations of these genes can facilitate elimination of the defects and determination of the precise mutation may allow for the development of control therapies. Deficit or absence of a homozygous genotype for a given locus may indicate lethal recessives. With the novel CRISPR/Cas technology gene defects may in the future become corretable and the same technology can also introduce mutations that confer desired traits (Menchaca, *et al.*, 2020), although genetic modification is not yet permitted in many countries.

Genomic information will rarely be the only tool available for the management of a breed, but can often complement factors such as productivity, breed demographics and cultural significance in making decisions on AnGR management.

INTERNATIONAL COORDINATION

Results from genomic characterization studies have potentially high value only for national management of the breed from which the samples originated, but also for other breeds and countries. Thus, sharing of the resources associated with such studies is strongly recommended whenever possible under the terms of use agreements and contracts and the Nagoya Protocol. As already mentioned, the data from such studies should be made available, preferably in a public repository such as DRYAD (2021) or Figshare (2021). DNA samples should be stored in a biobank for use in follow-up studies, including international collaborations. Once again, the rules of the source country and conditions imposed by the original owners of the animals must always be respected.

REFERENCES

Ball, R.D. 2013. Designing a GWAS: power, sample size, and data structure. *Methods in Molecular Biology*, 1019: 37-98. http://doi.org/10.1007/978-1-62703-447-0_3

CBD. 2008. Report of the meeting of the group of legal and technical experts on concepts, terms, working definitions and sectoral approaches, UNEP/CBD/ABSWG/7/2. (available at https://www.cbd.int/doc/?meeting=ABSGTLE-01).

CBD. 2011. *Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the Convention on Biological Diversity.* Secretariat of the Convention on Biological Diversity. Montreal. (available at https://www.cbd.int/abs/doc/protocol/nagoya-protocol-en.pdf)

Dryad. 2021. DRYAD – *for your research data* [online]. Durham, NC. [Cited 13 March 2021]. https://datadryad.org/

FAO. 2007. *Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration.* Rome (available at http://www.fao.org/docrep/010/a1404e/a1404e00.htm).

FAO. 2009. *Preparation of national strategies and action plans for animal genetic resources*. FAO Animal Production and Health Guidelines. No. 2. Rome. (available at http://www.fao.org/3/i0770e/i0770e00.htm)

FAO. 2011a. *Developing the institutional framework for the management of animal genetic resources*. FAO Animal Production and Health Guidelines. No. 6. Rome. (available at http://www.fao.org/3/ba0054e/ba0054e00.pdf)

FAO. 2011b. *Molecular genetic characterization of animal genetic resources*. FAO Animal Production and Health Guidelines. No. 9. Rome. (available at http://www.fao.org/3/i2413e/i2413e00.htm)

FAO. 2012. *Phenotypic characterization of animal genetic resources*. FAO Animal Production and Health Guidelines No. 11. Rome. (available at http://www.fao.org/3/i2686e/i2686e00.htm)

FAO. 2019. *ABS Elements: Elements to facilitate domestic implementation of access and benefit-sharing for different subsectors of genetic resources for food and agriculture – with explanatory notes*. FAO, Rome. (available at http://www.fao.org/3/ca5088en/ca5088en.pdf)

FAO. 2021a. *Domestic Animal Diversity Information System* [online]. Rome. [Cited 15 January 2021]. www.fao.org/dad-is

FAO. 2021b. *Synthesis progress report on the implementation of the Global Plan of Action for Animal Genetic Resources – 2020.* Rome (available at fao.org/3/cb4393en/cb4393en.pdf).

Figshare. 2021. *Figshare - credit for all your research* [online]. London [Cited 27 July 2021].

Groeneveld, L. F., Lenstra, J. A., Eding, H., Toro, M. A., Scherf, B., Pilling, D., Negrini, R., Finlay, E. K., Jianlin, H., Groeneveld, E., Weigend, S. & The GLOBALDIV Consortium. 2010. Genetic diversity in farm animals – a review. *Animal Genetics*, 41 Suppl 1: 6–31. https://doi.org/10.1111/j.1365-2052.2010.02038.x.

Hein, J. Schierup, M. H. & Wiuf, C. 2003. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, Oxford.

Kim, K., Kwon, T., Dessie, T., Yoo, D.A., Mwai, O.A., Jang, J., Sung, S., Lee, S., Salim, B., Jung, J., Jeong, H., Tarekegn, G.M., Tijjani, A., Lim, D., Cho, S., Oh, S.J., Lee, H.K., Kim, J., Jeong, C., Kemp, S., Hanotte, O. & Kim, H. 2020. The mosaic genome of indigenous African cattle as a unique genetic resource for African pastoralism. *Nature Genetics*, 52: 1099–1110. https://doi.org/10.1038/s41588-020-0694-2

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q. & Simianer, H. 2015. Properties of different selection signature statistics and a new strategy for combining them. *Heredity*, 115: 426–436. https://doi.org/10.1038/hdy.2015.42.

Madilindi, M.A., Banga, C.B., Bhebhe, E., Sanarana, Y.P., Nxumalo, K.S., Taela, M.G., Magagula, B.S. & Mapholi. N.O. Genetic diversity and relationships among three Southern African Nguni cattle populations. *Tropical Animal Health and Production*, 52: 753-762. http://doi.org/10.1007/s11250-019-02066-y.

Menchaca, A., Dos Santos-Neto, P.C., Mulet, A.P. & Crispo, M. 2020. CRISPR in livestock: From editing to printing. *Theriogenology*, 150: 247-254. http://doi.org/10.1016/j.theriogenology.2020.01.063.

Oklahoma State University. (2015). Breeds of Livestock [online]. Stillwater [Cited 1 April 2021]. afs.okstate.edu/breeds/

Porter, V., Alderson, L., Hall, S.J.G. & Sponenberg, P. 2016. *Mason's World Encyclopedia of Livestock Breeds and Breeding.* CABI Publishing. Wallingford, United Kingdom.

Yadav, D.K., Arora, R., Jain, A. 2017. Classification and conservation priority of five Deccani sheep ecotypes of Maharashtra, India. *PLoS One*, 12: e0184691. http://doi.org/10.1371/journal.pone.0184691

SECTION 3

Genomic tools and methods

GENOMIC TOOLS AND METHODS

Since the appearance of the previous version of these guidelines in 2012, there has been a spectacular growth in the use of genomic technologies, which still is becoming more and more powerful and accessible. As an initial step whole genome sequencing (WGS) was performed on selected animals from each major livestock species. This sequencing established and initial reference genome, which then allowed the development of multilocus bead arrays of single nucleotide polymorphisms (SNP) for most domestic species. These arrays fulfilled the objectives originally envisaged for microsatellite genotyping and have remained extremely useful. The next development was a much wider adoption of WGS technologies. More animals were fully sequenced and reference genomes were considerably improved. Approaches such as resequencing and genotyping-by-sequencing (GBS) were developed and applied and now allow the collection of WGS datasets for tens or hundreds of individuals, yielding vast multilocus genotype datasets. All theseprocesses are self-reinforcing, since the availability of many high-density genotypes allows the elevation of low- to high density datasets by imputation. This section explains these various developments and approaches in more detail.

SNP GENOTYPING

Single nucleotide polymorphism

A SNP is a deoxyribonucleic acid (DNA) sequence variation that occurs by substitution of a nucleotide at a specific position in the genome. The SNP is the most common type of polymorphism and is estimated to occur with a frequency of one SNP per 1 kilobase (kb) in most mammalian genomes. SNPs have increasingly become the marker of choice and have largely replaced microsatellite markers to assess genetic diversity, structure and relationship among populations; and particularly to identify genomic regions associated to economic traits. As noted in Section 1, SNPs have several advantages over microsatellites, including: (i) stable inheritance; (ii) distribution throughout the genome at a much greater density; (iii) location in coding regions that can possibly alter protein function and phenotypic expression; (iv) location nearby or within quantitative trait loci (QTL) of interest; and (v) suitability for high throughput genotyping.

Candidate SNP genotyping methodologies

Discovery of a SNP at the candidate gene level is done by targeted resequencing of specific regions of the genome in a subset of unrelated individuals representing the species or the populations of interest. Several methodologies are available to genotype one or a few SNPs identified in different regions of candidate genes like exons, introns, promoters and untranslated regions. Most of these methods are based on conventional or real-time polymerase chain reaction (PCR) procedures. A few examples include: (i) traditional gel-based approaches, such as amplification refractory mutation system (ARMS; Little 2001), restricted fragment length polymorphism (RFLP; Jarcho 1994), single strand conformation polymorphism (SSCP; Dong and Zhou 2005), denaturing gradient gel electrophoresis (DGGE; Strathdee and Free 2013); and (ii) medium throughput approaches, such as competitive allele specific polymerase chain reaction (KASP assays; He *et al*., 2014), exonuclease detection (Taqman assay; Holland *et al.,* 1991), mass spectrometry-based primer extension detection (Sequenom) and high resolution melting (Bradić *et al*., 2011).

SNP microarrays

Advances in next generation sequencing technologies and rapidly decreasing costs of WGS have made SNP discovery at the genome level possible in most livestock species. The process includes WGS of individuals from a subset of populations of interest and selection of polymorphic loci for additional genotyping in a larger pool of animals, often using high throughput SNP microarray technology. The SNP microarrays consist of allele-specific oligonucleotide probes fixed on a solid support, such as glass slides or silicon quartz wafers. The target DNA (whole genome) samples are fragmented, and then the fragments are amplified and hybridized to these oligonucleotide probes for single nucleotide extension or ligation. The labelled SNPs are visualized by an immunohistochemistry assay to increase

the signal intensity and infer the genotypes accurately. Box 4 provides technical details about the SNP genotyping platforms offered by two main commercial suppliers.

BOX 4

Commercial options for SNP genotyping

Two major commercial SNP array platforms are currently available for genome-wide SNP typing of livestock species: (i) Illumina's Infinium iSelect Microarray or BeadChip, based on single nucleotide extension or allele-specific primer extension (Illumina Inc., 2016), and (ii) Affymetrix's GeneChip or AxiomArray, based on molecular inversion probe hybridization (Affymetrix Inc., 2020). The two platforms differ in their array fabrication and protocols. The BeadChip uses 50-mer (i.e. 50 nucleotides) allele-specific probes while the GeneChip uses 25-mer and/or 30-mer probes for hybridization with target DNA samples.

The BeadChip is produced by etching out wells on a silicon wafer through photolithography and plasma etching methodologies. The beads linked to "oligos" (oligonucleotides, short strands of DNA) containing 29-mer address tags and 50-mer SNP specific primers are randomly dispersed and assembled into these wells etched on the silicon wafer. The GeneChip is produced by *in situ* synthesis of oligonucleotide probes on a chemically protected array surface. The method relies on photolithographic masking of array surface, ultraviolet light directed deprotection of specific array spots and synthesis of probes through nucleotide coupling, with one nucleotide at a time per spot, for many spots simultaneously. The sequential application of specific lithographic masks determines the order of the probe sequence synthesized on the array surface. At least four photolithographic masks are required to synthesize each nucleotide of the probe sequence (one mask to allow addition of the required nucleotide and three other masks to prevent light from deprotecting the same spot while the other three nucleotides are being added).

SNP array design

The process of SNP array design starts from SNP discovery and includes validation of *de novo* variants, selection of marker loci, validation of draft SNPs and synthesis of arrays. To fully understand the process of SNP array design, one must have a grasp of some basic concepts for WGS. These concepts and associated terminology. These concepts are explained in Box 5.

BOX 5

Basic concepts of whole genome sequencing

The process of WGS involves three basic steps:

(i) *Fragmenting the genomic DNA strands into segments.* The technology is rapidly improving and the length of an individual DNA sequence that can be sequenced varies. Depending, on the technology used, the DNA in a given sample is either kept as entire as possible or is broken into fragments. A DNA sample contains many copies of an animal's genome, which are fragmented by physical means (e.g. sonication) and thus break in random locations. This results in many fragments, each of which is fully or partly overlapping with other fragments.

(ii) *Sequencing individual segments.* The sequence of an individual segment is called a "read". Because a given sample includes many copies of DNA, sequencing all fragments is neither possible nor necessary. However, sequencing more fragments helps to ensure that as much of the genome is sequenced as possible and results in a greater accuracy of the final sequence. Two terms that characterize the quality and quantity of sequencing data are "coverage" and "depth". These terms are correlated and are often

used interchangeably, but also have distinct meanings. When expressed as a percentage, coverage refers to the proportion of genome that has been "read" (i.e. sequenced). Depth (or coverage depth) is essentially the average number of times that a random nucleotide in the genome had been included in one of the sequenced reads (by chance, some regions will have been sequenced more often, some less often). Depth is usually expressed as a "nx" and can be calculated by dividing the amount of data produced in the sequencing assay by the data size of a single genome sequence. For example, if a sequencing assay produced 30Gb of data for a single genomc of 3Gb, the depth would be 10x. Coverage is also used as direct synonym of depth. Increasing the depth of sequencing increases the accuracy but it also increases the cost. The cost and depth must therefore be balanced, and the appropriate depth depends on the sequencing objective.

(iii) *Converting the data for individual segments into a full genomic sequence.* The fact that individual segments are overlapping allows us to reconstruct the segments into full chromosomes and genomes. This process is accomplished by using sophisticated bioinformatic software and is based on "alignment" of fragments based on finding the commonalities between them and the reference genome for the species. When no reference genome exists, then an original genome must be "assembled", which is a more complex processing that usually relies on long read sequencing (see later in the section) and comparison of genomes with related species, when feasible.

SNP discovery. The ideal SNP discovery process for designing and developing a diversity array for a given livestock species involves WGS of a panel of unrelated individuals, preferably from diverse breeds across wide geographical locations, ensuring the detection of as many of the variants within the targeted species as possible. The low coverage sequencing of large numbers of individuals has proven to be more advantageous than deep sequencing of fewer individuals in improving the power of the SNP discovery process (Buerkle and Gompert 2013). SNPs identified from previous studies and databases (e.g., Hapmaps, BAC libraries, dbSNP) can also be utilized to increase the number of SNPs available for the design of the array (Matukumalli *et al.,* 2009).

Selection of SNPs. Errors in genome sequencing can lead to the detection of false SNPs. Hence *de novo* variants need to be validated based on quality parameters like: (i) minimum sequencing read depth; (ii) consensus base ratio; (iii) SNP quality score; and (iv) relative distance of the base from the 3′-end of the read, normalized by read length (Shen *et al.,* 2010). Generally, the SNP discovery process results in the identification of several million SNPs after initial validation. Among these, a subset of loci (~2-3 million) needs to be selected for drafting a marker panel (screening array) based on different criteria (Fan *et al.,* 2010; Kranis *et al.,* 2013):

- *in silico* design score for likelihood of success in the genotyping assay;
- type of polymorphism (i.e. transition versus transversion);
- minor allele frequency (MAF – frequency of the less-common allele for biallelic SNP);
- presence of nearby SNPs (e.g., exclusion of SNPs with nearby polymorphisms within 10-15 bases);
- linkage disequilibrium (LD) with other SNPs included in the panel;
- physical distribution of SNPs (e.g. equidistant spacing over the genome);
- polymorphism in multiple populations; and
- enrichment of specific regions of the genome (e.g. genomic regions potentially associated with QTL).

Validation of SNPs. The screening array with a draft panel of preselected SNPs must be validated to identify a subset of high-performance SNPs that show the potential for various downstream applications, such as diversity analysis, marker-trait association and gene mapping. During the process

of SNP validation, the following criteria are utilized to identify loci that are placed on routine genotyping arrays (Illumina Inc., 2016; Affymetrix Inc. 2020; Fan *et al.,* 2010; Kranis *et al.,* 2013):

- performance of SNPs in terms of high call rates (i.e. proportion of samples for which the genotype can be resolved);
- accuracy of genotyping (e.g. efficiency in clustering genotypes, to distinguish homozygous and heterozygous classes);
- informativeness of SNPs;
- association with traits of interest;
- tagging other variants based on LD;
- imputation of other variants in the genome;
- spacing and location with respect to known LD blocks (genomic regions with low diversity);
- functional significance of SNPs (e.g., exonic/intronic, synonymous/non-synonymous, coding DNA sequence/untranslated region).

Routine genotyping array and marker density. Arrays of different SNP densities may be used depending on the purpose of downstream applications and the genetic structure of the population investigated. Among the arrays routinely used in livestock species, the marker density ranges from low (<20 000 SNP), medium (50 000), medium high (150-200 000) to high (>500 000) SNP density. The cost per array increases with the number of SNPs, but the cost per marker decreases. The best-performing and most informative markers can be utilized to design smaller, cost-effective arrays for routine genotyping for a specific purpose. Low-density SNP arrays designed with markers having higher MAF and uniform spacing across the genome may be used to "impute" (see specific sub-section on this topic below) to a higher SNP density or even a complete sequence if a large set of reference animals are genotyped at higher density or whole-genome sequenced (Boichard *et al.,* 2012). However, with advancements in array fabrication and automated fluidic technologies, the cost of medium density arrays has significantly decreased, resulting in the phasing out of low-density arrays.

Ascertainment bias

When SNPs selected for designing and developing an array are discovered by sequencing only a few samples or from only a few selected breeds, an ascertainment bias (AB) may occur, affecting inferences about larger and distinct populations. Two major kinds of AB are associated with SNPs: (i) MAF bias; and (ii) sub-population bias. MAF bias occurs when the SNPs are preferentially selected based on intermediate frequencies and is nearly impossible to avoid with SNP selection. The MAF bias results in over representation of common polymorphic loci but under representation of low frequency SNPs. Consequently, estimates of population genetic parameters, allele frequency distribution and linkage disequilibrium can be biased (Albrechtsen *et al.,* 2010; Heslot *et al.,* 2013). Sub-population bias occurs when the SNP discovery panel is chosen from individuals belonging to a specific sub-population or a small group of breeds. This results in an overestimation of the variability present in that population but an underestimation of the variability in genetically dissimilar populations. Accordingly, this AB can inflate heterozygosity estimates in populations of breeds that are closely related to the breeds in the SNP discovery panel. The AB can increase or decrease the estimates of various genetic parameters as compared to the expected estimates from unbiased data, thus distorting information on population differentiation (Lachance and Tishkoff, 2013; McTavish and Hillis, 2015).

The phenomenon of AB is not unique to SNP markers and may also occur with morphological and microsatellite markers. The AB in microsatellites occurs when the most polymorphic markers are typically selected (Vowles and Amos 2006; Vali *et al.,* 2008). Such bias influences estimates of genetic variability of populations, but the influence of biased SNPs is much greater than that of microsatellites, by virtue of their large numbers and biallelic nature. One of the potential approaches to mitigate AB is to adopt SNP filtering strategies, such as LD pruning (see also later in the section). The

basic idea is to remove multicollinearity effects by removing SNPs that are highly correlated with other SNPs within a given genome window, thus reducing LD among the SNPs left after pruning. LD pruning has been effective in reducing AB when estimating genetic differentiation measures among populations including genetic distance, fixation index, inbreeding coefficient, kinship and principal components analysis (PCA) (Malomane *et al.*, 2018). Although LD-based pruning does not fully offset the AB, it does help to reduce its effects. Another potential approach to mitigate AB is to utilize ancestral SNPs (i.e. those polymorphic SNPs present in wild relatives of a given livestock species) while comparing highly divergent populations. The ancestral loci have relatively higher heterozygosity and are less likely to show population bias, but fine scale patterns of diversity among closely related breeds may be missed (Malomane *et al.*, 2018; Barbato *et al.*, 2020). If resources are available, the sequencing of a few samples and comparison of estimates of genetic variation from the sequence data versus the SNP data can help to determine the degree of AB.

How to choose your array?

Choosing a suitable array for genotyping depends on the purpose of the study. Although not exclusive, the following factors may be considered while choosing the array:

- SNP density (e.g., high, medium and low-density arrays);
- potential AB (e.g., genetic/geographic distance between discovery panel and the population to be investigated);
- tagging of specific genomic features (e.g. markers recommended by the International Society for Animal Genetics (ISAG) for parentage testing, copy number variants (CNV), markers for detection of recessive traits);
- SNPs in common with existing genotype data;
- cost effectiveness of genotyping (e.g., marker density *vs.* array cost); and
- performance of the array (genome coverage, genotype clustering).

Selecting a high-density array for genome wide association can help in capturing maximum possible variants that have potential influence on the trait (Tsai *et al*., 2015). In case of genetic evaluation (genomic selection) studies, where genotypes/sequences of several individuals from the reference population are available and population LD is generally high, choosing a low-density array followed by genotype imputation (see subsequent sub-section) may be cost-effective (Boichard *et al.*, 2012; Georges *et al.*, 2019). With the rapid growth of sequence/marker information over time, newly updated versions of arrays with special features keep evolving for many livestock species across different platforms (See Appendix 7). While choosing a new array for genotyping, checking the overlap of SNP with other available arrays can improve harmonization with existing data for comparative and meta-analysis across studies.

Merging SNP datasets

Unlike microsatellites, SNP markers facilitate merging and harmonizing datasets across different studies, generated in various laboratories and/or different platforms. The genome wide SNP genotypes can also be merged with the genotypes identified by genome re-sequencing data to generate information on specific loci of interest. Datasets are merged based on marker identification numbers and genomic positions, after alignment to a standard reference genome assembly of the species investigated. Care needs to be taken on the strand orientation of markers while merging different datasets. Open-source programs, such as PLINK (Chang *et al.*, 2015; Purcell 2010), are available to facilitate merging genotypes from different datasets and optimize resources available for genomic characterization studies in livestock.

GENOTYPING BY SEQUENCING

The principle

Although SNP arrays are arguably the most popular genotyping platform, genotyping-by-sequencing (GBS) can be a method of choice in certain settings, particularly in *de novo* applications (Scheben *et al.*, 2017), such as for species with no standard SNP chip. The GBS approach is a variant of resequencing that targets a small portion of the genome and in one step identifies SNPs, many of which may be novel, and yields the genotypes.

The principle of GBS is the targeting of a fraction of the genome in such a way that the same fraction is sequenced in all individuals, the so-called "reduced representation" of the genome. This can be accomplished, for instance, by hybrid-capture or size selection of fragments generated by restriction enzyme digestion. The fragments are then sequenced, sequence reads are mapped against a reference genome and alleles called from the reads (Figure 1). There is considerable variation in these steps, which ultimately control the number and quality of resulting markers. Fourteen different methods of GBS have been described (Scheben *et al.*, 2017) and new ones are continually being proposed (e.g., Rowan *et al.*, 2017).

Calling genotypes from alleles is reliable if fragments are sequenced many times. Obviously, the probability of correctly calling a heterozygote increases with the depth of sequencing (where "depth" refers to the number of times each fragment is sequenced, on average): 0.000 for x=1, 0.500 for x=2, 0.750 for x=3, 0.875 for x=4, 0.938 for x=5 and 0.998 for x=10. Generally, sequencing error is low, e.g. 0.01-0.001, but not zero, which makes genotype calling a bit more complex than schematically presented below (Figure 1). Available software tools account for sequencing errors (see Scheben *et al.*, 2017; Lou *et al.*, 2020; and the Software section).

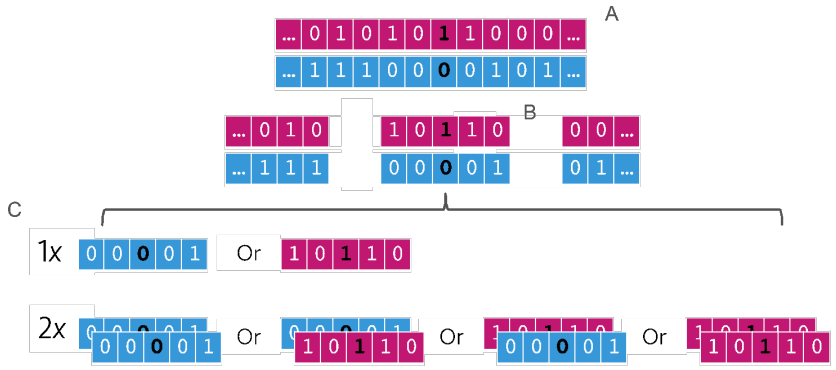

FIGURE 1

Characterizing variation at a fragment of DNA of a diploid individual (A) by focusing on a specific locus (B) and repeatedly characterizing its sequence (C)

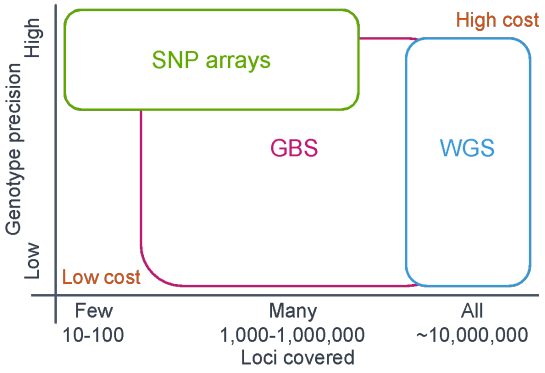

FIGURE 2

Schematic comparison of SNP array genotyping, genotyping-by-sequencing (GBS) and whole-genome sequencing (WGS) in terms of number of covered loci, precision of called genotypes and cost

Advantages and disadvantages

The GBS approach has several main advantages relative to SNP chips. First, GBS does not depend on a previously chosen set of markers or on the availability of whole-genome data and will reveal genetic variation within any livestock or wildlife population. Consequently, GBS may provide more relevant markers than a commercially available SNP array. Second, because only the partial genome is sequenced, GBS yields novel SNPs much faster than can be done by WGS and thusfor more immediate information on aspects such as MAFs and phylogeographic patterns within the panel of samples. Third, GBS completely avoids the AB (e.g., Schraiber and Akey, 2015; Simčič *et al*., 2015; Lou *et al*., 2020), except for GBS after hybrid capture. Fourth, users can tune GBS to their purpose and budget by choosing the number of sequenced fragments and the depth of sequencing (Figure 2) (see also Section 4). The GBS approach also provides an option for less common species that lack commercial SNP arrays.

There are, however, also disadvantages. First, GBS generally requires purified high-molecular weight DNA. Second, unique alleles in a population can be hard to distinguish from sequencing artefacts, whereas presence-absence variation due to deletions or insertions cannot be scored reliably at low sequencing depth. Third, because the number of markers depends on the sample and the output consists of allele calls instead of genotype calls, GBS requires additional bioinformatic attention in downstream analyses. Note that these considerations also apply to SNP calling on the basis of WGS assemblies. However, sequencing methods and tools for downstream analyses are continually improving.

Implementation and applications

Several commercial providers offer continually evolving custom GBS. Interested scientists are recommended to use an internet search engine to identify an appropriate service provider. The GBS approach has been adopted actively in several scientific communities (e.g., Buerkle and Gompert, 2013; Scheben *et al*., 2017; Lou *et al*., 2020). This adoption was fuelled by a lack of genomic resources or of SNP arrays in target species and/or unacceptably large AB in the available SNP arrays. GBS is less expensive than SNP bead arrays for a few individuals, but for high numbers of samples the streamlined SNP arrays become cheaper and offer the advantage of a more straightforward data analysis.

Several studies have leveraged GBS for population genetics studies. For example, Dodds *et al.* (2015) and Bilton *et al.* (2018) accounted for sequencing depth when estimating relatedness and LD, respectively. Dodds *et al.* (2019) and Whalen *et al.* (2019) applied GBS to parentage assignment. Interestingly, Lou *et al.* (2020) reviewed a number of standard population genetic analyses, most of which were based on GBS data without imputation and therefore had to develop specific approaches for standard analyses. An alternative approach is to infer genotypes from GBS allele calls from multiple individuals and loci and then follow standard analyses for genotypes obtained with SNP arrays.

Imputation of GBS datasets

The availability of GBS and other genotyping platforms has created the need for multiple types of imputation (Figure 3B-D). Imputation of genome-wide data is a large and a fast-evolving field and is in particularly suitable for populations with a known pedigree and a small number of parents genotyped by routine SNP array genotyping. For such populations, GBS strategies have been designed that enabled the imputation of whole-genome sequences in hundreds of thousands of individuals (Ros-Freixedes *et al.*, 2020a,b). Whalen *et al*. (2018, 2020) have developed methods, implemented in the programs ALPHAPEEL and ALPHAFAM, that quickly and accurately calls genotype, carries out phasing

and imputes whole-genome sequence data of any sequencing depth in pedigrees. GBS data can also be imputed without pedigrees by using software such as BEAGLE (Browning and Browning, 2016); STITCH (Davies *et al.*, 2016); and GLIMPSE (Rubinacci *et al.*, 2021). Whereas the imputation of SNP bead-array datasets to a higher density or to a WGS is at the expense of accuracy, imputation of GBS datasets improves the accuracy of genotypes that derived from low-depth datasets, which can have some inaccuracies due to missed alleles (Figure 3B-D).

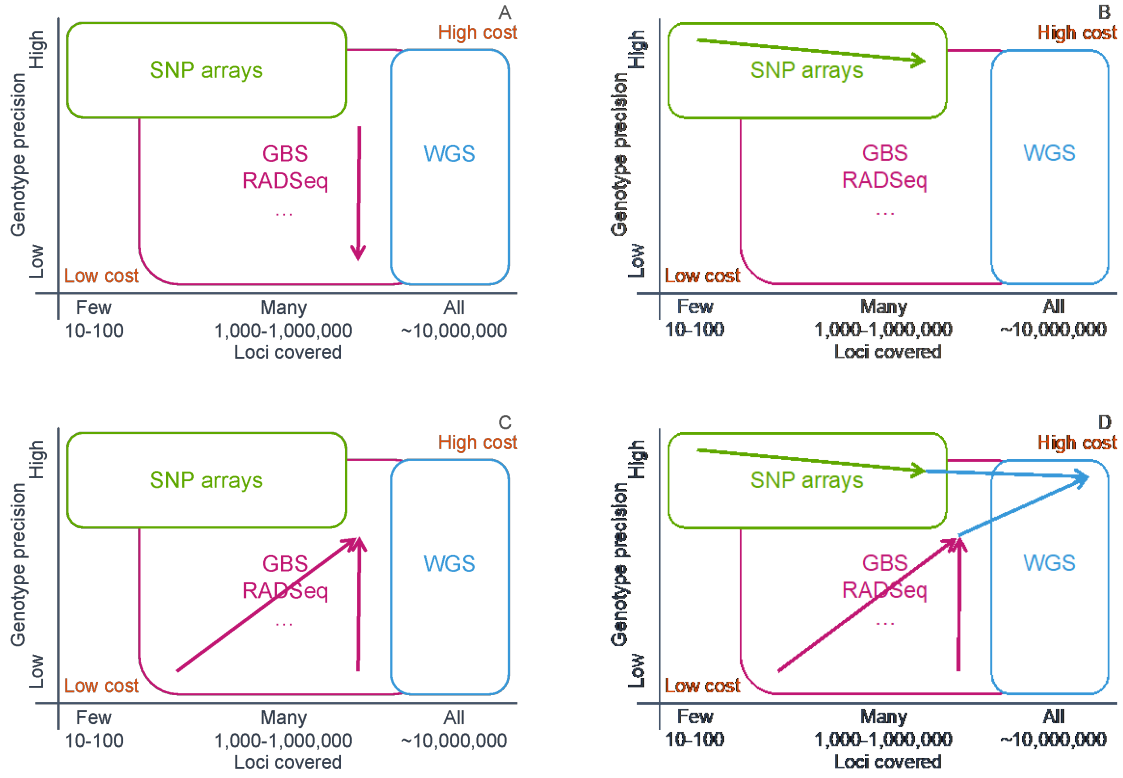

FIGURE 3

Strategies to increase volume of genomic data per unit of cost are indicated by arrows – to enable generating genomic data in the first place or genotyping more individuals at higher densities: A) reducing the depth in genotyping-by-sequencing; B) imputation from low-density to high-density SNP arrays; C) imputation from low-density and low-depth to high-density and high-depth in genotyping-by-sequencing; and D) imputation to whole-genome sequence.

PREPARING A WORKING MULTILOCUS DATASET

Genotyping and sampling errors can occur when performing genomic assays. The analysis of genetic data requires quality checks to identify such errors. Genotyping errors in SNP arrays can be due to insufficient DNA concentration, low quality of the DNA samples, sample contamination, poorly performing genotype probes, poor hybridization of DNA from very divergent populations or related species; and – in the case of WGS and GBS data – to low sequence coverage. Quality checks are performed at both the individual- and SNP-level and guide the "filtering" or the removal of the individuals and/or SNPs. The most common filtering approach attempts to preserve the number of individuals rather than the redundancy of SNPs. However, the optimal procedure and cut-off values for discarding of data should always be evaluated according to the dataset and the objectives of the study.

Although the manipulation and analysis of high-throughput genomic data, and interpretation of the results may require specialized bioinformatic competence (FAO, 2015), several well-documented and freeware bioinformatic tools are available, providing user-friendly tools and routines for the quality

checking procedures. Among many, these are the most popular and comprehensive: PLINK (Chang *et al.*, 2015; Purcell, 2010) and KING (Manichaikul *et al.*, 2010). Appendix 8 provides a step-by-step example of using PLINK (and R software) to perform quality checks and data editing.

Merging of datasets

As mentioned previously, a substantial advantage of SNPs relative to microsatellite markers is the option to combine data with existing datasets based on the same SNP panel or panels that have many SNPs in common. Since in livestock species most DNA arrays are provided by two major companies (i.e. Illumina and Affymetrix) with considerable across firm and backward compatibility of new versions, this is often the case. Merging of data is not trivial, but differences in SNP identification codes and genomic coordinates can be harmonized across the datasets to be merged by using the most updated assembly information for a given species. Chromosomes are composed of complementary strands of DNA and, as previously mentioned, opposing strands may have been genotyped between the datasets being merged. In such instances a "strand flip" must be undertaken to ensure proper mergiung. Such differences in strand orientation are identified by four alleles (A, C, G and T) recorded at a given locus, although a strict biallelic locus is generally expected. Such errors can be detected by PLINK during the merging operation and corrected by swapping A↔T and C↔G. However, different strand orientations in A/T and G/C SNPs cannot be detected, and it is advisable to remove all A/T and C/G SNPs if a strand flip must be carried out. It is also recommended that the datasets to be combined share at least one breed, to verify that in the combined dataset all animals from this breed cluster together and the diversity patterns are not confounded by systematic differences between the datasets that have been merged.

Mode of inheritance

With rare exceptions, mitochondrial DNA (mtDNA) markers are only inherited via the maternal lineages and markers on the male-specific part of the Y-chromosome (MSY) only via the paternal lineages. Neither mtDNA nor the MSY participate in meiotic crossing-over and both constitute a single non-recombining haplotype. As a consequence, they profoundly differ from autosomal markers in terms of mutation rate, evolutionary mechanisms, effect of genetic events and phylogeographic diversity patterns. X-chromosomal markers have a biparental mode of inheritance, but nonetheless differ in several properties form the autosomes:

- They have only one copy in males and thus a 25 percent lower effective population size Ne).
- Genes involved reproduction are overrepresented on the X-chromosomes (Vaiman, 2002).
- X-chromosomes are more involved in meiotic drives than are autosomes, the reasons for which are only partially understood (Hughes *et al.*, 2020).
- X-chromosomes also have a pattern of diversity that differs from the autosomal pattern (Wilson Sayres, 2018; da Fonseca *et al.*, 2019).

As a practical consequence for genetic diversity studies, mtDNA, MSY, X-chromosomes and autosomes should be analyzed separately. This happens automatically if LD pruning is applied, since this removes all three types of SNPs with a special mode of inheritance. For localization of genetic traits by GWAS (see Section 4), marker data from autosomes and X-chromosomes can be combined.

Missingness of data

The "missingness" of individual SNP and samples is defined as the percentage of missing genotypes for a given marker or individual, respectively, and is complementary to the "call-rate" (5 percent missingness equates to a 95 percent call-rate). High proportions of SNP missingness can be due to a probe design inadequacy for all or a part of the samples and in the latter case can lead to a bias in estimates of allele frequencies. A high proportion of missing genotypes for a sample indicates either a poor DNA quality, deviating species origin (e.g. *Bos indicus* rather than *Bos taurus*), or low sequence coverage for WGS- and GBS-based datasets.

For DNA array datasets, SNP-level filtering is typically performed first, followed by sample-level filtering, both allowing 5 percent missingness. A more refined approach can reduce the amount of data lost due to filtering by following a two-step approach. In the first step, a more relaxed threshold (such as 20 percent) to remove SNPs and samples with extremely high levels of missingness, followed by an additional filtering step with a more stringent threshold (e.g. 5 percent). Note that the SNP filtering step removes Y-chromosomal SNPs unless males account for at least 95 percent of the samples.

Sex discrepancy

When the sex of the animal providing the samples is known, discrepancies between the recorded and inferred sex may identify mislabelling of samples. Such an occurrence may indicate poor sample management and should be followed by a thorough check of the dataset and available metadata. The gender is derived by PLINK from the classical inbreeding coefficient [F = (expected homozygosity-observed heterozygosity)/expected homozygosity] for X-chromosomal SNPs; F >0.8 and <0.2 are often adequate cut-off values for males and females, respectively. Intermediate values may indicate either a mixed sample origin or, for male samples, the presence of SNPs from the pseudoautosomal region (PAR – region of commonality between sex chromosomes). The PAR can be localized by comparing the frequency of heterozygotes across the X-chromosome in males and females and can be relabelled as a separate chromosome not to be used for sex assessment. It is also essential to perform the sex-check assessments in PLINK on sets of loci in approximate linkage equilibrium (see section LD below). Furthermore, even if the sex inferred from analysis agrees with the sex recorded in the data, the presence of Y-chromosomal genotypes in a sample labelled as female is another clear warning of a problem that demands attention.

Minor allele frequency

As explained previously, the MAF for a biallelic SNP is the frequency of the least-commonly occurring allele at a given locus, hence MAF must be <0.5. The SNPs with a MAF of a few percentage points or less can represent true low-frequency variants but also be the result of genotyping errors. For this reason, it is common to retain only SNPs with MAF ≥0.05 for many analyses. Alternatively, with large data sets, the threshold MAF ≥10/N, based on the consideration that a larger sample size (N) reduces the likelihood that a small MAF has resulted from genotyping errors. However, some analyses (such as GWAS – see Section 4) are highly affected by low-frequency variants and thus more stringent thresholds are often applied (e.g. requiring MAF ≥0.05 even for N ≤ 10,000).

The MAF is particularly affected by sampling bias and AB. If a dataset has a large degree of homozygosity (hence, a small MAF on average) a larger sample size is necessary to obtain enough SNPs passing the MAF filtering because of the low overall frequency of heterozygotes. The effect of AB is proportional to the degree of divergence between the discovery and the study populations (Helyar *et al.*, 2011) and leads to an underestimation of the genetic diversity of the study population. This applies for instance to zebu cattle analysed by the popular arrays designed mainly for taurine cattle (Barbato *et al.*, 2020). As discussed above, the AB can be mitigated somewhat by filtering loci in high LD and by selection of SNPs that have a high MAF in the ancestral breeds or in the wild ancestors.

Allelic frequencies are used to calculate expected heterozygosity (*He*) and genetic distances between breeds. For optimizing breed comparison analyses, one must keep in mind that a low sample size for a given breed (<20 animals) can lead to underestimating the *He* and thus to inflating the genetic distances to other breeds.

Hardy-Weinberg equilibrium

The Hardy–Weinberg equilibrium (HWE) implies a direct relationship between allelic frequencies (p and q, with $p+q=1$ in the case of a biallelic SNP) and the genotypic frequencies (p^2 and q^2 for the homozygotes, and $2pq$ for the heterozygotes), which is supposed to remain constant over generations. The HWE is valid for a Wright-Fisher population of infinite size, with random mating, no selection, mutation, and migration. Significant departure from HWE for individual SNPs can be caused by population subdivision and substructure (e.g. Wahlund effect; Garnier-Géré and Chikhi, 2013),

recessive lethal mutations, selection, non-random mating, inbreeding, and genotyping errors (Chen *et al.*, 2017). For datasets of larger populations and datasets aimed at GWAS, it is a common practice to explain strong deviations from HWE by genotyping error and remove the corresponding SNPs. However, removal of such SNP must be done judicisously, as such deviations may be an indicator of selection or some other real phenomenon that is of interest in the analysis (e.g. Pausch *et al*., 2015).

The HWE can be tested by the so-called exact test, which provides a *P*-value for the significance of the HWE departure and is fairly robust to deviations from the theoretical assumptions. If the downstream analysis focuses on the identification of selection signatures (Section 4), it is recommended to use a permissive HWE filtering threshold (e.g. $P < 10^{-10}$) to avoid filtering out too many variants. In the case of GWAS applied to binary traits it is common practice to use more lenient thresholds in cases (e.g. $P < 10^{-10}$) than in controls ($P < 10^{-6}$), because deviations from HWE in cases can be due to true genetic association with the focal trait. When assessing quantitative traits, stricter thresholds can be used (e.g. 10^{-6}).

Non-random mating, which in its most extreme form leads to population subdivision (stratification into different sub-populations), is expected to cause HWE deviations across the whole genome, but such subdivision is commonly tested by comparing expected and observed heterozygosity (see below) and applying model-free multivariate statistics such as as PCA or multidimensional scaling (Section 4).

Heterozygosity

The observed heterozygosity of an individual (*Ho*) is the proportion of heterozygous genotypes, that is carrying two different alleles at a given locus, preferably considering only autosomal loci. Within populations, individual heterozygosity tends to be variable, but values for individuals with three standard deviations (SD) above the population average may indicate sample contamination or crossbreds, whereas values three SD below the average can be caused by inbreeding. Filtering H_o outliers is particularly important when the dataset is supposed to be genetically homogeneous, as in GWAS.

For checking the genetic constitution of a population, H_o averaged across the individuals of a population is an indicator of the diversity of the breed, although it should be borne in mind that it can be influenced by AB (see above) or be influenced by recent breeding decisions (e.g. sub-populations of inbred individuals within a relatively genetically diverse breed). It is further useful to compare H_o with H_e, the expected heterozygosity, calculated on the basis of the allele frequencies of the population. The H_o is generally lower than H_e as a consequence of sporadic non-random mating or sampling bias. Large differences between H_o and He indicate significant deviations from the Wright-Fisher ideal population, (e.g. $H_o > H_e$) can account for recent population admixture or outbreeding, while $H_o < H_e$ can indicate population subdivision or inbreeding. Both conditions may influence analyses that presume a homogeneous population. The extent of inbreeding/outbreeding can be quantified as $F_{IS} = (H_e-H_o)/H_e$, where F_{IS} represents the inbreeding coefficient of the population. F_{IS} is >0 when inbreeding is present and <0 when heterozygosity is greater than expected, such as with outbreeding or crossbreeding.

Linkage disequilibrium

SNPs are in LD if there is a non-random association of their alleles. This is to be expected for loci that are nearby on the same chromosome, for instance within or proximal to the same gene, so that recombination between them is a rare event. LD is also normal for the non-recombining mtDNA and the MSY. However, high LD of well separated genomic SNPs, that is not compatible with the expected recombination rate (about 1 percent per 10^6 base pairs (bp) per generation), may be caused by demographic processes such as admixture and genetic drift or by selective processes such as "hitchhiking" or background selection (McVean, 2007). Occasionally LD is observed between SNPs on different chromosomes (expected recombination rate 50 percent per generation), which is usually an artefact due to the sampling of a limited number of gametes or individuals, but may be a real phenomenon resulting from epistatic effects or unknown sources of population stratification.

A consequence of LD is that nearby markers are not independent. Therefore, it is common to implement LD-based pruning procedures to retain markers that are in approximate linkage equilibrium and are thus largely independent. This procedure is required when an analytical model does not explicitly take LD into consideration (see sections on MAF and relatedness). A common PLINK implementation of LD based pruning calculates correlation coefficients (r^2) within sliding windows across the genome and removes one of each pair of SNPs having LD greater than a given threshold. Typically, markers are selected in 2-Mbp windows that shift 0.2 Mbp forward (~10 percent of the window size) at each iteration while applying a LD threshold corresponding to $r^2 = 0.2$. This method could, however, induce additional bias if the LD removed is meaningful for a particular population. If the approximate average size of LD blocks is known for the species/breed of interests, another method to remove LD is to randomly select markers at a distance that is greater than the average level of LD specific to the breed. This method would then allow for the detection of extended LD that may be due to selective breeding or other biologically important characteristics of the population.

LD pruning can also be used for data size reduction, as it removes redundant SNPs by selecting only those SNP that are representative of the genetic haplotype blocks. Under this scenario it is possible to apply more conservative LD thresholds (e.g. $r^2 \geq 0.5$), which will remove the most highly correlated loci while preserving moderately correlated SNPs (e.g. due to low-frequency variants). Such an approach will retain the underlying genetic structure of the dataset while reducing the computational burden.

Note that LD-pruning removes mtDNA and Y-chromosomal SNPs. It should be carried out after removal of the SNPs with low call rates and bad-scoring or contaminated samples. LD pruning is beneficial for calculation of genetic distances, PCA or model-based ancestry analysis (e. g. as implemented in the popular software Admixture - Alexander *et al*., 2009), but not for haplotype-based studies, such as runs of homozygosity (ROH) and local ancestry analyses. An additional benefit of LD pruning is alleviating the AB in DNA array datasets (see above) (Malomane *et al.*, 2018).

Relatedness

Relatedness or kinship indicates the similarity of a pair of individuals due to common ancestry. Checking the pattern of kinship is essential for several reasons:

- Individuals may be found to be duplicated or be identical twins.
- The presence of highly related individuals introduces a bias towards a specific family in the dataset or reduces the statistical power of GWAS. It may be avoided during sampling by using pedigree information or a well-considered sampling scheme, but this is not always feasible.
- More in general, kinship may lead to population stratification, which confounds several analyses that assume a homogeneous population. For instance, it may easily cause false positives in single-gene association studies.
- Outlier animals with a much lower kinship with the other individuals than the population average should generally be excluded from analyses, as they arer likely to be crossbreds or mislabelled.
- If genetic distances between animals within the same breed are practically the same as between animals from different breeds, genetic clusters will no longer correspond to breeds. As a possible consequence, breed definitions may be reconsidered, especially if the lack of breed-level differentiation is due to uncontrolled gene flow between local populations kept under an extensive management regime.

A common method for the identification of related individuals is to compute the pairwise proportion of shared alleles inherited from a common ancestor (identity by descent, IBD) across all individuals. The magnitude of the resulting IBD score then reflects the relatedness of two individuals. This classic approach requires approximate linkage equilibrium (see the section on LD) and should be performed on autosomal markers. It relies strongly on the assumption of a homogeneous population structure. A matrix of IBD scores may be visualized by plotting the scores as a heat map.

A robust and efficient algorithm for inferring relatedness of individuals has been implemented in the software KING (Manichaikul *et al.*, 2010), which has been recently included in PLINK 2.0. The KING algorithm does not require LD among markers or a homogenous population structure. The KING relatedness coefficients are scaled to correspond to kinship coefficients (which are 0.5 × the relationships coefficients): monozygotic twins/duplicate samples have kinship=0.5; first-degree relations (parent-offspring, full siblings) have kinship=0.25; etc. As a cut-off value the geometric mean of the values corresponding to two degrees of kinship is used (e.g. a kinship coefficient >0.354, which is the geometric mean of 0.5 and 0.25) to identify monozygotic twins and duplicate samples.

Alternatively, a neighbour-joining phylogram of allele sharing distances between individuals can be generated to visualise identical samples, variable kinship, outliers, breed-level differentiation and, via the lengths of the terminal branches, the effect of genetic isolation on the diversity within breeds (Cardoso *et al.*, 2018). Such an analysis will also show the clustering of closely related breeds. For more distant breed relationships genetic distances between breeds instead of between individuals is more suitable.

Identical individuals should be removed from the dataset and, if possible, also those individuals with first or second-degree kinships (this is particularly important for GWAS), by removing from a closely related pair the individual with the lowest mean genotype call rate. Because one individual can be involved in several pairs, PLINK 2.0 has implemented an algorithm that maximizes sample size while pruning for relatedness. For calculation of relatedness on the basis of GBS datasets, see Dodds *et al.*, (2015).

Sample size

For diversity studies involving several breeds, availability of samples may be a limiting factor. Having a few samples of an interesting breed is still valuable, but its heterozygosity may be underestimated. Sample size does not have a large influence on the topology of phylogenetic networks or trees, but empirical evidence suggests that having <20 animals inflates the genetic distance to other breeds, as mentioned previously. Moreover, 20 animals may not be sufficient to detect low levels of breed admixture or for obtaining meaningful GWAS results (Ball, 2013).

On the other hand, more is not always better, especially when only some breeds have more data. When comparing multiple populations, breeds that are overrepresented may very well distort the results by dominating the principal components and the inference of clusters, respectively, especially if these breeds are also inbred. Hence, for these analyses it is recommended to reduce the number of samples from overrepresented breeds to obtain a dataset with a more balanced representation of breeds.

WHOLE GENOME SEQUENCING

Genotyping arrays are limited to assessing variants that have been preslected and are adaptable to array genotyping (generally, only SNP), such that the same sample may have to be run with multiple different assays to achieve desired goal(s) or to properly assess and account for AB. In contrast, WGS has the advantages that variants across the entire target genome are assessed and it is possible to identify all types of variants including SNP, indels (insertions and deletions), CNV, and structural variations (SV) like inversions or large deletions. The success of applying WGS for genome characterization depends on several factors including the type and depth of sequencing used and the analysis method(s). Here we present some background on the currently available approaches to using WGS for characterizing variation independent of the target species. Characteristics of the population (breed or species) such as Ne, or polyploid genomes, may affect the decision on the best approach to be applied for data generation and analysis.

De novo assembly

The most comprehensive evaluation of a genome comes from an independent, haplotype-resolved *de novo* whole genome assembly of the animal. This is the most expensive and technically demanding approach, although the cost has dropped by four orders of magnitude since the first livestock genomes were assembled 15 years ago, and the quality of the assemblies has dramatically increased. The basis

of this improvement is the advent of long-read sequencing (LRS) platforms, each of which has particular advantages and drawbacks (Amarasinghe *et al*., 2020; Logsdon *et al*., 2020). The LRS technology is developing rapidly, but the next paragraphs provides a brief overview of the state of the state of the art and most common technologies. Amarasinghe *et al*. (2020) provides a recent review.

The Pacific Biosciences (2015) platform, often referred to as PacBio, can operate in two "modes", one in which read length is maximized at the expense of read accuracy, and one with relatively shorter reads where read accuracy is maximized (the "HiFi" mode). The former produces contiguous assemblies at lower cost but with higher error rate, whereas the latter produces very high accuracy but somewhat less contiguous assemblies due to the shorter read length. The Oxford Nanopore Technologies (2008) sequencing platforms excel at generating extremely long reads that are very useful for assembling repetitive or duplicated regions of the genome but require relatively large amounts of input DNA and therefore substantial sample volume. Both LRS platforms are dependent on availability of suitable DNA template that consists mainly of very long fragments (10s to 100s of kb) of undamaged DNA. The characteristics of the two currently available LRS platforms differ so much that a single extraction method has yet to be formulated that is ideal for both platforms.

De novo assemblies of genomes based on short-read sequencing (SRS) (Bentley *et al*., 2008; Margulies *et al*., 2005; McKernan *et al*., 2009; Rothberg *et al*., 2011) are typically a less expensive option relative to LRS and can be suitable for population level analyses and transcriptomics. However, their accuracy (in terms of the lengths of contiguious genomic regions) is orders of magnitude lower than with LRS and all analysis of repetitive content and gene families must be done with consideration of their deficits. For these reasons, LRS has mostly replaced SRS for de novo assembly of genomes. There are approaches to improve SRS assemblies through large insert size libraries, or hybrid approaches that combine SRS and LRS. The most effective hybrid approach is linked-read technology (Wang *et al*., 2019; Zhang *et al*., 2017; Zheng *et al.,* 2016). In linked-reads, a library production technology is applied that adds barcodes to short fragments derived from a single, longer template molecule. Millions of these independent "micro-libraries" with different barcodes are then sequenced on standard short read platforms. Post-sequencing, the barcode information is used to partially reconstruct the longer DNA molecule and provides longer-range assembly continuity compared to standard, whole-genome SRS.

All platforms are limited in their ability to assemble genomic regions that undergo somatic cell rearrangements in blood cells (immune gene complexes like the T-cell receptor or immunoglobulin loci) if blood is the source of DNA.

Sequence data is first assembled into "contigs", which are stretches of DNA sequence without gaps. The SRS assemblies have N50 size (the length of the contig where the sum of all longer contigs is >50 percent of the total assembly size) in the range of 100 kb, while the contigs that can be created from LRS data can have N50 size over 70 megabases (Mb), or 700× as long. Generally, the genome sequence depth target is ≥100x for SRS and in the range of 50x for LRS to provide a quality assembly; however, if a haplotype-resolved assembly is desired then a greater depth is helpful. Also, if using the HiFi mode of sequencing, coverage can be as low as 20x although haplotype-resolution is steadily improved up to 40x or so. Contigs from LRS need to be "polished" to increase accuracy, with SRS data being useful for this step (e.g. Zimin and Salzberg, 2020). A second commonly used approach is "trio-binning, which utilizes divergence between two parental species or breeds and LRS of an F1 to create two, almost perfect haploid assemblies (Low *et al.*, 2020).

Y chromosomes remain challenging to assemble due to their highly repetitive nature. Even with LRS they often remain highly fragmented. Methods for the identification of these contigs include using the normalized ratio of female to male alignments (Hall *et al.*, 2013), and kmer based approaches (Carvalho *et al*., 2013; Rangavittal *et al*., 2019). Prior to polishing an assembly, it's important to make sure the mitochondrial genomc has been assembled. Failure to do so can result in over-polishing of nuclear insertions of mitochondrial sequence (NUMTs) which can lead to difficulties in identifying mitochondrial variants. There are methods to assemble mitochondrial genomes from both SRS and LRS (Al-Nakeeb *et al*., 2017; Dierckxsens *et al*., 2017; Formenti *et al*., 2020; Hahn *et al*., 2013; Meng *et al*., 2019). Polished contigs generated from LRS are sufficient to detect the majority of SNP, indels,

CNV, and SV in the genome by alignment to reference genome(s). Ideally, a breed- or population-specific reference or a pangenome representation would be used for such alignments, when available, to maximize mapping accuracy and variant detection. Alternatively, a full de novo assembly can be achieved by adding scaffolding data, either optical mapping (Hastie *et al.*, 2013; Nagarajan *et al.,* 2008) or chromatin conformation contact mapping (HiC) (Burton *et al*., 2013; Kaplan *et al*., 2013).

Resequencing

De novo assembly will nearly always yield the most detailed information about a single genome, however the cost and computational effort required is still impractical for analysis of multiple animals and populations in most circumstances. Resequencing, which involves SRS or LRS at much lower coverage than necessary for assembly and depends on mapping the reads to a reference genome to detect variants, is more amenable to application in larger populations. The goals of resequencing and the platform chosen will determine the depth of sequencing required. For discovery of SNP and indels, the high accuracy and low cost of SRS makes it the common choice. Tenfold mapped genome coverage is typically recommended as the minimum depth, and usually requires about 12-13x raw depth depending on platform. Extensive testing indicates that this is sufficient for accurate genotyping of ≥95 percent of SNP compared to very high coverage sequencing (≥30x coverage). At a low genome depth (<8x), some heterozygous genotypes will be scored as homozygous genotypes and while some sequencing errors will be recorded as heterozygous genotypes (Nielsen *et al.*, 2012), thus distorting the estimates of allele frequencies and genetic diversity. To address this problem, a genotype-likelihood based method has been developed to estimate nucleotide diversity for low sequencing coverage (Nielsen *et al.*, 2012). Resequencing by SRS can also accurately detect small (1-10 base) indels, but much higher coverage is needed to predict larger indels or CNVs, and some structural variants are very difficult to detect with only short read data. LRS can also produce quality SNP and indel calls but except for HiFi, sequencing will need to be done to a greater depth, around 20x. LRS excels when it comes to CNV and SV. With the same 10x coverage required for SNP and indels, HiFi data will yield many high quality CNV and SV calls. Regular LRS and SRS will benefit from a minimum of 20x coverage. LRS can call these larger variants with much greater frequency and accuracy than SRS and can capture larger variants than HiFi.

Prior to the advent of haplotype-resolved assemblies, it was common to choose female reference animals due to the drop in coverage associated with sequencing the heterologous portions of the sex chromosomes in males. When the sequenced animal is a male and the reference has no Y-chromosome to align to, identifying variants becomes more challenging. In this case, reads that do not map to the reference can be collected and aligned to an alternate dataset representing the Y chromosome. This can be the Y-chromosome of a closely related animal, or a collection of Y chromosome genes.

The success and accuracy of WGS for detection of variation is also highly dependent on the reference genome to which the sequencing reads are mapped. Currently, most studies use a single representative assembly agreed upon by the community of scientists working with the species in question, which provides a common basis for comparison among studies. However, many livestock species are now seeing high quality assemblies of multiple breeds emerging and studies comparing the success of using a breed-specific reference as opposed to "the reference" assembly have not been completed. Early results suggest that variant detection will be improved by breed-specific references, but at the cost of easy cross-study comparison (Crysnanto and Pausch, 2020). Recent efforts in cattle have focused on the creation of a "pangenome" reference including all genomic segments globally present among breeds, as a solution to this problem (Heaton *et al*., 2021).

Preparing a working multilocus dataset from genome resequencing

Regardless of the approach used (LRS or SRS), resequencing of an animal's genome initially yields a large quantity of individually sequenced DNA reads. These reads require substantial processing to finally provide a dataset of genetic variants positions on individual chromosomes. Figure 4 is a flow chart that illustrates the process to be taken to transform raw SRS data from a sample of genomic DNA into a dataset of genetic variants. For each step, the commonly used data formats and software modules are indicated. Other options for formats and software exist and may be used depending on

personal preference.

Many sequencing platforms output the raw sequence data for individual reads into a format called FASTQ. The FASTQ format is text based and includes a sequence identifier, the sequence, and a quality score. The raw sequence reads then undergo quality control and preprocessing. FASTQC (Andrews, 2010) is a software that undertakes a general overview to help identify problems in the sequencing process. QC3 (Guo *et al*., 2014) is a software with more features than FASTQC, such as identification of individual "good" and "bad" reads. PRINSEQ (Schmieder and Edwards, 2011) can perform both quality control and preprocessing, the next step in data preparation. Preprocessing often filtering and trimming of raw sequences. Individual sequences may be filtered based on quality based on parameters such as quality score, read length and content of guanine and cytosine nucleotides. Trimming removes sequence that corresponds to DNA adapters that may be incorporated during the sequencing process. Preprocessing improves the computing performance and accuracy of subsequent steps. In addition to PRINSEQ, Cutadapt (Martin, 2011), Trimmomatic (Bolger *et al*., 2014) and FastProNGS (Liu *et al*., 2019).

Following these steps, the individual reads are ready to be aligned to the reference genome. Several software programmes are available for this step, including BWA (Li & Durbin, 2009); Bowtie2 (Langmead & Salzberg, 2012); Novoalign (Novocraft Technologies, 2020) and GMAP (Wu & Watanabe, 2005). The alignment process yields more information for each read (e.g. chromosomal location) and thus requires a new output data format. The basic format is called SAM for "sequence alignment/map format", but given the large amount of data, the binary format of these data, "BAM" is more commonly used.

Finally, it is generally recommended to check and realign the sequence data due to indels and base quality. The GATK software (Van der Auwera & O'Connor, 2020) can be used for this task. This leads to the final step, variant calling. This step identifies the genomic locations of differences from the reference sequence. As this results in new information, another data format is used, VCF format, for "Variant Call Format". GATK can be used for variant calling, along with other software, including SAMtools (Li *et al*., 2009). The resulting data serves as the basis for further analyses.

As is the case for SNP assays, the resulting dataset should be subject to quality control. In particular, checks for the estimated level of relatedness of animals is important, both for the identification of putative duplicate samples and for possible removal of closely related animals that may distort the analysis. The quality control software described in Appendix 6 are generally applicable to sequence as well as SNP data.

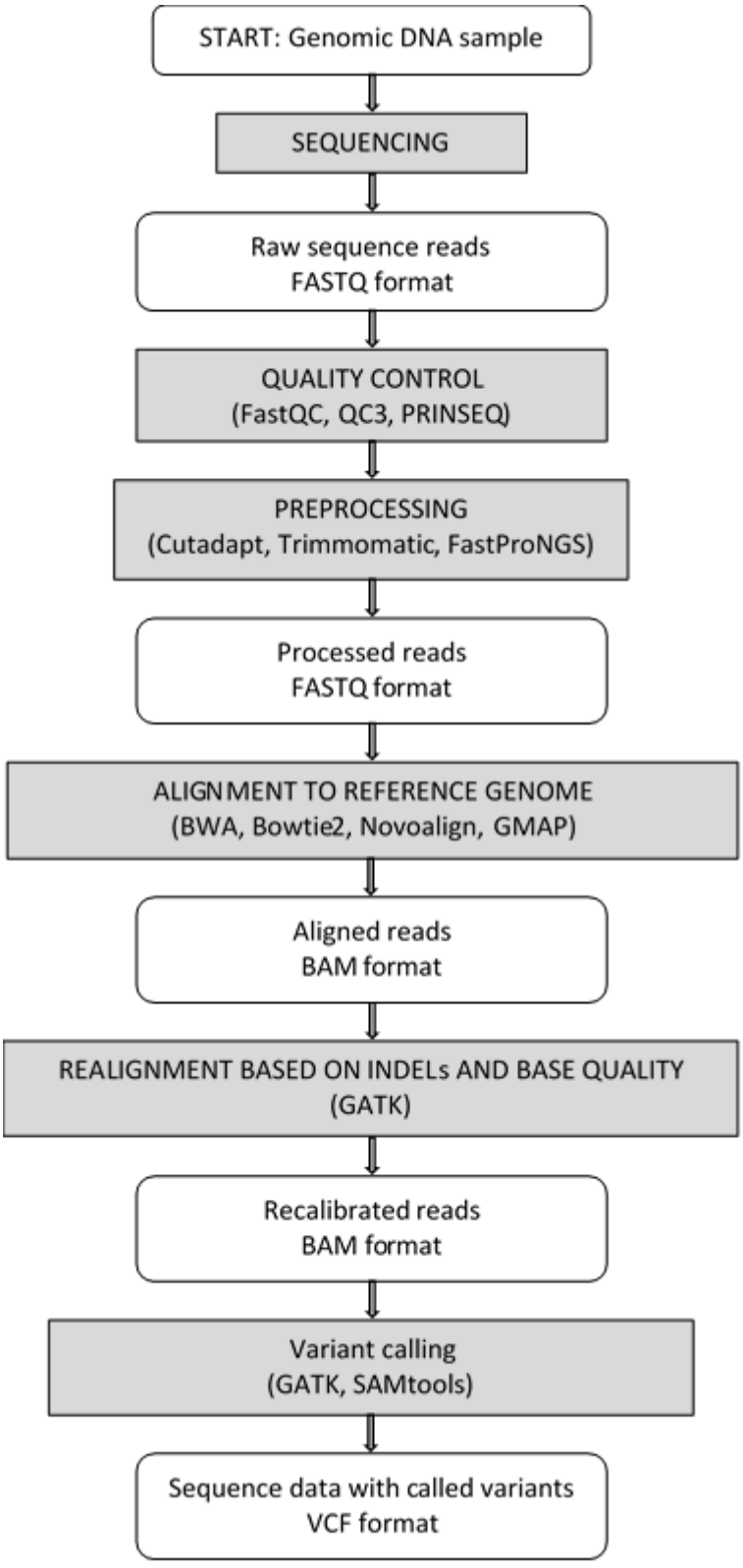

FIGURE 4

Steps to be taken in the processing of a sample of genomic DNA into a dataset of genetic variants (adapted from Carter & He, 2016). The white boxes indicate the material (DNA sample or data), whereas the gray boxes indicate the action undertaken. Common software for each action is shown in parentheses.

GENOTYPE IMPUTATION

Imputation: filling up missing values

Missing value imputation is the name given to the process of predicting incomplete values in a set of data. In many instances, missing values cannot be processed in an analysis and thus should be either removed or replaced by a guess beforehand. Although the first approach is simpler (e.g., the missing data could be handled by ignoring the missing information during computations), the second approach may help retain more statistical power in the analysis since it attempts to minimize data loss.

Imputation consists of taking advantage of correlations between variables in order to fill up the empty data. This approach involves the use of prediction methods adapted from statistics and machine learning, or even of deterministic approaches based on heuristics. As for any type of prediction, the imputation of missing values comes with uncertainty, and the accuracy of that prediction is dependent on a series of factors that are inherent to the data.

The number of errors that can be tolerated in an imputation procedure is dependent on the missingness rate. If a data set contains very few missing values, imputation with low accuracy could be performed with little impact on subsequent analyses, since the small number of affected values would be diluted over the rest of the data. This is frequently referred to as “minor imputation”. Conversely, if the number of missing values to be imputed is large, accuracy must be high in order to avoid the “garbage in, garbage out” effect - that is, poor quality input leading to faulty output.

If high missingness rates suggest poor quality data, it’s logical to question whether imputing the subsequent large numbers of missing values will be of value. However, in specific situations, large numbers of missing data are not caused by systematic errors but are rather deliberately planned during experimental design. It may be logical to “plan” to have missing data if a smaller subset of data can predict the missing data with a high accuracy but at a substantially decreased cost. Genomic data are a logical target for imputation, because DNA markers are inherited in linear sequences on chromosomes that remain intact from generation to generation (if recombination does not occur). Because of this, having data for the first and last nucleotide in a DNA sequence often allows accurate prediction of the entire DNA sequence between those nucleotides, based on full genome sequences of genetically related animals.

In a WGS experiment, for example, millions of variants are detected across samples. Despite drastic reductions in sequencing costs in the past decades, sequencing a whole genome (10x coverage depth) may still cost from US$ to more than US$ 1 000, depending on one’s location and the facilities and commercial providers available. In parallel, microarray technologies have matured substantially, making it possible to generate reliable data for a subset of these variants in a cost-effective manner. As discussed previously, the number of variants in these microarrays typically ranges from tens to hundreds of thousands. This is substantial data, but still a tiny fraction of the overall number of variants that could be captured by WGS. Therefore, one may design an experiment where a smaller number of animals have their whole genome sequenced, thus generating a reference panel of millions of variants. In turn, this reference panel is used as a guide to impute the bulk of the data that was genotyped using a microarray. The same rationale can be applied in imputing genotypes between microarrays, where a lower density panel (LDP) of markers is imputed to a higher density panel (HDP), or even to a full genomic sequence.

Imputation workflow

Genotype imputation is a straightforward procedure, albeit the mathematical and algorithmic mechanics behind it can be rather complex. First, the HDP and LDP are selected such that most – if not all – of the variants in the LDP are included in the HDP. Second, a set of animals is genotyped in the HDP in order to compose the reference panel. Third, animals to be imputed are genotyped in the LDP. Fourth, data quality control is applied to both the HDP and LDP sets. Finally, LDP data is imputed to the HDP using one of many methods available for genotype imputation.

Building a reference haplotype library

The selection of animals to be included in the reference panel must be optimized for improved imputation performance. If no genotypic data is available, one can use pedigree relationships in order to select key ancestors capturing a large proportion of the genetic variation in the population (Goddard and Hayes, 2009). Otherwise, reference animals could be selected based on the LDP data such that haplotype diversity is maximized. This is achieved by iteratively selecting individuals carrying large numbers of high frequency haplotypes (Butty *et al.,* 2019). Of note, if high imputation accuracy must be achieved for low frequency variants, selection of animals carrying rare haplotypes should also be considered (Bickhart *et al.*, 2016).

Imputation methods

Genotype imputation is divided into family-based and population-based methods. In the first, haplotypes from close relatives, typically identified via pedigree data, are used to impute the unobserved genotypes of LDP samples. In the second, pairs of individuals are assumed to share a common ancestor, such that LDP samples are interpreted as mosaics of haplotypes that are present in the HDP samples. Furthermore, there are methods that take advantage of both paradigms in order to achieve improved performance. Importantly, some methods require genotypes to be phased (i.e., that the parental haplotypes are separated) before imputation, while most phasing algorithms perform minor imputation on the fly.

The decision on imputation

Imputation of data can be attractive due to its potential to decrease costs, but accuracy will be sacrificed. When making the decision to imput data, the main factors to consider are the price difference between the low-density genotyping method and the high-density approach and the ramifications of possible errors. The accuracy of imputation will also rely heavily on the population to which it is applied. Accuracy is increased when a large amount of reference data is available, and when they are generally more closely related to the sample of animals to which imputation is to be applied. Accuracy of imputation is also increased by greater LD between the loci with known genotypes and the loci to be imputed. Increased mean LD of the breed will also increase imputation accuracy. For a small group of animals from one or more local breeds being genomically characterized for the first time, the conditions for imputation are will generally not be favourable. Imputation may however be quite attractive for further studies (such as GWAS) for a well-characterized population.

REFERENCES

Affymetrix Inc. 2020. *Mitigating Sequencing Errors, Monomorphs, and Poor Performing Markers during* de novo *SNP Selection for Genotyping Applications* [online]. Technical Note. [Cited 28 December 2020]. https://assets.thermofisher.com/TFS-Assets/LSG/brochures/mitigating_genotyping_appnote.pdf.

Albrechtsen, A., Nielsen, F. C. & Nielsen, R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, 27: 2534–2547. https://doi.org/10.1093/molbev/msq148.

Alexander, D. H., Novembre, J. & Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19: 1655–1664.https://doi.org/10.1101/gr.094052.109.

Al-Nakeeb, K., Petersen, T. N. & Sicheritz-Pontén, T. 2017. Norgal: Extraction and *de novo* assembly of mitochondrial DNA from whole-genome sequencing data. *BMC Bioinformatics*, 18: 510. https://doi.org/10.1186/s12859-017-1927-y.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. & Gouil, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21: 30. https://doi.org/10.1186/s13059-020-1935-5.

Andrews, S. 2010. FastQC: A quality control tool for high throughput sequence data [online]. Cambridge. [Cited 14 July 2021]. www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ball, R. D. 2013. Designing a GWAS: Power, sample size, and data structure. *Methods in Molecular Biology*, 1019: 37-98. doi: 10.1007/978-1-62703-447-0_3.

Barbato, M., Reichel, M. P., Passamonti, M., Low, W.Y., Colli, L., Tearle, R., Williams, J.L. & Ajmone-Marsan, P. 2020. A genetically unique Chinese cattle population shows evidence of common ancestry with wild species when analysed with a reduced ascertainment bias SNP panel. *PLoS ONE*, 15: e0231162. https://doi.org/10.1371/journal.pone.0231162.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., *et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456: 53–59. https://doi.org/10.1038/nature07517.

Bickhart, D. M., Hutchison, J. L., Null, D. J., VanRaden, P. M. & Cole, J. B. 2016. Reducing animal sequencing redundancy by preferentially selecting animals with low-frequency haplotypes. *Journal of Dairy Science*, 99: 5526–5534. https://doi.org/10.3168/jds.2015-10347.

Bilton, T. P., McEwan, J. C., Clarke, S. M., Brauning, R., van Stijn, T. C., Rowe, S. J. & Dodds, K. G. 2018. Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*, 209: 389–400. https://doi.org/10.1534/genetics.118.300831.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K. J., Hayes, B. J., Lawley, C. T., Sonstegard, T. S., Van Tassell, C. P., VanRaden, P. M., Viaud-Martinez, K. A. & Wiggans, G. R. 2012. Design of a bovine low-density SNP array optimized for imputation. *PLoS ONE* 7: e34130. doi: 10.1371/journal.pone.0034130

Bolger, A. M., Lohse, M., & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30: 2114-2120.

Bradić, M., Costa, J. & Chelo, I. M. 2011. Genotyping with Sequenom. *Methods in Molecular Biology*, 772: 193–210. doi: 10.1007/978-1-61779-228-1_11.

Browning, B. L. & Browning, S. R. 2016. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98: 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020.

Buerkle, A. C. & Gompert, Z. 2013. Population genomics based on low coverage sequencing: How low should we go?. *Molecular Ecology*, 22: 3028–3035. https://doi.org/10.1111/mec.12105.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O. & Shendure, J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31: 1119–1125. https://doi.org/10.1038/nbt.2727.

Butty, A. M., Sargolzaei, M., Miglior, F., Stothard, P., Schenkel, F. S., Gredler-Grandl, B. & Baes C. F. 2019. Optimizing selection of the reference population for genotype imputation from array to sequence variants. *Frontiers in Genetics*, 10: 510. doi: 10.3389/fgene.2019.00510.

Cardoso, T. F., Amills, M., Bertolini, F., Rothschild, M., Marras, G., Boink, G., Jordana, J., Capote, J., Carolan, S., Hallsson, J. H., Kantanen, J., Pons, A. & Lenstra, J. A. 2018. Patterns of homozygosity in insular and continental goat breeds. *Genetics Selection Evolution*, 50: 56. https://doi.org/10.1186/s12711-018-0425-7

Carter, T.C. & He, M.M. 2016. challenges of identifying clinically actionable genetic variants for precision medicine. *Journal of Healthcare Engineering*, 2016: 3617572. https://doi.org/10.1155/2016/3617572

Carvalho, A. B. & Clark, A. G. 2013. Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Research*, 23: 1894–1907. https://doi.org/10.1101/gr.156034.113.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4: 7. https://doi.org/10.1186/s13742-015-0047-8.

Chen, B., Cole, J. W. & Grond-Ginsbach, C. 2017. Departure from Hardy-Weinberg equilibrium and genotyping error. *Frontiers in Genetics*, 8: 167. https://doi.org/10.3389/fgene.2017.00167.

Crysnanto, D. & Pausch, H. 2020. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biology*, 21: 184. https://doi.org/10.1186/s13059-020-02105-0.

da Fonseca, R. R., Ureña, I., Afonso, S., Pires, A. E., Jørsboe, E., Chikhi, L. & Ginja, C. 2019. Consequences of breed formation on patterns of genomic diversity and differentiation: The case of highly diverse peripheral Iberian cattle. *BMC Genomics*, 20: 334. https://doi.org/10.1186/s12864-019-5685-2.

Davies, R., Flint, J., Myers, S. & Mott, R. 2016. Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, 48: 965–969. https://doi.org/10.1038/ng.3594.

Dierckxsens, N., Mardulyn, P. & Smits, G. 2017. NOVOPlasty: *De novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45: e18. https://doi.org/10.1093/nar/gkw955.

Dodds, K. G., McEwan, J. C., Brauning, R., Anderson, R. M., Stijn, T. C., Kristjánsson, T. & Clarke, S. M. 2015. Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, 16: 1047. https://doi.org/10.1186/s12864-015-2252-3.

Dodds, K. G., McEwan, J. C., Brauning, R., van Stijn, T. C., Rowe, S. J., McEwan, K. M. & Clarke, S. M. 2019. Exclusion and genomic relatedness methods for assignment of parentage using genotyping-by-sequencing data. *G3: Genes, Genomes, Genetics*, 9: 3239–3247. https://doi.org/10.1534/g3.119.400501.

Dong, Y. & Zhu, H. 2005. Single-strand conformational polymorphism analysis: Basic principles and routine practice. *Methods in Molecular Medicine*, 108: 149-157. doi: 10.1385/1-59259-850-1:149.

Fan, B., Du, Z.-Q., Gorbach, D. M. & Rothschild, M. F. 2010. Development and application of high-density SNP arrays in genomic studies of domestic animals. *Asian-Australasian Journal of Animal Science*, 23: 833–847. https://doi.org/10.5713/ajas.2010.r.03.

FAO. 2015. *The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture*. Rome, FAO. 562 pp. (also available at http://www.fao.org/3/i4787e/i4787e.pdf).

Formenti, G., Rhie, A., Balacco, J., Haase, B., Mountcastle, J., Fedrigo, O., Brown, S., *et al.* 2021. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biology,* 22: 120. https://doi.org/10.1186/s13059-021-02336-9.

Garnier-Géré, P. & Chikhi, L. 2013. Population subdivision, Hardy–Weinberg equilibrium and the Wahlund effect. In: *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Hoboken, United States of America.

Georges, M., Charlier, C. & Hayes, B. 2019. Harnessing genomic information for livestock improvement. *Nature Reviews Genetics*, 20: 135–156. https://doi.org/10.1038/s41576-018-0082-2.

Goddard, M. E. & Hayes, B. 2009. Genomic selection based on dense genotypes inferred from sparse genotypes. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics*, 18: 26–29. https://doi.org/10.1.1.599.3858.

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., Mchugh, C. P., Painter, I., Zheng, X., *et al.* 2012. Genetics and population analysis GWASTools: An R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, 28: 3329–3331. https://doi.org/10.1093/bioinformatics/bts610.

Gorjanc, G., Cleveland, M.A., Houston, R.D. & Hickey, J.M. 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genetics Selection Evolution*, 47(1): 12. https://doi.org/10.1186/s12711-015-0102-z

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R.C., Antolin, R. & Hickey, J.M. 2017. Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Science*, 57(3): 1404-1420. https://doi.org/10.2135/cropsci2016.08.0675

Guo, Y., Zhao, S., Sheng, Q., Ye, F., Li, J., Lehmann, B., Pietenpol, J., Samuels, D.C. & Shyr, Y. 2014. Multi-perspective quality control of Illumina exome sequencing data using QC3. *Genomics*, 103(5-6): 323-328. http://doi.org/10.1016/j.ygeno.2014.03.006.

Hahn, C., Bachmann, L. & Chevreux, B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. *Nucleic Acids Research*, 41: e129. https://doi.org/10.1093/nar/gkt371.

Hall, A. B., Qi, Y., Timoshevskiy, V., Sharakhova, M. V., Sharakhov, I. V. & Tu, Z. 2013. Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics*, 14: 273. https://doi.org/10.1186/1471-2164-14-273.

Hastie, A. R., Dong, L., Smith, A., Finklestein, J., Lam, E. T., Huo, N., Cao, H., Kwok, P. Y., Deal, K. R., Dvorak, J., Luo, M. C., Gu, Y. & Xiao, M. 2013. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. *PLoS ONE*, 8: e55864. https://doi.org/10.1371/journal.pone.0055864.

He, C., Holme, J. & Anthony, J. 2014. SNP genotyping: The KASP assay. *Methods in Molecular Biology*, 1145: 75–86. doi: 10.1007/978-1-4939-0446-4_7.

Heaton, M. P., Smith, T. P. L., Bickhart, D. M., Vander Ley, B. L., Kuehn, L. A., Oppenheimer, J., Shafer, W. R., *et al.* 2021. A reference genome assembly of Simmental cattle, *Bos taurus taurus*. *Journal of Heredity*, 112: 184–191. https://doi.org/10.1093/jhered/esab002.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R. & Nielsen, E. E. 2011. Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, 11(Suppl 1): 123–136. https://doi.org/10.1111/j.1755-0998.2010.02943.x.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.-L. & Sorrells, M. E. 2013. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLoS ONE*, 8: e74612. doi: 10.1371/journal.pone.0074612.

Holland, P. M., Abramson, R. D., Watson, R. & Gelfand, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, 88: 7276–7280. doi: 10.1073/pnas.88.16.7276.

Homburger, J. R., Neben, C. L., Mishne, G., Zhou, A. Y., Kathiresan, S. & Khera A. V. 2019. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Medicine*, 11: 74. https://doi.org/10.1186/s13073-019-0682-2.

Illumina Inc. 2016. *Infinium® iSelect® Custom Genotyping Assays. Guidelines for Using the DesignStudio™ Microarray Assay Designer Software to Create and Order Custom Arrays* [online]. Technical Note. [Cited 02 February 2021]. www.illumina.com/documents/products/technotes/technote_iselect_design.pdf.

Jarcho, J. 1994. Restriction fragment length polymorphism analysis. *Current Protocols in Human Genetics*, 1: 2.7.1-2.7.15. https://doi.org/10.1002/0471142905.hg0207s01.

Kaplan, N. & Dekker, J. 2013. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nature Biotechnology*, 31: 1143–1147. https://doi.org/10.1038/nbt.2768.

Karssen, L. C., van Duijn, C. M. & Aulchenko, Y. S. 2016. The GenABEL Project for statistical genomics. *F1000Research*, 5: 914. https://doi.org/10.12688/f1000research.8733.1.

Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., *et al.* 2013. Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics*, 14: 59. https://doi.org/10.1186/1471-2164-14-59.

Lachance, J. & Tishkoff, S. A. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*, 35: 780–786. doi: 10.1002/bies.201300014.

Langmead, B., & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4): 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25: 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16): 2078-2079. http://doi:org/10.1093/bioinformatics/btp352.

Li, J.H., Mazur, C.A., Berisa, T. & Pickrell, J.K. 2021. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Research*, http://doi.org/10.1101/gr.266486.120.

Little, S. 2001. Amplification-refractory mutation system (ARMS) analysis of point mutations. *Current Protocols in Human Genetics*, 9: Unit 9.8. http://doi.org/10.1002/0471142905.hg0908s07

Liu, T., Luo, C., Ma, J., Wang, Y., Shu, D., Su, G. & Qu, H. 2020. High-throughput sequencing with the preselection of markers is a good alternative to SNP chips for genomic prediction in broilers. *Frontiers in Genetics*, 11: 108. http://doi.org/10.3389/fgene.2020.00108

Liu, X., Yan, Z., Wu, C., Yang, Y., Li, X. & Zhang, G. 2019. FastProNGS: fast preprocessing of next-generation sequencing reads. *BMC Bioinformatics*, 20: 345. https://doi.org/10.1186/s12859-019-2936-9

Logsdon, G.A., Vollger, M.R. & Eichler, E.E. 2020. Long-Read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10): 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lou, R.N., Jacobs, A., Wilder, A. & Therkildsen, N.O. 2020. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Authorea*. http://doi.org/10.22541/au.160689616.68843086/v2

Low, W.Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D., Rosen, B., *et al.* 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, 11: 2071. https://doi.org/10.1038/s41467-020-15848-y

Malomane, D.K., Reimer, C., Weigend, S., Weigend, A., Sharifi, A.R. & Simianer, H. 2018. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics*, 19(1): 22. https://doi.org/10.1186/s12864-017-4416-9

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. & Chen, W.M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22): 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors." *Nature* 437(7057): 376–380. https://doi.org/10.1038/nature03959.

Martin A.R., Atkinson E.G., Chapman S.B., Stevenson A., Stroud R.E., Abebe T., Akena, D., *et al.* 2021. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *American Journal of Human Genetics,* 24: S0002-9297(21)00096-3. http://doi.org/10.1016/j.ajhg.2021.03.012

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17: 10-12.

Matukumalli, L.K., Lawley, C., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S. & Van Tassell, C.P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One*, 4(4): e5350. http://doi.org/journal.pone.0005350

McVean, G. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175(3): 1395-1406. http://doi.org/10.1534/genetics.106.062828.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R. *et al.* 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, 19(9): 1527–1541. https://doi.org/10.1101/gr.091868.109.

McTavish, E.J. & Hillis, D.M. 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC Genomics*, 16(1): 266. http://doi.org/10.1186/s12864-015-1469-5.

Meng, G., Li, Y., Yang, C. & Liu, S. 2019. MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Research*, 47(11): e63–e63. https://doi.org/10.1093/nar/gkz173.

Nagarajan, N., Read, T.D. & Pop, M. 2008. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, 24(10): 1229–1235. https://doi.org/10.1093/bioinformatics/btn102.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. 2012. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, 7: e37558. https://doi.org/10.1371/journal.pone.0037558.

Novocraft Techologies. 2020. Novoalign V4.03.01 [online]. Selangor. [cited 14 July 2021]. www.novocraft.com/novoalign-v4-03-01/

Oxford Nanopore Technologies. 2008. Products & Services. In: *NANOPORE* [online]. Oxford Science Park, United Kingdom. [Cited 8 April 2021]. https://nanoporetech.com/products

Pacific Biosciences. 2015. PACBIO [online]. Menlo Park, CA. [Cited 1 April 2021]. https://www.pacb.com

Palaiokostas, C., Clarke, S.M., Jeuthe, H., Brauning, R., Bilton, T.P., Dodds, K.G., McEwan, J.C. & De Koning, D-J. 2020. Application of low coverage genotyping by sequencing in selectively bred Arctic Charr (*Salvelinus alpinus*). *G3: Genes, Genomes, Genetics*, 10(6): 2069–2078. https://doi.org/10.1534/g3.120.401295

Pausch, H., Schwarzenbacher, H., Burgstaller, J., Flisikowski, K., Wurmser, C., Jansen, S., Jung, S., Schnieke, A., Wittek, T. & Fries, R. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. *BMC Genomics*, 16(1): 312. https://doi.org/10.1186/s12864-015-1483-7

Purcell, S. 2010. *PLINK (1.07) Documentation* [online]. [Cited 28 December 2020]. https://zzz.bwh.harvard.edu/plink/dist/plink-doc-1.07.pdf

Rangavittal, S., Stopa, N., Tomaszkiewicz, M., Sahlin, K., Makova, K.D. & Medvedev, P. 2019. DiscoverY: A classifier for identifying Y chromosome sequences in male assemblies. *BMC Genomics*, 20(1): 641. https://doi.org/10.1186/s12864-019-5996-3.

Ros-Freixedes, R., Whalen, A., Gorjanc, G., Mileham, A.J. & Hickey, J.M. 2020a. Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. *Genetics Selection Evolution*, 52(1): 18. https://doi.org/10.1186/s12711-020-00537-7

Ros-Freixedes, R., Whalen, A., Chen, C.Y., Gorjanc, G., Herring, W.O., Mileham, A.J. & Hickey, J.M. 2020b. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genetics Selection Evolution*, 52(1): 17. https://doi.org/10.1186/s12711-020-00536-8

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., *et al.* 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356): (July): 348–352. https://doi.org/10.1038/nature10242.

Rowan, B.A., Seymour, D.K., Chae, E., Lundberg, D.S. & Weigel, D. 2017. Methods for genotyping-by-sequencing. *Methods in Molecular Biology*, 1492: 221-242. https://doi.org/10.1007/978-1-4939-6442-0_16

Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J. & Delaneau, O. 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1): 120–126. https://doi.org/10.1038/s41588-020-00756-0

Scheben, A., Batley, J. & Edwards, D. 2017. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnology Journal*, 15(2): 149-161. https://doi.org/10.1111/pbi.12645

Schmieder, R. & Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27: 863-864.

Schraiber, J. & Akey, J. 2015. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12): 727–740. https://doi.org/10.1038/nrg4005

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., Liu, Y., Weinstock, G. M., Wheeler, D. A., Gibbs, R. A. & Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research*, 20(2): 273-280. https://doi.org/10.1101/gr.096388.109

Simčič, M., Smetko, A., Sölkner, J., Seichter, D., Gorjanc, G., Kompan, D. & Medugorac, I. 2015. Recovery of native genetic background in admixed populations using haplotypes, phenotypes, and pedigree information – using Cika Cattle as a case breed. *PLoS ONE*, 10(4): e0123253. https://doi.org/10.1371/journal.pone.0123253

Snelling, W.M., Hoff, J.L., Li, J.H., Kuehn, L.A., Keel, B.N., Lindholm-Perry, A.K. & Pickrell, J.K. 2020. Assessment of imputation from low-pass sequencing to predict merit of beef steers. *Genes*, 11(11): 1312. https://doi.org/10.3390/genes11111312

Strathdee, F. & Free, A. 2013. Denaturing gradient gel electrophoresis (DGGE). *Methods in Molecular Biology*, 1054: 145-157. http://doi.org/10.1007/978-1-62703-565-1_9

Tsai, H.Y., Hamilton, A., Tinch, A.E., Guy, D.R., Gharbi, K., Stear, M.J., Matika, O., Bishop, S.C. & Houston, R.D. 2015. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genomics*, 16: 969. https://doi.org/10.1186/s12864-015-2117-9

Vaiman, D. 2002. Fertility, sex determination, and the X chromosome. *Cytogenetic Genome Research*, 99: 224-228. http://doi.org/10.1159/000071597

Vali, U., Einarsson, A., Waits, L. & Ellegren, H. 2008. To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, 17(7): 3808–3817. http://doi.org/10.1111/j.1365-294X.2008.03876.x.

Vallejo, R.L., Leeds, T.D., Fragomeni, B.O., Gao, G., Hernandez, A.G., Misztal, I., Welch, T.J., Wiens, G.D. & Palti, Y. 2016. Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in Rainbow Trout: insights on genotyping methods and genomic prediction models. *Frontiers in Genetics*, 7: 96. http://doi.org/10.3389/fgene.2016.00096

Van der Auwera GA & O'Connor BD. 2020. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Vowles, E.J. & Amos, W. 2006. Quantifying ascertainment bias and species-specific length differences in human and Chimpanzee microsatellites using genome sequences. *Molecular Biology and Evolution*, 23(3): 598–607. http://doi.org/10.1093/molbev/msj065.

Wang, O., Chin, R., Cheng, X., Ka Yan Wu, M., Mao, Q., Tang, J., Sun, Y., *et al.* 2019. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and *de novo* assembly." *Genome Research*, 29(5): 798–808. https://doi.org/10.1101/gr.245126.118.

Whalen, A., Ros-Freixedes, R., Wilson, D.L., Gorjanc, G. & Hickey, J.M. 2018. Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. *Genetics Selection Evolution*, 50(1): 67. https://doi.org/10.1186/s12711-018-0438-2

Whalen, A., Gorjanc, G. & Hickey, J.M. 2019. Parentage assignment with genotyping-by-sequencing data. *Journal of Animal Breeding and Genetics*, 136(2): 102– 112. https://doi.org/10.1111/jbg.12370

Whalen, A., Gorjanc, G. & Hickey, J.M. 2020. AlphaFamImpute: high-accuracy imputation in full-sib families from genotype-by-sequencing data. *Bioinformatics*, 36(15): 4369–4371. https://doi.org/10.1093/bioinformatics/btaa499

Wilson Sayres, M.A. 2018. Genetic diversity on the sex chromosomes, *Genome Biology and Evolution*, 10: 1064–1078, https://doi.org/10.1093/gbe/evy039

Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9): 1859-75. http://doi.org/10.1093/bioinformatics/bti310.

Zhang, F., Christiansen, L., Thomas, J., Pokholok, D., Jackson, R., Morrell, N., Zhao, Y., *et al.* 2017. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nature Biotechnology*, 35(9): 852–857. https://doi.org/10.1038/nbt.3897.

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., *et al.* 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3): 303–311. https://doi.org/10.1038/nbt.3432.

Zimin, A.V. & Salzberg, S.L. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Computational Biology*, 16(6): e1007981. https://doi.org/10.1371/journal.pcbi.1007981

SECTION 4

Applications of genomics

APPLICATIONS OF GENOMICS

This section will describe how genetic and genomic information can be used to describe patterns (such as genetic diversity, differentiation, classification, and genetic clines) or to infer historical processes (expansions, contractions, admixture and gene flow) affecting animal genetic diversity. In the first case, one can quantify how genetic diversity changes per individual, location or environment or by domestication. The second group of approaches aims to identify, quantify and determine the date of ancient demographic events, possibly including direct or indirect selection acting on specific genomic regions. The principles and limits of several methods that have become available in the last couple of decades are noted. This section provides substantial the underlying genetic theory behind the most common analyses undertaken for genomic characterization and describes the methods used. The seminal research associated with each approach is cited. Appendix 9 provides a summary of the main steps and analyses that are frequently undertaken in genomic characterization and provides examples of the commonly-used software for each step.

Developments from the genomic era allow novel options to address previously unfathomable questions with increasingly sophisticated methods, including modelling. In some cases, the resulting patterns can only be interpreted by understanding the underlying models and processes. In addition, the analysis of patterns and processes should be integrated. For instance, most available approaches to identify selection typically make strong assumptions about the demographic history that generate "neutral" patterns, while methods reconstructing demographic events often ignore selection. Inferring simultaneously selection and demography, in order to explain spatial patterns, is one of the greatest challenges of evolutionary genomics.

ASSESSMENT OF GENOMIC VARIATION WITHIN-POPULATIONS

Measures of genetic diversity

Local well-adapted breeds are considered to be reservoirs of genetic diversity that are recognized to contribute to future traits of interest (Bruford *et al.*, 2015; Groeneveld *et al.*, 2010). Assessment of the within-breed component of genomic variation is essential for the effective management of breeds (Boettcher *et al.*, 2010; Caballero *et al.*, 2010; Ginja *et al.*, 2013; Lenstra *et al.*, 2012; Ruane, 2000; Toro and Caballero, 2005; Toro, Fernandez and Caballero, 2009) Assessment allows for monitoring, on one hand the effect of isolation on the genetic diversity and the ensuing risk of inbreeding depression, and on the other hand the effect of crossbreeding on the genetic constitution of breeds. In addition, a comparison of the diversity of domestic and wild populations, and the geographic pattern of within-bred diversity may reveal the serial population bottlenecks that occurred in the past in association with domestication and migration, respectively.

The genetic diversity can be quantified via the observed (H_o) and expected (H_e) heterozygosities, nucleotide diversity (π) and effective population size (N_e, see below). The H_o is based on the actual distribution of genotypes, whereas H_e is based on allele frequencies (Nei, 1973; 1978). Both of these measures are easily obtained from datasets of genetic markers. For nucleotide sequence data, nucleotide diversity (π) gives the average number of nucleotide differences per site between two DNA sequences chosen randomly from the studied population (Nei and Li, 1979) and is equivalent to H_o if all nucleotides are considered as loci. The parameter N_e represents the number of reproducing individuals given a number of conditions (see subsection on Effective population size).

Inbreeding and runs of homozygosity

Inbreeding results from the mating of related individuals and it is estimated by the probability (F) that two alleles at a locus are identical. Classically, inbreeding was calculated using pedigree records, but genomic data and H_o and H_e values allow for a more accurate quantification of inbreeding without using pedigree data (Curik, Ferenčaković and Sölkner, 2014; Druet and Gautier, 2017; Howrigan, Simonson and Keller, 2011; Purfield *et al.*, 2012). Inbreeding increases the homozygosity in individuals and thus decreases H_o, but it does not change H_e based on the allele frequencies, resulting in a heterozygote deficit and a departure from the Hardy-Weinberg equilibrium (HWE) of

homozygous and heterozygous genotypes. Thus, Wright's *F*-statistics based on comparing H_o and H_e can be used to infer inbreeding (F_{IS}) from molecular data (see Box 6 for more details on F-statistics). Because F_{IS} is insensitive to inbreeding in more ancient generations, the genomic coverage by ROHs has become a commonly used measure for inbreeding if single nuclueotide polymorphism (SNP) and whole-genome sequencing (WGS) data are available.

BOX 6

F-statistics

Wright's (1951) fixation indices can be defined in at least three hierarchical levels that measure the probability of two randomly sampled alleles being the same in a set of subpopulations (e.g. farms) that together form a population (e.g. breed). For instance, the probability of sampling the same allele twice in a subpopulation (a measure or inbreeding) can be obtained from the difference between the average H_o of the individuals in the subpopulation ($\overline{H_{oI}}$) and the HWE expectation of the heterozygosity for the subpopulation (H_e):

$$F_{IS} = \frac{H_S - \overline{H_o}}{H_e} = 1 - \frac{\overline{H_o}}{He_S},$$

F_{IS} measures deviations in random mating within a subpopulation from one generation to the next, with negative values (minimum is -1) indicating an excess of heterozygotes in the subpopulation, while positive values (maximum is +1) indicate a deficit of heterozygotes. It is worth noting, that in a subpopulation deviating from HWE, a single generation of random mating would restore the HWE proportions, rendering F_{IS} equal to zero. Similarly, the correlation between allele frequencies in a pair of subpopulations can be expressed as the difference in the average H_e of the two subpopulations ($\overline{H_e}$) and the expected heterozygosity estimated for the two subpopulations together (H_T, with $H_T = 2\,\overline{p} - \overline{(1-p)}$ and $\overline{p}$ being the average frequency of one of the two alleles in a biallelic locus):

$$F_{ST} = \frac{H_T - \overline{H_e}}{H_T} = 1 - \frac{\overline{H_e}}{H_T},$$

F_{ST} varies between 0 (indicating no differences in allelic frequencies between the two subpopulation) and 1 (indicating that the allelic frequencies in the two subpopulations are completely different), with the latter indicating that each of the two populations is fixed for a different allele, thus, within population diversity must be zero. F_{ST} was originally defined as a measure of fixation, but it is often seen as a measure of genetic differentiation, and some authors even use it as a measurement of divergence that works like a genetic distance. In practice, it cannot be used as a distance since it has a maximum value of 1, however Reynolds *et al.* (1983) suggested a distance measure that can be seen as measuring drift accumulating between populations in the absence of mutations contributing new genetic variation to the population. For small F_{ST} values, Reynolds distance is approximately the same as F_{ST} and linearly increases with time since separation (Reynolds, Weir and Cockerham, 1983; Weir and Cockerham, 1984).

Lastly, the probability of sampling the same allele twice can also be obtained for two samples randomly collected from the total population (i.e. ignoring the breakdown of groups of samples in subpopulations). F_{IT} thus measures deviations between the average H_o across all individuals in the population ($\underline{H_o}$) and the HWE expected heterozygosity for the population (H_T):

$$F_{IT} = \frac{H_T - \overline{H_o}}{H_T} = 1 - \frac{\overline{H_o}}{H_T},$$

As for F_{IS}, F_{IT} varies from -1 to +1, with the obtained value representing the deviation from HWE expected genotypic frequencies due to non-random mating and divergence in allele frequencies between subpopulations. The three fixation indices shown above relate to each other following: $1 - F_{IT} = (1 - F_{ST}) * (1 - F_{IS})$. This relationship enables for additional hierarchical levels to be included as required by the study design, e.g. should there be colonies within subpopulations, the relationship between the fixation indices would become: $1 - F_{IT} = (1 - F_{ST}) * (1 - F_{IS})*(1 - F_{IC})*(1 - F_{CS})$, with the corresponding F_{IC} and F_{CS} being:

$F_{IC} = \frac{H_C - \overline{H_o}}{H_C} = 1 - \frac{\overline{H_o}}{H_C}$ and $F_{CS} = \frac{H_e - \overline{H_C}}{H_e} = 1 - \frac{\overline{H_C}}{H_e}$, with H_C being the HWE expected heterozygosity for the colony, and with F_{IC} varying between -1 and +1, while F_{CS} varies between 0 and 1.

Inferring the frequency and length of ROHs in the genomes of farm animals helps to understand population histories, including the occurrence of bottlenecks, estimate inbreeding and identify signatures of selection (Bruford *et al.*, 2015; Ceballos *et al.*, 2018; Meyermans *et al.*, 2020; Peripolli *et al.*, 2017). For example, long ROH segments suggest recent inbreeding (consanguinity) and low genetic diversity. This has been observed for Holstein-Friesian cattle (Doekes *et al.*, 2019; Peripolli *et al.*, 2017; Purfield *et al.*, 2012; Upadhyay *et al.*, 2019) and is possibly associated with deleterious homozygous variants (Pryce *et al*., 2014). In contrast, relatively many short ROHs with only a few long ROHs indicate a reduced population size in the past and little recent inbreeding, as found for wild boars (Peripolli *et al.*, 2017). Conversely, hybridization reduces the ROH coverage and increases genetic diversity, as in cattle of admixed taurine-zebu ancestry (Kim *et al.*, 2017; Purfield *et al.*, 2012).

A high frequency of ROHs detected around a specific locus in several individuals can be indicative of recent positive selection (Curik, Ferenčaković and Sölkner, 2014; Zhang *et al.*, 2015). ROHs fixed in a population indicated a selective sweep by rapid selection for a beneficial mutation. However, appropriate analyses and demographic modelling are needed to disentangle the processes behind specific genomic patterns and the associated parameters (MacLeod *et al.*, 2013; see sections below on Relationships between breeds and Reconstruction of population history and demographic modelling).

Effective population size

The effective size (N_e) of a population corresponds to the number of individuals in an idealized Wright-Fisher population (i.e. reproducing with random mating, even sex ratio and non-overlapping generations) that would become inbred or lose diversity at the same rate as this population (Frankham *et al.*, 2002). The concept of N_e is central to population genetics, as it quantifies genetic drift and allows for comparisons across species or populations, but it is difficult to estimate. The census size of a population and the number of breeding animals (N_b) are usually larger than N_e, because actual populations do not follow the assumptions of an idealized population. Non-random mating and uneven sex-ratios are common in domestic animal populations that are under strong artificial selection, especially with the use of reproductive biotechnologies, and may also result from admixture events. Thus, N_e can be small even for a large population and through fixation of alleles it implies loss of genetic variability, fast genetic drift, a high level of inbreeding and possibly decreased viability (Kristensen *et al.*, 2015; Peripolli *et al.*, 2017).

Estimation of N_e depends substantially on the indicators of genetic diversity considered, such as inbreeding, and on variation of genetic diversity through time. For instance, from a dataset of sequences N_e can be estimated as θ / μ (where θ , is the population mutation rate estimated on the basis of the number of segregating sites, and μ is the individual mutation rate). Whole-genome data provide the least biased estimation of diversity (and thus N_e). Linkage disequilibrium (LD) based methods can account for sample size, mutation, phasing (assigning alleles to the paternal and maternal chromosomes), and recombination rate (Barbato *et al.*, 2015; Orozco-terWengel *et al.*, 2015). ROH may also serve as a robust genome-scale N_e estimator, although interpretation and scaling depend on local recombination (Bruford *et al.*, 2015).

However, computed values should be treated as estimates that are prone to biases, especially if the population is subdivided. For instance, Wakeley (1999) has shown that a structured population in which all subpopulations increase in size while exchanging more migrants will exhibit a signal of decrease of N_e despite the increase in the actual size of the population. Similar results were obtained by Mazet *et al.* (2016) for genomic data from single individuals analysed using the pairwise sequentially Markov coalescent (PSMC) method (see below). If N_e is used for translating various measures of genetic diversity within a specific model, it should be interpreted within that model. Various complex coalescent-based methods have been developed that allow to infer the population demography over time, including the variation of N_e or connectivity, such as the PSMC, the multiple sequentially Markov coalescent (MSMC) method and the site frequency spectrum (see section on Reconstruction of population history and demographic modelling).

Increasing the N_e of a breed or population is challenging, as it may require a balanced number of breeding males and females contributing to subsequent generations, thus limiting directional selection pressures, and avoiding inbreeding (Curik, Ferenčaković and Sölkner, 2014).

Several program packages available for the analysis of WGS data allow for the estimation of within-breed genetic diversity summary statistics. These software packages include the following: *PLINK* (Chang *et al.*, 2015; Purcell *et al.*, 2007); *VCF Tools* (Danecek *et al.*, 2011); R package *diveRsity* (Keenan *et al.*, 2013); *BEAGLE* specifically for detecting IBD (Browning and Browning, 2010); *SNeP* specifically to estimate N_e (Barbato *et al.*, 2015); *ANGSD* (Korneliussen, Albrechtsen and Nielsen, 2014) useful for low coverage WGS data. For more information, see Appendix 6.

The take home message is that estimation of parameters related to within-breed genetic diversity such as N_e is important, but this information needs to be interpreted properly and placed in the proper context. Estimation of such parameters should be complemented with clustering methods and demographic modelling (see subsequent sub-sections) to better understand evolutionary processes. By combining different approaches, it will be possible to use genomics to make a more informed management of breeds and other populations of livestock by monitoring N_e and practices such as crossbreeding, while avoiding high levels of inbreeding and ultimately preserving adaptive variation.

ASSESSMENT OF POPULATION STRUCTURE AND BETWEEN-BREED GENOMIC VARIATION

Principal component analysis Assessment of the genetic variation within breeds is usually an important objective for all genomic characterization studies. Such studies often involve multiple breeds or populations, however, so the analysis of the structure of the multiple populations and how they relate genetically to each other is also a common research application and provides information that is useful for management of the populations. The foremost way of visualizing genetic structure of a multi-population (multi-breed) sample is to perform a principal component analysis (PCA) on the individual multilocus genotypes. PCA is a statistical method to capture the variability associated with many variables (such as marker genotypes) into a much smaller number of variables that still contain most of the original variation. The resulting variables (the "principal components" – PC) are ordered in terms descending order according to the amount of variation they explain. Thus, PC1 explains the most variation, PC2 the second most, and so forth.

The theoretical ground of the PCA approach has been described by several studies. In a sample made of two distinct genetic groups, the PC1 separates these two groups and the proportion of variance explained by this first component corresponds exactly to the F_{ST} of the sample (McVean, 2009). Distinguishing the two groups is possible if (and only if) the product of the number of samples and markers is larger than $(1 / F_{ST})^2$ (Patterson, Price and Reich, 2006). Similar arguments suggest that more complex patterns of population structure also will be detected if the dataset is large enough. For instance, for a sample made of p distinct genetic groups, the number of PCs with significant contributions to the total observed variance is equal to p - 1 (Patterson, Price and Reich, 2006). The PCA may also suggest about past admixture events: admixed individuals are on the PCA located between the two parental populations while their distances to these parental populations reflect the percentage of admixture (McVean, 2009; Patterson, Price and Reich, 2006). When population structure arises from a continuous isolation-by-distance model, the PC often show genetic clines that correlate with geographical clines and the two-dimensional patterns may even resemble the geographic map of the area under study (Novembre *et al.*, 2008).

A standard procedure is to plot projections of the samples on the plane defined by two PCs, usually PC1 and PC2 or PC1 and PC3, thus providing a useful survey of the dataset. Although the individual PC have no biological or other significance *a priori*, plotting of the data frequently reveals patterns that suggest a logical interpretation. Figure 4 (P. Ajmone-Marsan, personal communication, 2021) shows an example of a PCA plot of cattle breeds and wild relatives. The wild relatives (Banteng and Gaur) appear at the top of the graph, separated from cattle breeds along the lower part. The Hong Kong feral (HKF) cattle, which seems to include some common ancestry with wild relatives, falls closest to the middle of the plot. The other cattle breeds are separated horizontally according to

subspecies. European taurine (EU-tau) breeds group on the far lower left, whereas Asian indicine (AS-ind) breeds are on the far right. African taurine (AF-tau) breeds are just to the right of the European taurine group, whereas Sanga cattle, which are considered to be a composite of African taurine and indicine breeds, are grouped in the bottom centre of the graph, between their two ancestral groups.

However, several limitations, potential sources of bias and associated misinterpretations should be avoided. Because the PCA plots are two-dimensional projections of samples from a multidimensional space, they visualize only a small proportion of the total variation. For instance, they do not unambiguously indicate a close relationship of samples and do not clearly indicate duplicates. Published plots are often too dense and thus only allow a superficial survey of the dataset. In addition, the use of many different symbols can be confusing. This can be remedied by plotting breed averages instead of individuals and by positioning the breed codes within the plot, just as always done with geographical names on a road map. Unbalanced sample sizes between genetic groups may strongly affect the position of these groups on the plot; those with large sample sizes being typically pushed towards the centre of the plot (McVean, 2009). If the dataset contains a breed that by genetic isolation differs substantially from all others, the major PCs mainly reflect this difference, that is, the bias due to inbreeding (Lenstra *et al.*, 2012). This result can be prevented by excluding such breeds while computing the PCs but not from the plot, an approach referred to as "supervised" PCA (Ciani *et al.*, 2020). Note also that several higher-order PCs typically each correspond to the contrast of one breed with the others (Kijas *et al.*, 2012).

Menozzi *et al.* (1978) introduced PCA for the study of human genetic variation across continental regions. It has since become a standard of population genetic data analysis since the 2000s, especially with the advent of high-density genotyping and WGS technologies. It can be applied using generic R functions or population genetics software such as *PLINK* (Chang *et al.*, 2015) or *eigensoft* (Price *et al.*, 2006). (See Appendix 6.)

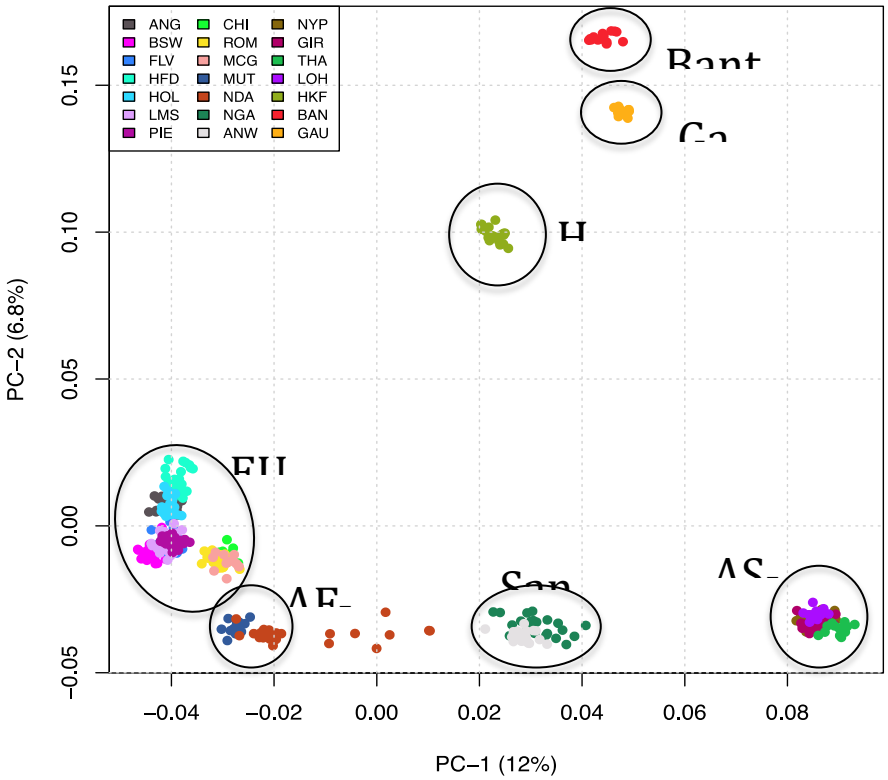

FIGURE 4

A principal components analysis plot of groups of bovine populations, including Banteng, Gaur, Hong Kong feral cattle (HKF) and groups of European taurine (EU-tau), African taurine (AF-tau), Sanga and Asian indicine (AS-ind) breeds.

While genetic clines that correlate with geographic clines often result from past migration waves, alternative explanations should be explored, such as isolation by distance (Novembre and Stephens, 2008). Ancient DNA (aDNA) samples with heterogeneous sample ages may also distort PCA representations, which can be amended by factor analysis (FA) instead of PCA (François and Jay, 2020). Another common method for comparing ancient and modern DNA samples is a supervised PCA in which low-coverage aDNA samples are projected in a plot of coordinates computed on the basis of high coverage depth DNA samples (Daly *et al.*, 2018), but this does not correct all potential biases related to sample age heterogeneity (François and Jay, 2020). For more details on the analysis of aDNA specimens see Box 7. Another mode of supervised PCA is the calculation of PCs that show the difference between geographical (or genetic) extremes, such the northern-most and the southern-most breeds, in order to test the clines between these extremes (Ciani *et al.*, 2020).

BOX 7

Archaeogenetics applied to farm animals

The recovery of genome variation from archaeological livestock samples is now eminently feasible, thanks to methods for DNA extraction and high throughput sequencing, particularly using instruments that give high read numbers from small fragments. Application of these technologies has helped solve the problems of contamination in the field - the presence of modern livestock DNA within an extract is now measurable through bioinformatic analysis of the resulting data and problem samples can simply be excluded. Also, it is possible to identify damage patterns to authenticate aDNA sequences and validate results.

Ancient genomics is an invaluable tool for understanding modern diversity and its origins. Archaeological bones from the cradle of domestication in the Near East have yielded genome data for both cattle and goat (Daly *et al.*, 2018; Verdugo *et al.*, 2019) showing that the initial domestication of both species must have been followed by secondary input from additional wild populations. A more extreme example is clear in pig ancestry where the Near Eastern genome which first enters Europe was almost completely replaced by introgression from local wild boar (Frantz *et al.*, 2019). Because DNA degrades with time, more recent genomes will be more accessible, and the early development of modern breeds will become visible. For example, in the horse, a sharp drop in genome heterozygosity within the last 250 years has been shown (Fages *et al.*, 2019). Ancient genomics can also reveal past domestic animal diversity that has since disappeared; for example ancient Iberian horse genomes from over four millennia ago have revealed the presence of a now-extinct lineage which is not ancestral to modern horse populations (Fages *et al.*, 2019).

One enduring aim is to understand genome variation in terms of impact and function and ancient genomics has the potential to add to this in two ways: in giving temporal resolution to the search for genome regions which have been under selection and in ancient epigenomics. A group of 8 000-year-old goat populations from Serbia and Iran showed evidence for selection of genetic variation involved in pigmentation and some production traits; the earliest discovered direct evidence of human manipulation of the genomes of herded (Daly *et al*., 2018). Inference of the methylation patterns of ancient molecules is also possible, although requiring high coverage sequencing, a route to directly mapping epigenetic markers in the past (Hanghøj *et al*., 2019).

Because of inherent properties of ancient DNA (particularly low concentration, short fragments and chemical damage) the number of methods that are useful is limited. Some suggested guidelines are as follows:

- Only high throughput sequencing is regarded as reliable and current state of the art; PCR-based assays including microsatellite size calls and Sanger sequences are no longer deemed reliable data.

- The best assay is the highest coverage genome retrieved by shotgun high throughput sequencing. This is a standard that the field should aim for, particularly with rare irreplaceable ancient specimens.
- However, many successful population genetic analyses can also be performed using low coverage depth genomes (i.e. <1x), and some samples may not feasibly be sequenced to higher density because of poor preservation.
- Methods such as RNA bait capture may also be useful (particularly for less well-preserved samples) for target read enrichment prior to sequencing. Examples of useful and achievable bait targets are either whole mitochondrial genomes, or specific autosomal loci such as selected genes or panels of thousands of informative SNPs.
- Additionally, it is highly desirable to use uracil–DNA glycosylase (UDG) treatment of sequencing libraries to remove C/G→ T/A misincorporations that have occurred as a result of age-related damage to ancient molecules.

Source: Dan Bradley

Multidimensional scaling (MDS) is an alternative method to survey the dataset by reducing the number of dimensions (Tzeng, Lu and Li, 2008). In practice MDS and PCA patterns are largely identical, but MDS cannot be run in a supervised mode.

Model-based clustering

Another common approach to summarize the genetic structure observed in a set of individuals is the clustering method first developed by Pritchard *et al.* (2000) and implemented in the software *Structure*. This method assumes there are *K* genetic groups (or clusters) and all polymorphic sites are at HWE and linkage equilibrium within these groups. Each individual *i* is modelled as a mixture of these *K* groups with specific proportions q_{ij} with $j = 1$ to *K*. *Structure* estimates these proportions from the observed individual genotypes using a Monte Carlo Markov Chain (MCMC) algorithm. In the most standard application (unsupervised clustering), the *K* clusters and their associated allele frequencies are inferred and the biological meaning of cluster number *j* is deduced *a posteriori* from the set of individuals such that q_{ij} is close to one. Ideally, this reveals the major ancestral populations and the composition of admixed individuals or populations. Supervised clustering allows one to define *a priori* the ancestral clusters and infer directly the admixture proportions of the other individuals, similar to assignment methods like *GeneClass* (Piry *et al.*, 2004). The original approach of Pritchard *et al.* (2000) was later extended to account for sampling location (Hubisz *et al.*, 2009), linked loci and shared ancestry between clusters (Falush, Stephens and Pritchard, 2003). More efficient inference approaches of the same model were also developed in the software *Frappe* (Tang *et al.*, 2006) and *Admixture* (Alexander, Novembre and Lange, 2009), allowing the analysis of much larger datasets and larger values of *K,* albeit with the constraint that loci are independent from each other, that they are in linkage equilibrium.

Again, there are caveats and pitfalls to be avoided. Many scientists choose to interpret the results of model-based clustering in terms of population history. For instance, the separation of one cluster into two sub-clusters at high *K* values is often interpreted as a past divergence event. More generally, inferred clusters are equated to ancestor populations and any breed that is not assigned to one cluster is assumed to be admixed. However, *Structure* is essentially a clustering approach for genetic data, but does not model evolutionary processes such as drift, mutation, migration or divergence (see sections on Relationships between breeds: genetic distances and Reconstruction of population history and demographic modelling for such methods). The implicit assumption that the ancestral populations are present in the dataset is often not met. Thus, the inferred clusters do not necessarily correspond to real past or present populations and strongly depend on the composition of the dataset. Applying *Structure* to populations along a genetic cline formed by preferential mating of proximate parents (isolation-by-distance) may infer homogeneous clusters that contain the extreme populations and mixed compositions for the central populations (Engelhardt and Stephens, 2010). This may be remedied by modelling continuous spatial variation between (Corander, Waldmann and Sillanpää, 2003; Francǫis, Ancelet and Guillot, 2006; Guillot, Mortier and Estoup, 2005) or within clusters (Bradburd, Coop and

Ralph, 2018), the latter option allows the analysis showing both discrete and continuous genetic structure. By the sampling and the inbreeding bias, inferred clusters tend to correspond to overrepresented and inbred breeds, respectively (Lenstra *et al.*, 2012).

However, in spite of a wide-spread over interpretation, model-based clustering does contribute data analysis by showing graphically: (i) groups of related individuals from one or more breeds; (ii) ancestral proportions, provided inferred clusters correspond to ancestral populations, which may be accomplished in the supervised mode; and (iii) subdivision of breeds not inferred by PCA or genetic distances.

As an alternative to the supervised mode, breed-specific admixture analysis (BSAA) is based on the selection of 300 ancestry informative markers (AIM) specific for a breed suspected to be introgressed, for instance by markers that have the highest F_{ST} between the breed and a group of other breeds. It was demonstrated that *Structure* analysis with 300 to 400 AIM markers provides a more sensitive detection of (pre)historic admixture than *F*4 analysis (see below).

Both *Structure* and PCA attempt to summarize observed genetic diversity using a reduced number of *K* latent factors, namely the clusters for *Structure* and the PCs for the PCA (Engelhardt and Stephens, 2010). More specifically, both methods decompose the genotype matrix *X* as $X = QF$, where *Q* is a matrix of *n* individuals × *K* factors that quantify the individual compositions. The two methods impose different constraints to the matrices *Q* and *F*; the coefficients of *Q* are positive and must sum per individual to one with *Structure*, but not with PCA. Other latent factor approaches were proposed to analyse genetic data (Engelhardt and Stephens, 2010; Frichot *et al.*, 2014) and represent interesting alternatives to PCA and *Structure*. For instance, the sparse Nonnegative Matrix Factorization (sNMF) developed by Frichot *et al.* (2014) and implemented in the *LEA* R package estimates admixture proportions with the same accuracy as the *Admixture* software while reducing computation time by a factor from 3 to 30 and being less affected by the inbreeding of the analysed populations.

Genetic distances

Genetic distances between individuals

The third common approach to visualize genetic structure uses a matrix of genetic distances between all pairs of individuals. For SNPs of a diploid species, distances are based either on the identity by state (IBS) of markers or on genomic relationships of individuals (Grünwald *et al.*, 2017; Yang *et al.*, 2011), which both can be computed using *PLINK*. These distances can be visualized via MDS analysis, the alternative to PCA described earlier (Tzeng, Lu and Li, 2008) or via Neighbour-Joining trees (Saitou and Nei, 1987; see section on Relationships between breeds: genetic distances). Such trees unambiguously show the identity or close relationship of samples and are recommended for quality control of SNP datasets (see Section 3). They also show the level of inbreeding of breeds via short terminal branches. However, individual-based trees can be misleading on the deeper phylogenetic relationships of the breeds, which are better explored by genetic distances based on allele frequencies of breeds (see section on Reconstruction of population history and demographic modelling). A representation of genetic distances in a network that connect each individual to its ten nearest neighbours ("supermagnetic clustering") shows genetic structure within populations (Neuditschko, Khatkar and Raadsma, 2012), which may be related to the uneven use of breeding sires (Neuditschko *et al.*, 2017).

So far, the approaches described assume markers to be independent, but several recent studies take advantage of haplotype information of phased datasets, that is, data with information about which alleles of different markers are on the same haploid chromosome copy (maternal or paternal). Such information is not provided by dideoxy-Sanger or short-read sequencing. Without data from the (expensive) PacBio or Oxford Nanopore long-read sequencing (see Section 3), phasing must be inferred statistically by exploiting observed LD patterns in the population or genotype information from close relatives (Al Bkhetan *et al.*, 2019; Browning and Browning, 2007; Delaneau, Marchini and Zagury, 2012).

The *ChromoPainter* approach implemented in the program *FineSTRUCTURE* (Lawson *et al.*, 2012) builds a coancestry matrix of individuals based on local haplotype similarity along the genome and uses this matrix for PCA and clustering (similar to *Structure*). For both applications, the use of haplotype information significantly improved the resolution of the inferred population structure. The method was developed for high-density SNP datasets and its performance with low-density (<50 000 SNPs) datasets is not yet clear. Therefore, when using this method, it is essential to show that the linked mode, using haplotypes, gives better results than the unlinked mode (using separate SNPs). Going beyond this genome-wide description, the *Relate* (Speidel *et al.*, 2019) and *tsinfer* (Kelleher *et al.*, 2019) methods reconstruct the phylogeny of observed haplotypes, i.e. the most-probable inheritance at every SNP position in the genome, accounting for the haplotype diversity observed among sampled individuals around this position. All these methods are either based on or inspired by the model of Li and Stephens (2003), which iteratively considers one sampled haplotype and tries to reconstruct it as a mixture of the others.

In a study from The Bovine HapMap Consortium (Gibbs *et al.*, 2009), genotype data at 37 470 SNPs was available for 497 cattle from 19 geographically and biologically diverse breeds. This dataset was analysed using a *Structure* like (Figure 5A) and a PCA approach (Figure 5B). PC1 separates zebu and taurine breeds, similar to the *Structure* analysis with $K = 2$ clusters. PC2 further separates African versus European taurine breeds, similar to the *Structure* analysis with $K = 3$ clusters. Breeds resulting from admixture between these three main groups are indicated by dashed rectangles and clearly stand out in both analyses. The *Structure* analysis with $K=9$ identifies several clusters corresponding to single breeds.

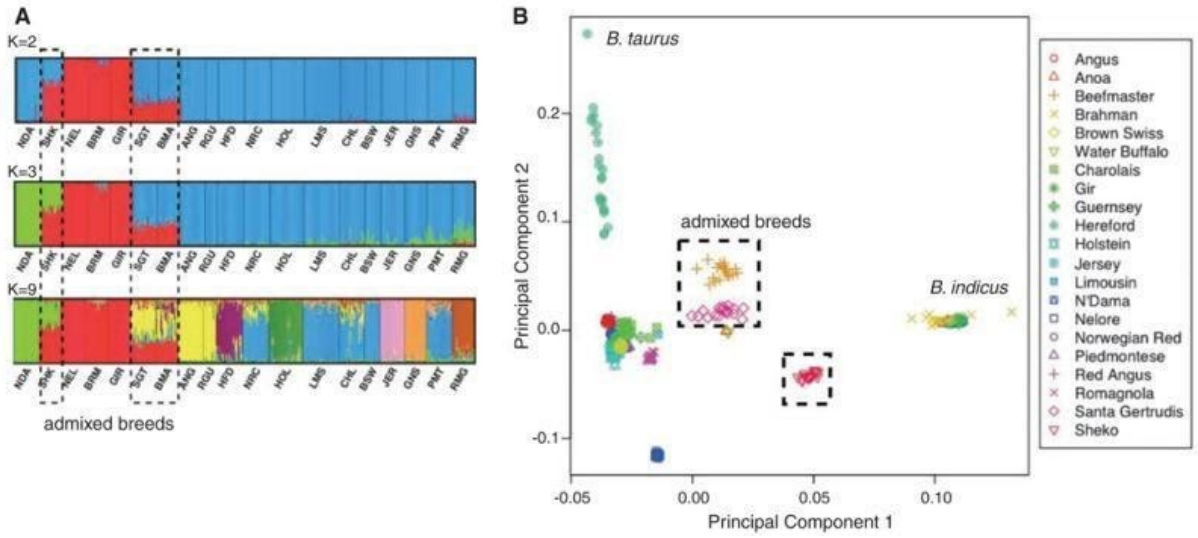

FIGURE 5

Genetic structure of worldwide cattle breeds, according to model-based clustering (A) and principle component analysis (B) (Adapted from The Bovine HapMap Consortium, 2009).

Genetic distances among breeds

Since the development of selective breeding in the eighteenth century, the mating of animals with similar desired characteristics has resulted in the establishment of breeds, which are usually defined by a range of morphological, economic and behavioural traits (Taberlet *et al.*, 2008). Depending on the degree of genetic isolation, animals from the same breed are genetically more similar to each other than to animals from different breeds. These differences among animals from different breeds can be quantified using genetic distances between individuals (see corresponding section) and also by determining the components of the genetic variation within and between breeds using an Analysis of Molecular Variance (AMOVA; Excoffier, Smouse and Quattro, 1992; Michalakis and Excoffier, 1996).

There are two important differences between phylogenies at the species (macroevolution) and breed (microevolution) levels. First, the bulk of molecular differences between species are quantified via fixed differences in the DNA sequence. However, the mammalian substitution rate, approximately 1.0 to 1.25 × 10^{-8}/bases/generation (-The 1000 Genomes Project Consortium, 2010; Kong *et al.*, 2012; Venn *et al.*, 2014), is too slow to generate many of the DNA differences observed between breeds that diverged at most 10 000 years ago and in most cases only 50 to 200 years ago (Laval, SanCristobal and Chevalet, 2002). Instead, the major DNA-based differences between breeds depend on genetic drift, the mechanism behind the change of allele frequencies in genetic markers as a function of the N_e. Second, many species do not interbreed and hybrids of related species are often infertile. As a consequence, the phylogeny of species can be largely represented by a tree topology without connections between branches. In contrast, crossbreeding of breeds is a rule rather than an exception and trees of breeds are at best a practical visualization than an estimation of the true phylogeny of breeds.

The extent of genetic drift separating populations A and B can be quantified as the additive statistic f_2, or the mean of $(P_A - P_B)^2$ with P_A and P_B corresponding to the frequencies of allele P for a SNP in populations A and B, respectively. The f_2 is the average of the squared differences across all genetic markers studied in the two populations (Patterson *et al.*, 2012; Peter, 2016). Several genetic distances have been derived from f_2, such as the Reynolds distance or D_R (Reynolds, Weir and Cockerham, 1983), which is the preferred genetic distance for breed comparisons (Laval, SanCristobal and Chevalet, 2002) and which correlates very well with the F_{ST} fixation index (Wright 1951). Because it approximates the average inbreeding coefficient of the two breeds compared, D_R trees may represent the loss of within group genetic diversity (Laval, SanCristobal and Chevalet, 2002). However, with SNP panels this is likely to be confounded by the ascertainment bias (AB), which may increase the estimates of diversity of the industrial breeds.

Note that the accuracy of allele frequencies depends on an adequate sample size per breed. In practice, 20 animals per breed are considered to be sufficient for biallelic markers; in fact, a sample size of 20 diploid animals has a probability of 95 percent of including the whole genealogy of the breed (Hein, Schierup & Wiuf, 2003). As with any statistical test, larger amounts of data will increase precision. If at least 20 samples cannot be obtained, a lower sample size tends to increases the proportion of homozygote genotypes and thus the genetic distances to other breeds but usually does not alter the topology of a tree (see the following section). The F_{ST} distance as estimated by Weir and Cockerham (1984) accounts for low and uneven sample sizes. Using medium-density array datasets, Ramjak *et al.* (2018) calculated genetic distances between short haplotypes consisting of the alleles of up to four SNPs located within a stretch of 150 kilobases.

Phylogenetic trees and networks. The neighbor-joining (NJ) algorithm (Saitou and Nei, 1987) is the most common and one of the simplest methods to reconstruct phylogenetic trees on the basis of a matrix of $n \times n$ pairwise genetic distances between n breeds. It works by choosing the pair of breeds that after being combined gives the largest reduction of the total branch length of the tree, after which the distance matrix is recalculated and reduced to $(n - 1) \times (n - 1)$ breeds, continuing iteratively until all breeds have been paired. Such a tree complements PCA and model-based methods and represents a hierarchical clustering of breeds, with the length of the terminal branches indicating the degree of genetic isolation of a breed. In contrast to the (non-recommended) UPGMA (unweighted pair group method with arithmetic mean) approach, the NJ algorithm does not assume an even rate of change across breeds, which agrees with the effect of genetic drift being breed specific. The effect of genetic drift can be observed when the NJ tree is visualised like a phylogram where the trees' branch lengths are proportional to the amount of change in each breed. Other visualisations such as cladograms are not suitable to visualise the effect of genetic drift as the branches in such methods are not proportional to the amount of change.

Conceptually, the topology of a tree directly shows genetic events as breed divergence, genetic drift and breed isolation. In practice, this is realized mainly for the most closely related breeds, but is not shown by PCA or model-based clustering. Trees of individuals based on genetic distances also show the genetic drift, but do not show very well the relationships between breeds. However, phylogenetic trees of breeds have a few inherent limitations:

- Testing the significance of clusters is problematic and bootstrapping values are low.
- The deeper bifurcations are supported by short branch lengths and may not be realistic.
- As shown by their hierarchical topology, trees cannot visualize reticulations and are thus confounded by gene flow between breeds leading to admixture. Breeds that have been derived by crossbreeding, such as taurine–indicine composites, tend to cluster in the NJ tree in between the source breeds (Ginja *et al.*, 2019; Kopelman *et al.*, 2013; Mei *et al.*, 2018). Whereas this still may be a useful visualization of a genetic cline with again the terminal branches showing the breed-specific genetic drift (Decker *et al.*, 2014), it should be realized that the genetic differences are dominated by the degree of crossbreeding. Thus, bifurcations and branch lengths no longer correspond to divergence events and genetic drift, respectively. If crossbreeding occurred recently, the genealogy of the alleles depends on the position of the alleles in the genome.

Various methods have been developed to introduce reticulation in a tree by construction of an admixture graph. The most popular methods are the *NeighbourNet* and *Treemix* algorithms. *NeighbourNet* graphs (Bandelt and Dress, 1992; Dopazo, Dress and Von Haeseler, 1993; Dress, Huson and Moulton, 1996) can be constructed via the easy program *Splitstree* (Huson, 1998) with many graphical options. It is based on the NJ algorithm, but considers three instead of two breeds before reducing the difference matrix. *NeighborNet* graphs share the main advantages of NJ trees by reproducing the terminal branches and unambiguous clusters of breeds but replace the often meaningless bifurcations of a tree by networks (Ciani *et al.*, 2020; Kijas *et al.*, 2012; Pitt *et al.*, 2019). These are not to be interpreted in terms of specific genetic events and may correspond to ancestral breed pools, and at least visualize the uncertainty of the phylogenetic reconstruction. However, *NeighborNet* graphs rarely show reticulations corresponding to specific cases of crossbreeding. Kijas *et al.* (2012) and Ciani *et al.* (2020) showed that deeper phylogenetic splits, i.e. between breed clusters rather than between breeds, are resolved better if: (i) the most recent and documented crossbreds are removed, and (b) the unambiguous monophyletic breed clusters are combined in one taxon. All this may contribute to the formulation of specific hypotheses to be tested statistically or by modelling.

The sophisticated program *Treemix* (Pickrell and Pritchard, 2012) constructs a maximum-likelihood tree based on the covariances of the allele frequencies, which overall agrees with NJ trees. Subsequently, the agreement between the covariances explained by the tree and the real covariances can be improved by adding vectors between branches that represent gene flow (migration) between those branches. Taking a user-defined number of migrations, *Treemix* optimizes a new tree topology, the start and endpoint of the migrations and their weights (Decker *et al.*, 2014; Fonseca *et al.*, 2018; Orozco-terWengel *et al.*, 2015). However, the graphical output has a poor quality that cannot be adapted easily. More importantly, we recommend that the migrations inferred on the basis of an overall match of graph-based and real covariances are tested, for instance, using the f_3 or f_4 statistics, which is rarely done. Ciani *et al.* (2020) found that plausible admixture confirmed by both f_4 analysis and by BSAA were only partially reconstructed by *Treemix*. Other methods to construct admixture graphs are discussed in the next section.

Detection of admixture using f_3, f_4 and *D* statistics- The availability genome-wide datasets obtained by SNP arrays (e.g. The Bovine HapMap Consortium, 2009; Kijas *et al.*, 2012) and WGS (e.g. Alberto *et al.*, 2018; Bosse *et al.*, 2014; Fan *et al.*, 2020) led to the understanding that hybridization and admixture are more common than previously thought. The density of data now allows for an accurate estimation of the new statistics for the detection of gene flow between well-diverged breeds (Patterson *et al.*, 2012; Peter, 2016). These statistics can be used with both individuals and breeds Figure 6 shows the most common applications of the f_3 and f_4 statistics based on allele frequencies of three (f_3) or four (f_4) breeds. As is shown, they can be decomposed as the sum or difference of distances between nodes as quantified via the f_2 genetic drift statistic. For a more extensive theoretical treatment involving path lengths and admixture proportions we refer to Patterson *et al.* (2012) and Peter (2016).

Calculation of the f_3 statistic while using an outgroup is useful for quantifying the relative affinity of individual animals or breeds, for instance of ancient and modern DNA samples (Daly *et al.*, 2018). The properties of f_3 for detection of admixture can be understood by considering that the average of $(P_C - P_A)(P_C - P_B)$ becomes negative if P_C is intermediate between P_A and P_B for many SNPs, which

indeed reveals recent admixture. However, a positive value of f_3 does not exclude admixture, which limits its use.

The statistic f_4 is more predictable since admixture follows directly from systematic differences from zero. These deviations can be understood by considering that sharing of characteristic alleles by source and target or, for a more general case, by A and C (Figure 6), but not by the other possible pairs of breeds, leads to a correlation of P_A and P_C, which makes the average of $(P_A - P_B)(P_C - P_D)$ positive; decomposing f_4 into a function of four f_2-scaled distances as indicated in Figure 6 results in a total length of zero without gene flow, but becomes greater than zero if gene flow shortens the $f_2(A, C)$ distance.

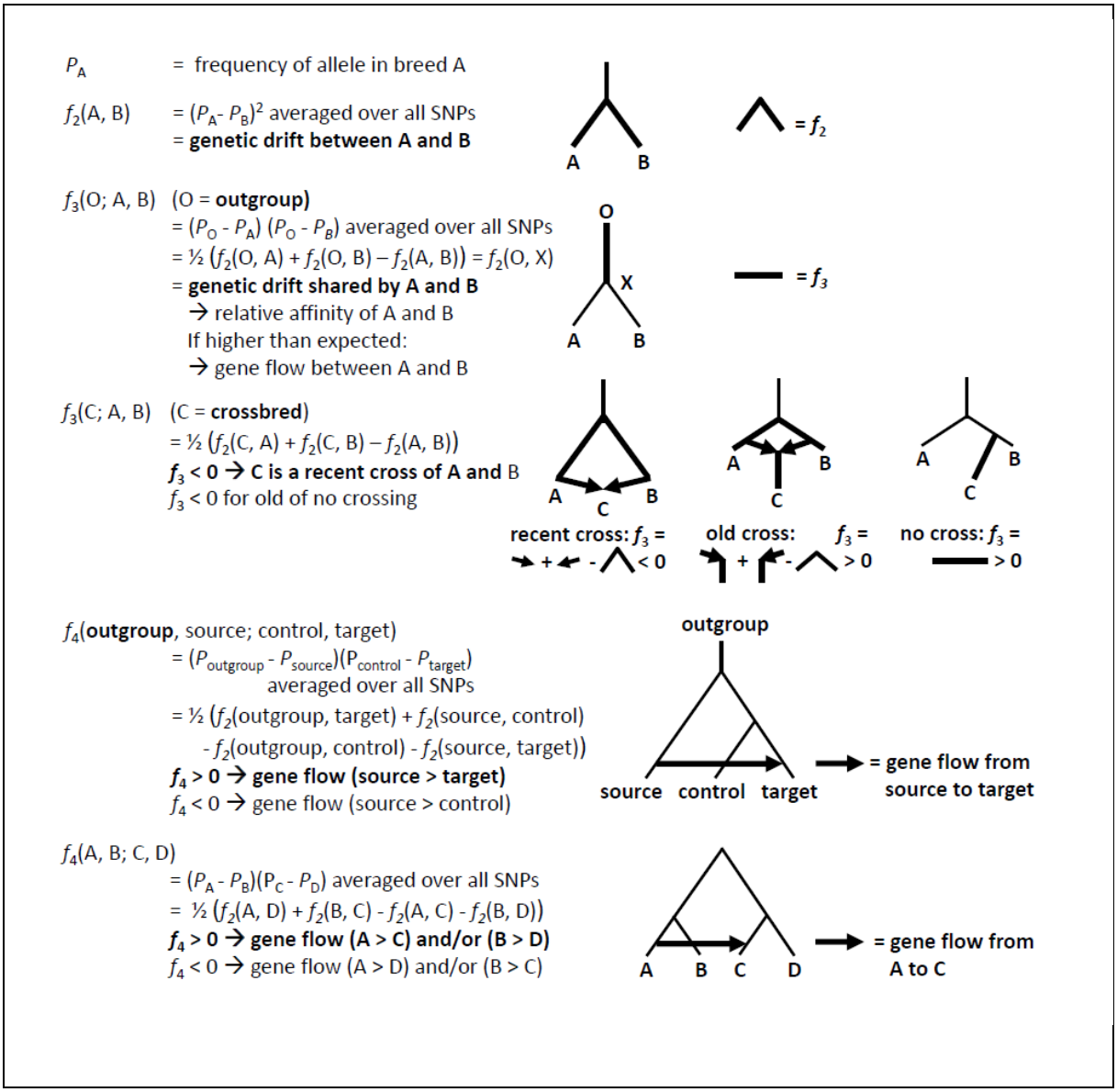

FIGURE 6

Definitions of the f_2 statistic, the outgroup f_3 statistic, the admixture f_3 statistic and the admixture f_4 statistic with an outgroup and the more general case with four breeds. It is shown how the statistics correspond to branch lengths and how they are used as test statistic for the occurrence of gene flow.

The f_4 statistic is related to the *D* statistic from the ABBA/BABA algorithm. This method considers the sister taxa one and two, the donor three and the outgroup (O). For a given SNP, *A* is the ancestral allele fixed in *O* while *B* is the derived allele fixed in the donor three. For a haploid genome, the possible allele combinations for one, two, three and O are AABA, ABBA, BABA and BBBA. Ignoring the non-informative AABA and BBBA, a higher admixture from three into two than from three into one follows from a majority of SNPs with the ABBA pattern:

$D = [n(ABBA) - n(BABA)]/[n(ABBA) + n(BABA)]$, in which n(*ABBA*) and *n*(*BABA*)are the counts of SNPs with the ABBA and BABA patterns, respectively. For a collection of samples, the use of allele frequencies instead of counts and an adapted formula allow to consider also SNPs for which *A* and *B* are not fixed in the outgroup and the donor, respectively (Patterson *et al.*, 2012). For statistical testing, *D* is preferred above f_4 (Patterson *et al.*, 2012). However, we stress that both methods strictly depend on a tree-like evolution of the breeds being considered

The *f*-statistics can be used to construct more realistic admixture graphs than a *NeighborNet* or *Treemix* graph. However, because this procedure is very specialized and only partially automated by software, the methods described in the literature are not generally used. The programs *QPTOOLS* from the *AdmixTools* package (Patterson *et al.*, 2012) and *ADMIXTUREGRAPH* (Leppälä, Nielsen and Mailund, 2017) take as input a user-specified Admixture graph and all possible f_2, f_3 and f_4 values and then optimize the branch lengths and admixture proportions. *MIXMAPPER* (Lipson *et al.*, 2013) selects breeds for which a hierarchical tree can be constructed (i.e. no negative f_3 values, additivity of f_2 values) and then sequentially and interactively adds admixed breeds while optimizing topology, sources of gene flow, branch lengths and admixture proportions.

Phylogeny across the genome. So far we compared individuals, or groups, on the basis of genome-wide genetic variation. However, as a consequence of genetic recombination, the genome consists of segments with widely different genealogical histories (Li and Durbin, 2011). For instance, within crossbred individuals, portions of the genome may originate from different breeds (e.g. Fan *et al.*, 2020), which result from various demographic processes such as migration, inbreeding and selection. Software products have been developed that attempt to overcome the deviating phylogenetic signals from markers and thus establish the true genealogy that reflects the breeds' evolutionary processes. Among these, *SNAPP* (Bryant *et al.*, 2012) uses a likelihood-based approach to estimate the groups' true tree based on the genetic variation of markers across the genome. Hobolth *et al.* (2007) established a framework that enables simultaneous modelling of divergence or speciation times, and demographic parameters such as the ancestral N_e, which has been successfully applied to model the divergence history of the European bison, *Bison bonasus* (Gautier *et al.*, 2016).

RECONSTRUCTION OF POPULATION HISTORY AND DEMOGRAPHIC MODELLING

Several methods presented above can identify evolutionary processes in the history of species, populations and breeds, while representing patterns of differentiation. The subsequent sections will focus on methods that explicitly aim at detecting, dating and quantifying population size changes and other events such as admixture or population splits. Because the complexity of the evolutionary history of species and breeds probably precludes its full reconstruction, even with genomic data, the interpretation of an inferred demographic history should thus always be done with caution. Also, in the case of domesticated animals one must keep in mind that this evolutionary history integrates the recent artificial selection, the domestication process, and the history of the wild species from which they were derived. Genetic and genomic data enable testing alternative interpretations and models by comparing the genetic consequences of alternative evolutionary scenarios with observed data. This is particularly true with the approximate Bayesian computation (ABC) framework (Beaumont, 1999; Beaumont, Zhang and Balding, 2002).

An initial step is to describe the coalescent theory, which is at the heart of modern population genetics (Hudson, 1990, 2002; Wakeley, 1999). It was developed in the 1980s and 1990s to study the statistical properties of gene trees for non-recombining DNA fragments (Kingman, 1982) evolving under a simple demographic model that assumes panmixia and isolation from other populations (i.e. no population structure), non-overlapping generations, neutrality, and constant population size (e.g. the Wright-Fisher model). It has since been extended to account for population size changes (Griffiths and

Tavaré, 1994), population structure (Herbots, 1994), recombination (Griffiths and Marjoram, 1997) and different sampling times to integrate aDNA or viral samples. Some types of selection can also be integrated. The coalescent theory takes a retrospective approach that differs from modelling or simulation approaches used before the 1980s. Coalescence and classical simulation are now commonly denoted as backward and forward simulation, respectively. A coalescent approach works backward in time and focuses on the statistical properties of the gene trees of the samples. If a population is large, the probability that lineages coalesce (have a common ancestor) is low, whereas in a small population this probability is high. As a consequence, starting from a particular sample size the length of the branches connecting these samples will depend on the changes in population size, as one goes backward in time, and as the number of remaining lineages decreases. In a more general way, the coalescent theory studies how gene trees depend on the parameters of the demographic models of interest (past population sizes, migration rates, age of a bottleneck or an admixture event). Coalescent-based methods are sample-based and thus computationally less expensive than the population-based forward methods, which require the simulation of the whole population. The development of programs such as *ms* (Hudson, 2002), allowed the development of simulation-based inference methods including ABC. Forward simulations are still used, for instance for selection, and are being revived after recent improvements in efficiency (e.g. *SliM* from Haller and Messer, 2019).

The coalescent theory played a central role in the development of inferential methods in the 1990s and 2000s. It was shown that bottlenecks and expansions have different effects on the shape of trees, hence allowing the use of coalescence-based inference to distinguish between the two types of events. This led to the development of likelihood-based methods to compute the conditional probability $P_M(Data \mid Param)$ of generating the observed *Data* under a predefined demographical model *M*, defined by a set of parameters $Param = (Par_1, \ldots, Par_k)$, which may correspond to divergence times, population sizes, or migration rates. Likelihood methods apply optimization algorithms, often via a sophisticated search of the parameter space, to find parameter values that best explain the observed *Data*, i.e. the maximum likelihood estimates of the parameters. A Bayesian perspective can also be used to estimate probability density functions for the parameters, assuming prior distributions for these parameters, $P_M(Param)$ and multiplying them by the likelihoods to obtain posterior distributions using Bayes formula: $P_M(Params \mid Data) = P_M(Param) * P_M(Data \mid Param) / P_M(Data)$. Likelihood-based methods were developed between the mid-1990s and mid-2000s and applied to different demographic models to consider: population size change (Beaumont, 1999); population structure and gene flow (Beerli and Felsenstein, 2001); population split and post-split migration rates (Nielsen and Wakeley, 2001); or admixture (Chikhi, Bruford and Beaumont, 2001).

In the last decade several important methods were developed that allow users to infer either simple or complex demographic models with genomic data. One of the most popular is the PSMC method of Li and Durbin (2011), which uses the genome of a single diploid individual (or two haploid genomes) and infers a histogram or "skyline plot" usually interpreted as a history of population size changes. The PSMC uses information on mutation and recombination rates to analyse and interpret the distribution of heterozygous sites along the genome. Under panmixia (i.e. absence of genetic subdivision) the method estimates the history of population size changes that best fits the distribution of heterozygous sites. Schiffels and Durbin (2014) extended the method by allowing the use of multiple phased genome sequences simultaneously (MSMC). Under panmixia the MSMC also estimates the history of population size changes. It is important to note that both the PSMC and MSMC only use information from the first coalescence time, *Tk*, where *k* is the number of haploid genomes and thus ignores most of the information contained in the whole coalescent tree. These methods provide only information on events that took place within the time frame defined by the *Tk*. Thus, the more haploid genomes one uses with MSMC, the earlier the start and end of the reconstructed curves.

Other methods use the site or allele frequency spectrum (AFS) as a way to summarize genomic data across many independent loci. For instance, the *Fastsimcoal2* (Excoffier *et al.*, 2013) and *dadi* (Gutenkunst *et al.*, 2009) software allow the user to infer the parameters of tree models in which populations are allowed to have different population sizes, and may be connected by gene flow. For computational reasons, the number of populations is limited, typically two to five populations

depending on the methods. *Fastsimcoal2* uses a likelihood approach and can thus be used to compare alternative tree models, with or without gene flow or admixture.

A general inferential framework for comparing alternative models is ABC, which is particularly well-adapted to study complex demographic datasets and evolutionary models for which the likelihood may be difficult or impossible to compute. In such cases, the principle of ABC is to use large numbers of simulations under different demographic models and parameter values together with two approximations. First, the data are summarised by a limited number of summary statistics (such as n_A, Tajima's D or H_e), and the objective is to estimate $P_M(SummData \mid Param)$, where *SummData* is a vector of the summary statistics. Second, simulated datasets are selected with a *SummData* that, according to a tolerance threshold, are close enough to the *SummData* of the real dataset. The parameter values corresponding to the selected simulations (usually the best 0.1 or 1 percent) are kept for inference. The number of simulations used to estimate the posterior distributions of the parameters of a particular model is typically $>10^5$-10^6. The coalescent theory played a central role in the development of ABC methods by enabling an efficient simulation of data. The simplest ABC algorithm is usually called "rejection" because it only requires filtering the best simulations according to the tolerance threshold and rejecting the others.

The ABC framework is extremely flexible and can handle complex datasets and other situations for which classical models do not perform well. For instance, Boitard *et al.* (2016) used the AFS together with summary statistics measuring LD at different distances to reconstruct a history of population size changes in large samples obtained from a set of four cattle breeds (Angus, Fleckvieh, Holstein and Jersey). ABC can also be used to compare alternative demographic models by computing the relative proportions, among alternative models, of simulated *SummData* closer to the observed *SummData*. It can also test the quality of the inference process by generation of pods (pseudo observed data) on the basis of the different models and check if an ABC analysis of the pods indicates the corresponding model and generates a meaningful posterior distribution of parameters. However, the large number of simulations required by ABC methods can become computationally intensive with genomic data and complex alternative models.

Beaumont *et al.* (2002) demonstrated that a regression algorithm achieved the same quality of the inference as the rejection algorithm but with ten times fewer simulations, allowing a much lower tolerance threshold. Other approaches have been proposed to improve the regression step by using non-linear regression and neural networks (Csilléry *et al.*, 2010). More recently, a machine-learning approach called "random forests" has been proposed as a way to avoid the selection step of summary statistics, and the need to define a tolerance threshold (Raynal *et al.*, 2019). One powerful and versatile software package for ABC analyses is the ABCtoolbox (Wegmann et al., 2010)

MITOCHONDRIAL DNA AND THE SEX-CHROMOSOMES

The genetic diversity of autosomes, sex chromosomes and mtDNA)is affected in different ways by demography, and their combined analysis provides insight into the various historical processes that shaped present-day populations. Because of their lack of recombination, the evolution of mtDNA and Y-chromosomal DNA is less complex than the autosomal evolution and can be reconstructed largely by obtaining a phylogenetic tree and coalescence analysis. Generally, it reveals links between the domestic species and their wild ancestors, and shows ancient bottlenecks that contributed to differentiated haplogroup distributions, e.g. in (sub)continents (Lenstra *et al.*, 2014), but is considerably less informative than autosomal DNA for reconstructing breed histories. In addition, mito-nuclear discordances occur frequently simply due to incomplete lineage sorting or by sex-biased introgression, such as for the European bison carrying cattle-like mtDNA (Wang *et al.*, 2018), zebu outside Asia carrying taurine mtDNA (Ginja *et al.*, 2019), or possibly the Grey jungle fowl that carries a domestic/Red jungle fowl mitochondrial haplotype (Lawal *et al.*, 2020).

Before the availability of WGS data, the analysis of Y-chromosomal variation was limited by a lack of informative markers. It shows the paternal history, which by the small male population size for most domestic species has led to a considerable breed-level differentiation (see Box 8). For many species, in current genome assemblies the Y chromosomes are partially represented by unplaced scaffolds that are only present in males, and for a large part consist of Y-chromosomal repeats and multicopy genes.

Note that single-copy male-specific SNPs are hemizygous, that is, they have one allele per individual without heterozygosity scores (H_o and H_e) and should not have scores in females.

BOX 8

Y-chromosomal variation

The mammalian Y chromosome is small, carries only a few genes and mainly consists of repetitive DNA. Due to its complex DNA structure, it is often omitted from genome assemblies. But the Y chromosome has a powerful property: its male-specific, non-recombining part (MSY) is transmitted as a tight linkage group (a haplotype) from the father to the son. Hence, genetic variation on the MSY reflects the paternal history of populations (Jobling and Tyler-Smith, 1995).

The MSY diversity is best studied in the humans, where deep insights derive from several decades of steady progress in variant discovery and haplotype analyses. A diverse spectrum of polymorphisms on the human MSY, ranging from small- and large-scale rearrangements, SNPs, indels, and CNVs as single tandem repeats, now represent a clear picture of the human male genealogy (reviewed in Jobling and Tyler-Smith, 2017). The wealth of MSY markers available in humans, enables the study of evolution over different time scales. The work in humans has also demonstrated that a community accepted MSY phylogeny, based on defined haplotype determining variants, is pivotal to consolidate results from different studies (Van Oven *et al*., 2014).

The tremendous amount of WGS data becoming available is now empowering comprehensive fine-scaled MSY haplotype analysis in domestic species. To infer the MSY pattern in a population of interest, proper MSY variant discovery is a prerequisite. As a first step, one usually needs to define MSY sequences that can be used as the reference for variant calling from NGS data. The potential of MSY draft assemblies, combined with an approach to define MSY regions appropriate for reliable variant calling, has been recently demonstrated in the horse (Wallner *et al*., 2017), camel (Felkel *et al.*, 2019) and sheep (Deng *et al*., 2020). After mapping WGS data to the reference and variant ascertainment, the allelic states of MSY variants can be catenated into haplotypes and a robust haplotype phylogeny can be built under the principle of maximum parsimony. Once the MSY phylogeny is defined, haplotype dispersal should be screened by genotyping haplotype determining markers in larger sample collections (Deng *et al*., 2020).

Defining the MSY haplotype signatures that are diagnostic for a given breed or even a single influential individual enables to univocally trace their influence. This process has been conducted in the horse to assess the influence of the Thoroughbred breed (Wallner *et al.*, 2013; Cosgrove *et al*., 2020). From horse studies the importance of an appropriate ascertainment panel has become evident, as haplotypes that are private in rural populations are still not fully defined (Han *et al*., 2019) and ancient remains enlighten significant turnover of MSY haplotype spectra through time and space (Fages *et al*., 2019).

Establishing MSY phylogenies is computationally demanding and obtaining inferences on haplotype distribution is labour intensive. The building of international consortia to coordinate MSY research in livestock would be extremely beneficial.

Source: Barbara Wallner

Deviations from the expected levels of genetic diversity and divergence of the X chromosomes have been observed in several species. For mammals, the X chromosome is present in two copies in females and one copy in males, therefore the expected ratio between the genetic diversities of chromosome X and the autosomes is 0.75 (Ellegren, 2009; Wilson Sayres, 2018). However, ratio has this varied widely in several breeds of European and African cattle (Da Fonseca *et al.*, 2019). Increased levels of X-chromosomal diversity have been shown to result from growth in population size (Van Belleghem *et al.*, 2018; Pool and Nielsen, 2007). Conversely, a reduction in relative diversity follows an overall reduction in N_e (Ellegren, 2009; Pool and Nielsen, 2007). In livestock populations, this is often a consequence of breeding practices that involve the selection of a few individuals with desirable

phenotypes. A reduction in population size would result in relatively strong genetic drift of X-chromosomal alleles, which has a lower N_e than on autosomes. In the case of cattle, molecular signatures of a bottleneck were detected in several cattle breeds (Boitard *et al.*, 2016; Da Fonseca *et al.*, 2019; Ginja, Telo Da Gama and Penedo, 2010; Kim *et al.*, 2017; Martín-Burriel *et al.*, 2011).

Another process that is expected to impact the relative diversity of the sex chromosomes is sex-biased gene flow (Ellegren, 2009; Wilson Sayres, 2018). This is especially relevant for cattle, for which artificial insemination of many cows with semen of very few sires is a widespread practice. Furthermore, it is known that historically female populations were more likely to be geographically constrained and human-driven hybridization/crossbreeding may have been carried out mainly using males (Lenstra *et al.*, 2014). Generally, introgression is more efficient in the autosomes than in the sex chromosomes, as observed for the domestic chicken (Lawal *et al.*, 2020). This pattern can be explained by stronger species barriers on the sex chromosomes (Meiklejohn *et al.*, 2018).

As shown by Da Fonseca *et al.* (2019), the use of whole-genome sequences led to two observations that are probably related and may very well be of fundamental significance. First, a ratio of F_{ST} values for autosomal and X-chromosomal sequences depends dramatically on the cattle being compared. For taurine vs indicine cattle, the X-chromosomal F_{ST} values are ~25 percent larger than the autosomal values. In contrast, within taurine cattle, the autosomal F_{ST} values are about twice those of the X chromosome, and for Iberian cattle with autosomal F_{ST} values ranging from 0.1 to 0.26, the X-chromosomal range is 0 to 0.04. Second, in model-based clustering at K=2 the clusters correspond to taurine and indicine cattle. In this case, the autosomes of African taurine cattle show an indicine component of 18 to 36 percent, whereas the X chromosomes remained purely taurine. It is not yet clear if these effects can be explained by a relatively ineffective X-chromosomal introgression; there may well be consequences of genetic conflicts underlying the meiotic drive (Meiklejohn *et al.*, 2018). Since LD-based pruning removes X-chromosomal SNPs, this was never noticed with bead-arrays, but it is certainly worth following up.

IDENTIFICATION OF GENOMIC REGIONS SUBJECT TO SELECTION

Domestication, adaptation, breed formation and selective breeding have left specific genetic patterns, "selection signatures", in genomic regions of farm animal breeds (e.g. Raudsepp *et al.*, 2019). The identification of selection signatures is a central goal in evolutionary and population genetic studies but has importance also in characterization of animal genetic resources (AnGR) for food and agriculture. With methods for the detection of selection signatures it may be possible to identify, for example, genes related to economically important traits or for adaptation to challenging environments (Saravanan *et al.*, 2020; Weldenegodguad *et al*., 2019). Breeds displaying special adaptive traits have typically high priority in the conservation. Selection signature studies can also provide knowledge of quantitative trait loci (QTL) for production characters and important causal mutations (Saravanan *et al*; *et al*., 2020). Moreover, genome-wide searches for regions associated with phenotypic traits may promote our understanding of function of genomes. For example, several whole-genome data sets have shown that a great proportion of SNPs exhibiting selection signatures correspond to non-coding genomic regions, indicating that selection occurs specifically via the regulatory elements of genomes (Librado *et al*., 2015).

Types of selection

During the domestication process, genomes of animals have been shaped by natural selection and artificial selection. Human-initiated artificial selection (or selective breeding) conducted over generations for productive, morphological, fertility and other economically traits has had remarkable effects on genetic and phenotypic variation within and between breeds of farm animal species (e.g. Felius, 1995; Forutan *et al.,* 2018). Natural selection driven by environmental circumstances involves differential reproduction of genetically diverse types so that some types leave more offspring than others (e.g. Falconer and MacKay, 1996). Three forms of natural selection are typically classified: (i) positive selection; (ii) purifying selection (negative or background selection); and (iii) balancing selection.

The various forms of selection have specific effects on allelic and genotypic frequencies, as well as on genomic structural variation in animal populations (de Simoni Gouveia *et al.,* 2014). When artificial selection or natural positive selection increases frequency of an advantageous allele, the frequencies of neutral variants that are physically linked to the advantage allele tend to change (de Simoni Gouveia *et al.,* 2014 and references therein). This process is termed "genetic hitchhiking". When a substantial (or complete) reduction of genetic variation at the target of selection and its surrounding genomic regions occurs, it's called a "selective sweep" (Maynard-Smith, 1974). Moreover, in these chromosomal regions an increase in the average LD is expected to occur, leading to long haplotypes. These changes can typically be observed in statistical parameters related to genetic diversity within and between populations (de Simoni Gouveia *et al*., 2014 and the references therein). The within-population diversity is typically decreasing in these regions, while between-population diversity tends to increase. In negative (purifying) selection, the selective pressure operates against disadvantageous alleles at the target of selection, thus not always affecting the surrounding genetic variation, especially when there are multiple advantageous or disadvantageous alleles. However, negative selection can also result in loss of genetic variation at the region of the genome which, although neutral, is linked to the negatively selected sites; this process is known as "background selection" (Charlesworth *et al*., 1993). The third main form of natural selection, balancing selection, maintains multiple alleles in a locus and high genetic diversity (Oleksyk *et al*., 2010). Major histocompatibility complex genes are classical examples of genes shaped by balancing selection (Hedrick, 1998). More information on various forms of selection and their effects on nonneutral and linked neutral genetic variation is provided in the literature (de Simoni Gouveia *et al*., 2014; Horscroft *et al*., 2019; Saravanan *et al.,*2020). Figure 7 is a graph showing trends in diversity of the KITLG genic region of sheep and goats (Alberto *et al.*, 2018). This region includes subregion that show decreased diversity in domesticated sheep but decreased diversity in goats, presumably due to species-specific selection processes associated with domestication.

Methods of detecting selected loci

Methods to detect selection signatures have been recently reviewed, including by Utsunomiya *et al*. (2015), Horscroft *et al.* (2019) and Saravanan *et al.* (2020). The methods commonly used in livestock studies can be classified into two main groups: methods based on intra-population statistics and those on inter-population statistics (Saravanan *et al*., 2020). In intra-population methods, genomic or DNA-marker data are compared within populations. In this group, methods are based on the site frequency spectrum, LD or the identification of genomic regions with reduced variation compared to the genome average (de Simoni Gouveia *et al*., 2014; Alachiotis and Pavlids 2018; Saravanan *et al*., 2020). Alachiotis & Pavlids (2018) introduced a composite evaluation test, μ statistic for detection of positive selection, which examines genomic regions by quantifying the site frequency spectrum, the levels of LD and the amount of genetic diversity along a chromosome. PCAdmix (Brisbin *et al*,, 2012) is a sophisticated and popular program that can be used to detect adaptive introgression. It uses phased genotypes applies PCA and a Hidden Markov Model on sliding windows along the genome in order to detect the ancestry of the individuals as this varies across the genome. Local deviations of the ancestry shared by all individuals of a breed than may indicate that the region is involved in adaptation

The inter-population approaches are based on allele frequency differences and the degree of differentiation between the populations (Saravanan *et al.*, 2020). The level of genetic differentiation can be analysed using single site differentiation or haplotype-based differentiation (Saravanan *et al.*, 2020). Coding sequences of orthologous genes among animal species can also be compared and tested based on synonymous and non-synonymous substitution rates applied to detect the neutrality or the form of selection (positive or negative - de Simoni Gouveia *et al.*, 2014). In addition, "landscape" genetics and genomics (Manel *et al.*, 2003; 2010) approaches may reveal associations between genomic data and environmental variables and thereby environmental adaptative signatures (e.g. Mdladla *et al.*, 2018; Vajana *et al.*, 2018). See Box 9 for more information on landscape genomics.

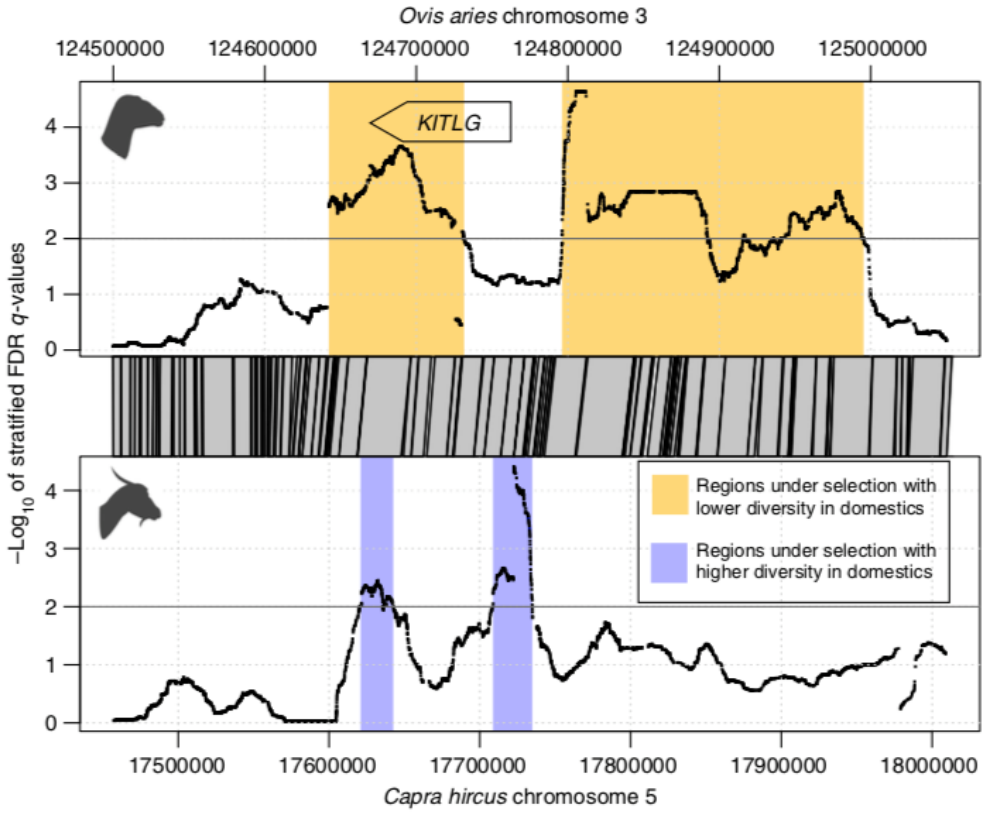

FIGURE 7

Chromosomal regions under selection within the KITLG gene for sheep and goats. The regions have decreased genetic diversity in domesticated sheep and a greater diversity in goats (from Alberto *et al.*, 2018)

BOX 9

Landscape genomics

Landscape genomics is a research field that combines population genomics, landscape ecology, computer science and bio-geoinformatics to explicitly quantify the effects of environmental heterogeneity on neutral and adaptive genetic variation and underlying processes. Landscape genomics has potential for addressing research questions in various research fields, including ecology, evolution, and conservation of livestock breeds. The discipline is based on the idea that environmental conditions can substantially affect the genetic variation of local populations, with important consequences for evolutionary processes.

The landscape genomics approach consists of linking genome-wide information to environmental variables identify potentially valuable genetic material, like genomic regions conferring adaptive advantages. Typically, data obtained from genome-wide scans carried out on a number of animals from populations living in different habitats or across ecological clines will be compared to geo-environmental information characterizing these habitats (yearly amount of precipitation, monthly temperatures, number of days with ground frost, etc.). Comparisons are carried out by means of association models calculated between allele or genotype frequencies and the values of the environmental variables. The parallel and simultaneous processing of up to billions of statistical models (i.e. hundreds of environmental variables times millions of SNPs) makes it possible to identify genomic regions of interest, for instance linked to genes involved in adaptation to heat, or genes involved in processes of resistance against parasites.

In developing countries, small ruminants play an important role in the livelihood of a large proportion of farmers. In these countries, conserving traditional breeds is essential since the latter are able to survive in changing habitats showing often harsh conditions to which their rich genomic resources make it possible to adapt. Therefore, landscape genomics is an important tool to identify the key genetic features of local adaptation in order to support sustainable breeding of low-input livestock.

In the current context of global warming, this genetic adaptive potential can be directly integrated into conservation management using SPatial Areas of Genotype probability (SPAGs). This tool is able to transpose landscape genomic results into an adaptive potential spatial prediction framework. SPAGs can be integrated with climate change projections to forecast the future spatial distribution of genotypes. The analysis of the mismatch between current and future SPAGs ("genomic offset") makes it possible to identify vulnerable populations lacking the adaptive genotypes necessary for future survival.

Environmental data are necessary to compute landscape genomics models. The WorldClim (2020) database contains monthly minimum, maximum and average temperature and total precipitation together with a series of bioclimatic variables computed from these variables. The Climatic Research Unit (CRU, 2021) in Norwich, United Kingdom, a recognised world's leading institutions concerned with the study of natural and anthropogenic climate change, also produces several datasets to be used in the context of landscape genomics studies.

For computation, several open-source software implement landscape genomics algorithms. These are provided in Appendix 6.

Source: Stephane Joost

Appendix 6 provides examples of software and includes a section for applications that have been used in domestic animal studies to detect selection signatures. Because various methods have different strengths and weaknesses, it is typically recommended that more than one method is used for analyses of selection signatures (de Simoni Gouveia *et al*., 2014; Utsunomiya et al. 2015; Yurchenko *et al*., 2018; Horscroft *et al*., 2019; Saravanan *et al.*, 2020). In animal populations, demographic events, such as population expansions, genetic bottlenecks, and population subdivision, can lead to false selection signals that mimic signatures of selection (de Simoni Gouveia *et al.*, 2014; Utsunomiya *et al*., 2015). de Simoni Gouveia *et al.* (2014), Utsunomiya *et al*. (2015); Ahrens *et al.* (2017) and Horscroft *et al.* (2019) have presented approaches to avoid effects of demography or other factors causing bias in selection signature analyses.

GENOME-WIDE ASSOCIATION STUDIES

The development and application of dense marker panels that can be genotyped at a high-throughput frequency by using microarray technology (see Section 3) has facilitated the generation of large-scale genotype data for many domestic animal species. The most widely applied microarrays provide genotypes for between 50 000 and 800 000 SNPs that are evenly spread across the genome. Millions of domestic animals have been genotyped with these arrays in the past decade with the primary goal to implement genomic selection and predict genomic breeding values (Georges *et al*., 2019). The availability of dense genotypes and a diverse array of phenotypes, often available over multiple generations, makes domestic animal populations amenable to genome-wide association testing. In genome-wide association studies (GWAS), statistical tests are applied to examine if molecular markers are associated with the expression of phenotypes. Because such analyses effectively involve the testing of multiple hypotheses simultaneously, the markers need to pass stringent significance thresholds to be considered as significantly associated with the trait of interest. Various statistical methods have been developed to account for multiple testing (Rice *et al*., 2008; Joo *et al*., 2016). Marker panels typically do not contain causal mutations, so the primary aim of GWAS is to identify regions likely to harbour causal variants through an association with a physically linked genotyped variant. The precision to localize causal variants through GWAS with anonymous marker panels depends on the marker density. Sparse marker panels may result in large confidence intervals for QTL, rendering the identification of causal variants a difficult task. Due to extensive LD between adjacent

markers in populations with small N_e, QTL confidence intervals may also be large when dense marker panels are employed. Functional investigations are required to differentiate between causal variants and anonymous markers in LD with the causal variants (Karim *et al*., 2011).

Apart from the density of the marker panel, the power of the GWAS, i.e., the probability to detect trait-associated markers, depends on: (i) the LD between molecular markers and QTL; (ii) the heritability of the trait; and (iii) the size of the mapping cohort (Goddard and Hayes, 2009). The most widely applied genotyping arrays in domestic animal species have been developed for cosmopolitan breeds (Matukumalli *et al*., 2009), thus they are less informative for local breeds that are greatly diverged from the breeds used to develop the marker panels. Moreover, the genotyped markers are often depleted for low-frequency variants. This may result in AB (see Section 3), which can compromise the power of GWAS, particularly in local breeds. While large mapping cohorts may be established easily in cosmopolitan breeds, the small census size of local breeds may also be a limiting factor for powerful GWAS to detect markers associated with within-breed variation. Across-breed GWAS may reveal variants that are associated with breed differences (Schoenebeck and Ostrander, 2013). However, even in large cosmopolitan breeds, the large sample size required to detect trait-associated variants may not be readily available for novel traits, thus requiring efforts to coordinate collaborative projects to establish powerful mapping cohorts (Lu *et al*., 2018).

Both binary (discretely distributed) and complex (continuously distributed) traits may be subjected to GWAS testing. Association testing for binary traits requires genotypes for a cohort of individuals that expresses a particular phenotype ("cases") and for a cohort of control individuals that don't express the phenotype ("controls"). Chi-square tests and Fisher tests of allelic association may be implemented to test the association between markers and binary traits (Balding, 2006). It is important though, that both cohorts are matched because systematic differences in allele frequencies between cases and controls due to stratification may confound the GWAS and lead to false-positive associations. The highest-ranking PCs of a genomic relationship matrix can be effectively used to account for population structure and cryptic relatedness in case-control GWAS (Price *et al*., 2006; Nosková *et al*., 2020). Case-control association testing of genome-wide markers has been successfully used to reveal causal variants for many Mendelian traits in domestic animals, and most of them are curated at the Online Mendelian Inheritance in Animals (OMIA) database (OMIA, 2021).

Most complex traits are highly polygenic and recent empirical evidence suggests that they are predominantly determined by additive effects (Hayes *et al*., 2010; Hivert *et al*., 2021). Each causal variant explains only a small fraction of the phenotypic variation of complex traits. To avoid an inflation of false-positive association signals arising due to population stratification (see above), software (Appendix 6) for complex trait GWAS typically rely on mixed model-based approaches that fit a (genomic) relationship matrix. While QTL that explain a relatively large fraction of the trait variation can be detected in small mapping cohorts (Pausch *et al*., 2011; Signer-Hasler *et al*., 2012), the discovery of QTL with small effects usually requires tens-of-thousands of genotyped animals (Bouwman *et al*., 2018).

The first GWASs were performed in 2005 for monogenic traits in human populations (DeWan *et al*., 2006; The Wellcome Trust Case Control Consortium, 2007). Likewise, the first SNP-based GWAS in domestic animals were performed for monogenic traits using genotypes for approximately 25 000 SNPs (Karlsson *et al*., 2007; Charlier *et al*., 2008). Since then, both the size of the mapping cohorts and the marker density have increased continuously. The largest GWASs in domestic animals have been performed in cattle including the genotypes at more than 25 million variants for more than 94 000 samples (van den Berg *et al*., 2020). With an ever-increasing size of mapping cohorts and more markers being tested, efficient software tools are required to undertake marker quality control and the association analyses. Since its implementation, the PLINK software has undergone a major overhaul and has become an indispensable tool for genotypic data analysis at the population scale (Purcell *et al*., 2007; Chang *et al*., 2015). Moreover, the .bed/.bim/.bam-format and its derivatives used by PLINK to store genotypic data is accepted as an input for most downstream analyses. Basic association analyses may be performed with PLINK. However, more sophisticated statistical methods that are based on Bayesian or mixed model-based approaches may be conducted efficiently with the EMMAX (Kang *et al*., 2010), GCTA (Yang *et al*., 2011) and BayesR (Erbe *et al*., 2012; Moser *et al*., 2015) and LFMM2

(Caye *et al*., 2019) software tools (see Appendix 6). Latent factor mixed models (LFMMs) are in particular currently gaining popularity. This method has been designed for detection of gene-environment associations and identifies SNPs with allele frequencies that correlate with environmental clines or with phenotypes.

With an increasing number of animals being sequenced, GWAS nowadays often rely on imputed sequence variant genotypes (see Section 3). Informative haplotype reference panels are required to accurately infer sequence variant genotypes for mapping cohorts that have array-derived genotypes. Such reference panels have been established for a variety of animals, including cattle (Daetwyler *et al*., 2014; and Hayes and Daetwyler, 2019), sheep (https://sheepgenomesdb.org/), goats (http://www.goatgenome.org/vargoats.html) and dogs (Plassais *et al*., 2019). Genotypes imputed from informative haplotype reference panels may readily reveal causal variants in GWAS (Pausch *et al*., 2017). The integration of functional data is required to fine-map QTL regions and reveal causal variants (Xiang *et al*., 2019). Cross-species approaches where information from humans is used to prioritize causal variants in livestock has also been explored (Costilla et al 2020 and Raymond et al 2020). Novel approaches that are based on low-pass sequencing and imputation-based genotype refinement offer a cost-effective opportunity to perform sequence-based GWAS also in populations that either lack informative haplotype reference panels or for which dense microarrays are not available (Li *et al*., 2020; Rubinacci *et al*., 2020; Fuller *et al*., 2020), which offers opportunities for less common livestock species.

APPLICATION OF GENOMICS IN SUSTAINABLE USE AND CONSERVATION

These guidelines emphasize the use of genomics for the characterization of AnGR and therefore are designed to facilitate the implementation of Strategic Priority Area (SPA) 1 of the Global Plan of Action for Animal Genetic Resources (FAO, 2007) on Characterization, Inventory and Monitoring of Trends and Associated Risks. Genomics can also play an important role in the sustainable use and development (SPA2) and conservation (SPA3) of AnGR. Those roles for genomics are somewhat out of the scope of these guidelines, so they will not be addressed in detail. The subsequent sub-sections will briefly introduce some of the opportunities that genomics provide for these SPA. Oldenbroek (2017) addresses the contributions of genomics to the management of AnGR in more detail.

Every breed has a history

Never the same; livestock breeds are not static entities but have continuously evolved. Changes have been substantial after domestication (Larson and Fuller 2014; Zeder, 2015), but perhaps the most consequential changes have taken place during the last 200 years through the formation of breeds and the systematic breeding. This can now be reconstructed by the analysis methods outlined in these guidelines. The following are a few examples showing how the evolution of breeds is directly or indirectly relevant for their management and conservation.

Local and global breeds

Most breeds originate from local populations. Thus, they potentially harbor unique adaptive traits and often belong to the local cultural heritage. However, influence from other populations has often been undocumented. The influence from other breeds may not necessarily affect their adaptation and traditional status, but reconstruction of their local and non-local ancestry by genomic analysis will be most revealing. A recent study in from Brazil revealed some enexpected information about Latin American sheep breeds (Box 10). The genomic characterization of groups of breeds also may reveal the sources of adaptive introgression, which contributes to both resilience and uniqueness (Barbato *et al*., 2017; Chen *et al*., 2018). However, the introduction of exotic highly productive breeds, such as the Holstein dairy cattle or the Merino wool sheep, does replace robust unique animals by a highly productive uniform populations that require intensive management.

BOX 10

Application of basic genomic diversity studies: A case study of Brazilian hair sheep breeds

A recent study in Brazil (Paim *et al*., 2021) provides an example of how genetic diversity data may advise national conservation and breeding programs. Genotypes (50 000 SNPs) from seven Brazilian sheep breeds (five hair and two coarse wool types) and 87 worldwide breeds were used to evaluate population structure and phylogenetic origin and quantify genetic diversity. The main result shows that Brazilian hair sheep, contrary to other breeds, have shared alleles with both breeds from both African and European continents (Admixture and PCA). This mixture of genetic sources, plus adaptation to local conditions has resulted in high genetic variability (neutral and functional) that should be considered in conservation and breeding programmes. For example, Brazilian hair sheep breeds have high frequencies of a specific allele (FecGE) in the GDF9 gene (Silva *et al*., 2011) linked to litter size, which has not been reported in most breeds outside the American continent. This trait has already been used to enrich the Brazilian National Gene Bank and has provided farmers with an additional tool for in-herd selection. The study also helped to identify rare or genetically distant national breeds for which germplasm collection should be a priority. The Brazilian Somali was the first breed prioritized for germplasm collection, and the genebank has already begun cryopreservation of semen and embryos. The endangered Brazilian Fat-tail sheep was also identified as a priority for cryoconservation. Analyses of genetic diversity and population structure were also used to identify the breeds with more potential for breeding programs. The number of animals in the Santa Ines (hair) breed has expanded over the last decade, in part due to crossbreeding with other populations, and is currently the primary commercial hair breed in Brazil (McManus *et al*., 2014). With high genetic diversity and low inbreeding compared to other Brazilian breeds, it has potential to achieve high genetic gains in production traits within a well-designed breeding programme.

Source: Samuel Paiva

For goats Colli *et al*. (2018) observed on the basis of genome-wide SNP genotypes reported a strong geographic partitioning of goats. In addition, several ancient goat DNA samples are related to modern goats from the same location (Daly *et al.*, 2019), indicating a high degree of local ancestry. In contrast, The International Sheep Genomics Consortium (2012) reports a high level of geneflow among global sheep populations, which was presumably driven by the desire for increased wool or mutton production and has eroded the geographic differentiation.

What's in a name?

Several successful breeds have been exported to other countries, most notably from Europe into the Americas and Australia. Populations of these "transboundary" breeds often have been developed separately on different continents. This has led, for instance, to the development in America of black varieties of European beef breeds, and genomic analyses show that several American breeds form a cluster separate from their European namesakes (Davenport *et al.*, 2020). In France, this divergence of a breed into distinct units has even occurred with the national population of the Alpine goata breed, detectible by genomic characterization (see Box 11). Conversely, closely related breeds with different herdbooks, but regularly exchanging sires, may be different in name only.

BOX 11

Genomic characterization as tool the breed recognition process

In France, a ruminant breed is officially recognized by the Ministry of Agriculture if it fits a set of criteria validated by a commission of experts. The items to provide to the commission include a

description of the breed's history, its phenotypic description, the existence of a breed register, a list of breeders, and the common set of rules for organizing the selection or conservation process etc.

The most prominent French goat breed is the Alpine, which originated in the northern part of the French Alps. For 50 years the breed has been intensively selected for dairy traits. The main selection nuclei are now located in the west and centre of France. Meanwhile, some farmers kept the traditional population in the Alps and claimed to have limited exchanges with the rest of the breed. A breeders' association was created 20 years ago to preserve this population which they called "Savoie". The official recognition was complicated to achieve because the experts doubted that the population was significantly different from the Alpine breed.

In the mid-2010s, a medium SNP goat chip (by Illumina) became available in France for a reasonable price (about 40 $/genotype). The breeders' association, in collaboration with a research and development institute (IDELE), took this opportunity to sample and genotype about 40 goats to compare it with available Alpine genotypes. IDELE performed a 'PCA based on a kinship genomic matrix which showed that the Savoie breed is indeed a close relative to the Alpine breed, but that the population differs now from the Alpine (Figure B). The analysis was also an opportunity to detect animals that were clearly crossbred recently with the Alpine, leading to their withdrawl from the conservation program. The results of the genomic analysis were provided to the expert commission and were one of the factors that led to the official recognition of the Savoie breed in 2020.

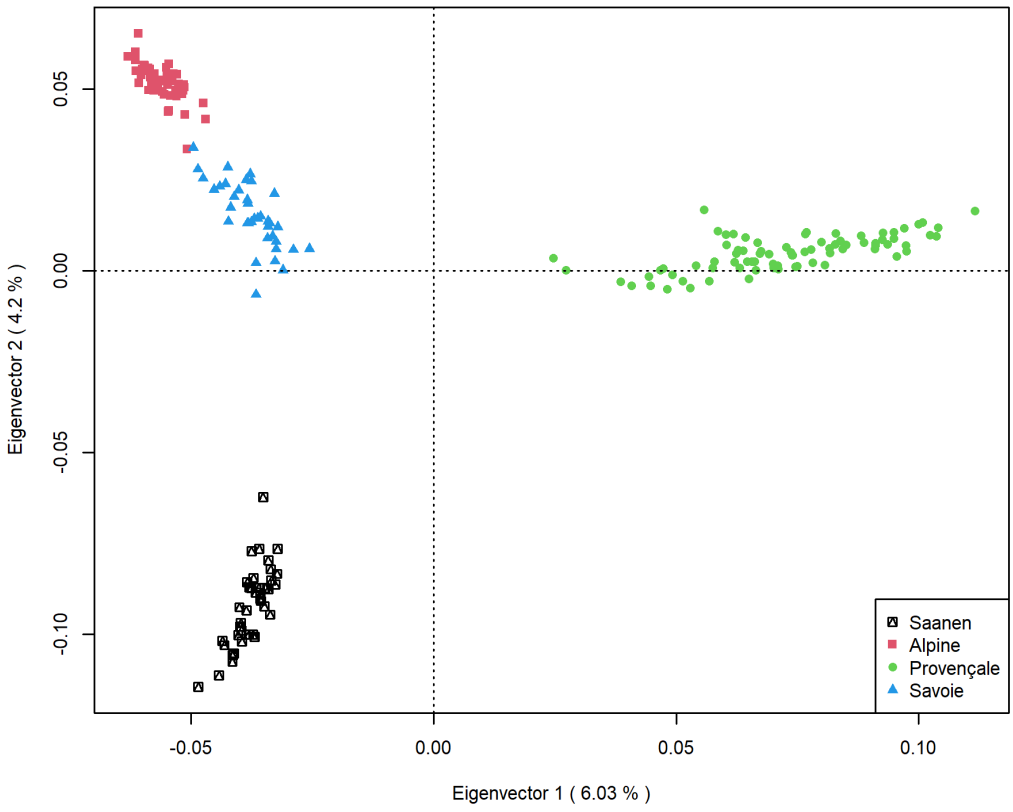

Figure B11

First two principal components using four French selected goat populations from the Alps

Source: Coralie Danchin-Burge

Valuable DNA or valuable breeds?

Genome analysis allows an accurate assessment of the diversity, both across the genome or at the level of individual genes. However, it is not straightforward to link this to this to the value of a breed. Differences between breeds are quantified via allele frequencies instead of the number of mutations, as for species. A unique allele frequency profile by genetic drift does not in itself make a breed valuable (European Cattle Genetic Diversity Consortium, 2005), but may be relevant for breed management in different ways. First, genetic distances on the basis of the allele frequencies may reveal that a breed underwent a long separate history in a given environment and thus is expected to possess unique adaptive traits. Second, levels and patterns of homozygosity as influenced by small population sizes and/or assortive mating are warning signs of inbreeding depression affecting health and fertility. Third, as mentioned, it may indicate a unique adaptive influence of cross-fertile species (Chen *et al.*, 2018). On the other hand, short genetic distances do not exclude large differences in phenotype. For instance, the unique Merino wool sheep are closely related to the coarse-wool Churra breed (The International Sheep Genomics Consortium, 2012) despite their vastly different fibre-quality characteristics.

It may be argued that livestock breeds carry largely overlapping portions of the available adaptive variation. Thus, it may be questioned if in general breeds are to be considered as unique units of conservation (Felius *et al.*, 2015). On the other hand, the diversity of livestock species is primarely managed on the level of breeds, which may include conservation measure, adjustment of breeding objectives and crossbreeding. Thus, livestock breeds remain the units for management of diversity, which can now more and more be supported by genomic analysis.

Sustainable use and development

The long-term sustainability of a given breed will be enhanced if that breed has greater survival and longevity, is more productive and profitable, or otherwise contributes competitively to the livelihood of its keepers (FAO, 2013). Improvement through selective breeding is one opportunity to increase the average genetic merit of breeds for heritable traits associated with sustainability. Although its technically not “genomic characterization” and will thus not be covered in depth in these guidelines, “genomic selection” (Meuwissen *et al.*, 2001) is an approach that statistically associates the phenotypes of animals within a population to their genotypes, as defined by large panels (typically thousands) of SNPs. Operationally, genomic selection can be performed either by predicting the substitution value of each of the alleles of the panels of SNPs or by using multi-locus genotypes to obtain more accurate relationship matrices than can be obtained by using pedigrees and replacing the matrices in typical methods based on best linear unbiased predictions (BLUP). When large amounts of historical performance and pedigree data are available, the breeding values of animals without phenotypes can be predicted much more accurately than with pedigrees alone. These “genomic breeding values” may also allow selection to occur prior to sexual maturity, thus increasing response to selection by decreasing the generation interval. Genomic selection is currently being routinely practiced in several livestock species (Georges *et al.*, 2019).

Alas, because the accuracy of genomic selection increases with the amount data available, breeds with small population sizes and/or a lack of historical performance and pedigree data may have less opportunity to benefit than do larger breeds (Mészáros *et al.*, 2015; Obšteter *et al.*, 2019). Breeds with small population sizes tend to also have small N_b and small N_e, meaning that intense selection to maximize genetic response may quickly decrease genetic variation to unacceptably low levels. Genomic data provides the possibility to balance selection response and maintenance of genetic variation at the molecular level, such as by applying genomic-based optimum contribution selection (Obšteter *et al.*, 2019; Sanchez-Molano, Pong-Wong and Banos, 2016).

Genomics can also contribute to the management of genetic diversity as it relates to one or a few loci. Methods for selection signatures and GWAS can be used to identify targets for introgression or purging of deleterious loci. Genomic or marker-assisted selection can then be applied to increase/decrease the frequencies of the alleles associated with the trait in the population (e.g. Gaspa *et al.*, 2015). Genomics may also be used to decrease the proportion of “foreign” genes from a population that had been historically subjected to crossbreeding (Wellmann, Hartwig & Bennewitz,

2012). Finally, genomic characterization can be applied to individual crossbred animals to determine their genetic composition, potentially leading to improved management (see Box 12),

BOX 12

Estimation of ancestral breed proportions to improve management of admixed dairy cattle in developing countries

In many developing countries, imported semen has been used for decades to try to increase the milk production of local cattle populations. This process has increased productivity in many cases but has also tended to erode the genetic background of the original breeds (Leroy *et al.*, 2020). In many instances, the livestock keepers are unsure about the breed composition of their cattle (Manirakiza *et al.*, 2017). Assessment of the admixture among the breeds contributing to these populations allows the estimation of the proportion of ancestry from local and imported genetics for populations and individual animals. The technology can also be applied for determining parentage and evaluation of genetic diversity of the populations. Knowledge about the breed composition of individual animals can be used by farmers and other stakeholders to identify the best-performing genotypes in the local production environment and to inform the management and mating of individual animals. Recent studies have applied this approach in Africa (Gebrehiwot *et al.*, 2021) and India (Strucken *et al.*, 2021).

Conservation

Genomics may also contribute to reaching conservation objectives. As a first step, genomic estimates of breed-wise genetic distances (distinctiveness) or kinship (diversity) can be used to prioritize genetically valuable breeds for conservation (e.g. Ginja et *al.*, 2013). The methodology for these approaches was originally developed for simpler markers, such as microsatellites, but in general can be applied to genomic data. For a review of these approaches, see Boettcher *et al.* (2010). The genetic assessments may then be considered together with other factors such as risk of extinction and cultural significance for a comprehensive assessment of conservation priority (FAO, 2013).

Toro, Villanueva and Fernández (2014) reviewed the possible contributions of genomics for management of livestock conservation programmes. They discuss genomics-based approaches to evaluate genetic diversity of populations according to identity by descent, rather than identity by state. Approaches applied to *in vivo* populations may also be applied for the management of genebank collections. Ideally, *in vivo* and cryoconserved populations should be managed together. Genomics are currently being used to evaluate and manage populations within the Chinese national conservation programme for chickes, as described in Box 13.

BOX 13

Application of genome-wide SNPs in monitoring the impact of long-term *in situ – in vivo* conservation scheme on the genetic diversity of Chinese indigenous chicken breeds

Based on the latest national wide survey on animal genetic resources (CNCAGR, 2011), 114 indigenous chicken breeds, 81 improved chicken breeds/lines and 35 exotic chicken breeds/lines were identified and recorded in China (2020a). Twenty-eight of the indigenous chicken breeds have been recognized and included into the national directory for conservation of livestock and poultry genetic resources (2020b). Currently, two comprehensive national *ex situ - in vivo* conservation genebanks, one national cryoconservation genebank for somatic cells and DNAs, and 13 national *in situ – in vivo* conservation farms for individual breeds have been established and approved for the management and sustainable utilization of Chinese indigenous chicken genetic resources (China, 2014). To establish a conservation farm for an indigenous chicken breed, at least 300 breeding hens and 30 breeding cocks (each from a different family) are recommended to be maintained over generations (China, 2006). As it was recognized, 21 indigenous chicken breeds were at the risk of

extinction due to their very small breeding population sizes (CNCAGR, 2011); fortunately, at least four such breeds have been successfully rescued by the conservation measures (CNCAGR, 2016).

At one of national *ex situ - in vivo* conservation genebanks – the National Chicken Genetics Resources program established in Jiangsu, three breeds, Baier Yellow chicken (BEC), Beijing You chicken (BYC) and Langshan chicken (LSC), have been maintained since 1976, 1998 and 1998, respectively, among other Chinese indigenous chicken genetic resources. They began with a recommended flock size including 300 hens and 30 cocks, which have been continuously maintained with a family rotational mating regime. To evaluate the effectiveness and impact of this scheme on the genetic diversity of these populations, Zhang *et al*. (2018) performed a investigation using genome-wide SNPs for 270 birds from three generations per breed and 30 birds per generation (10 cocks and 20 hens). Up to 716 373 SNPs were screened. Results are shown in Figre BX. Both *Ho* and *He* were generally lower in these nine populations than other populations of chicken from around the world (Cendron *et al*., 2020; Luo *et al*., 2020; Malomane *et al*., 2021; Talebi *et al*., 2020; Zhang *et al*., 2020). Estimates of F_{IS} estimates based on either pedigree information or genome-wide SNPs have been gradually increasing over generations. The proportions of polymorphic SNPs have decreased in BEC and BYC The loss of genetic variability and the accumulation of inbreeding over generations was probably due to the rapid reduction of population sizes when the conservation lines were established. Therefore, a combination of both *ex situ* and *in situ - in vivo* conservation populations was recommended to maintain acceptable neutral and functional genetic diversity in the indigenous chicken breeds in the future.

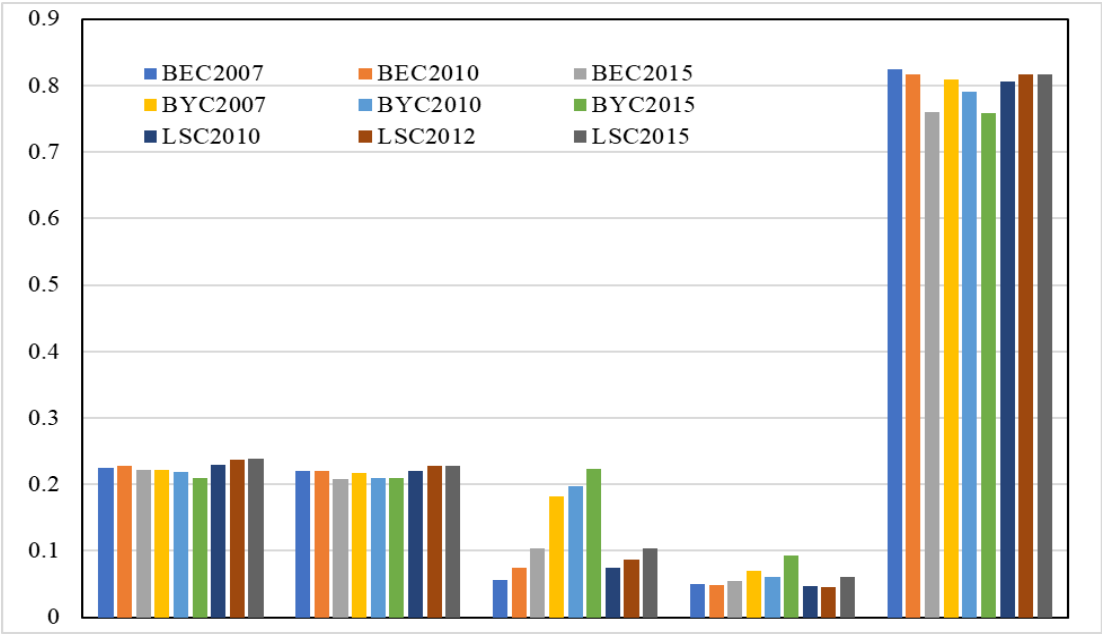

FIGURE B13

Changes in population genetic diversity parameters over generations of the three *in situ – in vivo* conserved Chinese indigenous chicken breeds. F_{IS} was estimated using pedigree records while F_{ROH} was measured by genome-wide SNPs (adapted from Zhang *et al*., 2018).

Source: Han Jianlin

REFERENCES

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467: 1061–1673. https://doi.org/10.1038/nature09534.

Ahrens C.W., Rymer P.D., Stow, A., Bragg, J., Dillon, S., Umbers, K.D.L. & Dudaniecet, RY. 2018. The search for loci under selection: trends, biases and progress. *Molecular Ecology*, 27: 1342– 1356.

Alachiotis, N. & Pavlidis, P. 2018. RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, 1: 79.

Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K. & Goudey, B. 2019. Exploring effective approaches for haplotype block phasing. *BMC Bioinformatics*, 20: 540. https://doi.org/10.1186/s12859-019-3095-8

Alberto, F.J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., de Villemereuil, P., Benjelloun, B. *et al.* 2018. Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, 9: 813. https://doi.org/10.1038/s41467-018-03206-y

Alexander, D.H., Novembre, J. & Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9). https://doi.org/10.1101/gr.094052.109

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7: 781–791. https://doi.org/10.1038/nrg1916

Bandelt, H.J. & Dress, A.W.M. 1992. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92: 47-105. https://doi.org/10.1016/0001-8708(92)90061-O

Barbato, M., Hailer, F., Orozco-terWengel, P., Kijas, J., Mereu, P., Cabras, P., Mazza, R., Pirastru, M. & Bruford, M.W. 2017. Genomic signatures of adaptive introgression from European mouflon into domestic sheep. *Scientific Reports*, 7: 7623. https://doi.org/10.1038/s41598-017-07382-7

Barbato, M., Orozco-terWengel, P., Tapio, M. & Bruford, M. W. 2015. SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics*, 6: 109. https://doi.org/10.3389/fgene.2015.00109.

Barbato, M., Reichel, M. P., Passamonti, M., Low, W. Y., Colli, L., Tearle, R., Williams, J. L. & Ajmone-Marsan, P. 2020. A genetically unique Chinese cattle population shows evidence of common ancestry with wild species when analysed with a reduced ascertainment bias SNP panel. *PLoS ONE*, 15: e0231162. https://doi.org/10.1371/journal.pone.0231162

Beaumont, M. A. 1999. Detecting population expansion and decline using microsatellites. *Genetics*, 153. https://doi.org/10.1046/j.1471-8286.2003.00351.x

Beaumont, M. A., Zhang, W. & Balding, D. J. 2002. Approximate Bayesian computation in population genetics. *Genetics*, 162: 2025–2035. https://doi.org/10.1093/genetics/162.4.2025.

Beerli, P. & Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, 98: 4563–4568.https://doi.org/10.1073/pnas.081068098.

Boettcher, P. J., Tixier-Boichard, M., Toro, M. A., Simianer, H., Eding, H., Gandini, G., Joost, S., Garcia, D., Colli, L., Ajmone-Marsan, P. & the GLOBALDIV Consortium. 2010. Objectives, criteria and methods for using molecular genetic data in priority setting for conservation of animal genetic resources. *Animal Genetics*, 41: 64–77. https://doi.org/10.1111/j.1365-2052.2010.02050.x.

Boitard, S., Rodríguez, W., Jay, F., Mona, S. & Austerlitz, F. 2016. Inferring population size history from large samples of genome-wide molecular data - an approximate Bayesian

computation approach. *PLoS Genetics*, 12: e1005877. https://doi.org/10.1371/journal.pgen.1005877.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S. & Sancristobal, M. 2010. Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, 186: 241–262. https://doi.org/10.1534/genetics.110.117275.

Bosse, M., Megens, H. J., Madsen, O., Frantz, L. A. F., Paudel, Y., Crooijmans, R. P. M. A. & Groenen, M. A. M. 2014. Untangling the hybrid nature of modern pig genomes: A mosaic derived from biogeographically distinct and highly divergent *Sus scrofa* populations. *Molecular Ecology*, 23: 4089–4102. https://doi.org/10.1111/mec.12807.

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., Sahana, G. *et al.* 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature Genetics*, 50: 362–367. https://doi.org/10.1038/s41588-018-0056-5.

The Bovine HapMap Consortium. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, 324: 528–532. https://doi.org/10.1126/science.1167936.

Bradburd, G. S., Coop, G. M. & Ralph, P. L. 2018. Inferring continuous and discrete population genetic structure across space. *Genetics*, 210: 33–52.https://doi.org/10.1534/genetics.118.301333.

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J.G. & Bustamante, C.D. 2012. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology*, 84: 343–364. https://doi.org/10.3378/027.084.0401

Browning, S. R. & Browning, B. L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81: 1084–1097. https://doi.org/10.1086/521987.

Browning, S.R. & Browning, B.L. 2010. High-resolution detection of identity by descent in unrelated individuals. *American Journal of Human Genetics*, 86: 526–539. https://doi.org/10.1016/j.ajhg.2010.02.021.

Bruford, M. W., Ginja, C., Hoffmann, I., Joost, S., Wengel, P. O., Alberto, F. J., Amaral, A. J. *et al.* 2015. Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. *Frontiers in Genetics*, 6: 314. https://doi.org/10.3389/fgene.2015.00314.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & Roychoudhury, A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29: 1917–1932. https://doi.org/10.1093/molbev/mss086.

Caballero, A., Rodriguez-Ramilo, S. T., Avila, V. & Fernández, J. 2010. Management of genetic diversity of subdivided populations in conservation programmes. *Conservation Genetics*, 11: 409–419. https://doi.org/10.1007/s10592-009-0020-0.

Caye, K., Jumentier, B., Lepeule, J. & François, O. 2019. LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies, *Molecular Biology and Evolution*, 36: 852–860. https://doi.org/10.1093/molbev/msz008

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. 2018. Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, 19: 220–234. https://doi.org/10.1038/nrg.2017.109.

Cendron, F., Perini, F., Mastrangelo, S., Tolone, M., Criscione, A., Bordonaro, S., Iaffaldano, N., Castellini, C., Marzoni, M., Buccioni, A., Soglia, D., Schiavone, A., Cerolini, S., Lasagna, E. & Cassandro, M. 2020. Genome-wide SNP analysis reveals the population structure and the conservation *status* of 23 Italian chicken breeds. *Animals*, 10: 1441. https://doi.org/10.3390/ani10081441

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M. & Lee, J. J. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4: 7. https://doi.org/10.1186/s13742-015-0047-8.

Charlesworth, B., Morgan, M. T. & Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134: 1289–1303. https://www.genetics.org/content/134/4/1289.

Charlier, C., Coppieters, W., Rollin, F., Desmecht, D., Agerholm, J. S., Cambisano, N., Carta, E. *et al.* 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics,* 40: 449–454. https://doi.org/10.1038/ng.96.

Chen, N., Cai, Y., Chen, Q., Ran, L., Wang, K., Huang, Y., Hu, S. *et al.* 2018**.** Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nature Communications*, 9: 2337. https://doi.org/10.1038/s41467-018-04737-0

Chikhi, L., Bruford, M. W. & Beaumont, M. A. 2001. Estimation of admixture proportions: A likelihood-based approach using Markov Chain Monte Carlo. *Genetics*, 158(3): 1347–1362. https://doi.org/10.1093/genetics/158.3.1347.

China. 2006. *Departmental Regulations-Management Measures for the Protection Areas and Gene Banks of Livestock and Poultry Genetic Resources* [online - Chinese] Beijing [Cited 25 July 2021]. http://www.zzj.moa.gov.cn/flfg/202009/t20200903_6351442.htm

China. 2014. *The General Office of the Ministry of Agriculture issued the "National Broiler Genetics Notice of Improvement Plan (2014-2025)* [online - Chinese] Beijing [Cited 25 July 2021]. http://www.moa.gov.cn/nybgb/2014/dsiq/201712/t20171219_6110244.htm

China. 2020a. *Ministry of Agriculture and Rural Affairs publishes National Directory of Genetic Resources for Livestock and Poultry* [online – Chinese]. Beijing. [Cited 25 July 2021]. http://www.gov.cn/xinwen/2020-05/29/content_5515954.htm

China. 2020b. *The list of breeds to be conserved* [online - Chinese] Beijing [Cited 25 July 2021]. http://www.nahs.org.cn/gk/tz/202101/P020210115497616621650.pdf

China National Commission of Animal Genetic Resources (CNCAGR). 2011a. *Animal Genetic Resources in China – Poultry*. China Agriculture Press, Beijing.

China National Commission of Animal Genetic Resources (CNCAGR). 2011b. *National "Twelfth Five-Year Plan" for the Protection and Utilization of Livestock and Poultry Genetic Resources* [online - Chinese] Beijing [Cited 25 July 2021]. http://www.moa.gov.cn/govpublic/XMYS/201611/t20161111_5360757.htm

China National Commission of Animal Genetic Resources (CNCAGR). 2016. *National "Thirteenth Five-Year Plan" for the Protection and Utilization of Livestock and Poultry Genetic Resources* [online - Chinese] Beijing [Cited 25 July 2021]. http://www.nahs.org.cn/xxcm/zybhly/201905/t20190516_339678.htm

Ciani, E., Mastrangelo, S., Da Silva, A., Marroni, F., Ferenčaković, M., Ajmone-Marsan, P., Baird, H. *et al.* 2020. On the origin of European sheep as revealed by the diversity of the Balkan breeds and by optimizing population-genetic analysis tools. *Genetics Selection Evolution*, 52: 25. https://doi.org/10.1186/s12711-020-00545-7.

Colli, L. & the AdaptMap Consortium. 2018. Genome-wide SNP profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes. *Genetics Selection Evolution*, 50: 58. https://doi.org/10.1186/s12711-018-0422-x

Corander, J., Waldmann, P. & Sillanpää, M. J. 2003. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163: 367–374. https://doi.org/10.1093/genetics/163.1.367

Cosgrove, E. J., Sadeghi, R., Schlamp, F., Holl, H. M., Moradi-Shahrbabak, M., Miraei-Ashtiani, Salma Abdalla, S. R., Shykind, B., *et al.* 2020. Genome diversity and the origin of the arabian horse. *Scientific Reports,* 10: 1–13. https://doi.org/10.1038/s41598-020-66232-1.

Costilla, R., Kemper, K.E., Byrne, E.M., Porto-Neto, L.R., Carvalheiro, R., Purfield, D.C., Doyle, J.L., Berry, D.P., Moore, S.S., Wray, N.R. & Ben J. Hayes 2020. Genetic control of temperament traits across species: association of autism spectrum disorder risk genes with cattle temperament. *Genetics Selection Evolution*, 52: 51. https://doi.org/10.1186/s12711-020-00569-z

CRU. 2021. Climatic Research Unit. [online]. University of East Anglia, Norwich, United Kingdom. [Cited 18 March 2021]. https://lr1.uea.ac.uk/cru/cruhome.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. & François, O. 2010. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, 25: 410–418. https://doi.org/10.1016/j.tree.2010.04.001.

Curik, I., Ferenčaković, M. & Sölkner, J. 2014. Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livestock Science*, 166: 26–34. https://doi.org/10.1016/j.livsci.2014.05.034

Da Fonseca, R. R., Ureña, I., Afonso, S., Pires, A. E., Jørsboe, E., Chikhi, L. & Ginja, C. 2019. Consequences of breed formation on patterns of genomic diversity and differentiation: The case of highly diverse peripheral Iberian cattle. *BMC Genomics*, 20: 334. https://doi.org/10.1186/s12864-019-5685-2.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., Liao, X., *et al.* 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 46, 858–865. https://doi.org/10.1038/ng.3034.

Daly, K.G., Delser, P. M., Mullin, V.E., Scheu, A., Mattiangeli, V., Teasdale, M.D., Hare, A.J., *et al.* 2018. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. *Science*, 361: 85–88. https://doi.org/10.1126/science.aas9411.

Daly, K.G., Mattiangeli, V., Hare, A.J., Davoudi, H., Fathi, H., Zeder, M.A. & Bradley, D.G. 2021. Herded and hunted goat genomes from the dawn of domestication in the Zagros Mountains. *Proceedings of the National Academy of Sciences*, 118: e2100901118. http://doi.org 10.1073/pnas.2100901118 /

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G. & Durbin, R. 2011. The variant call format and VCFtools. *Bioinformatics*, 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Davenport, K.M., Hiemke, C., McKay, S.D., Thorne, J.W., Lewis, R.M., Taylor, T. & Murdoch, B.M. 2020. Genetic structure and admixture in sheep from terminal breeds in the United States. Animal Genetics, 51: 284-291. https://doi.org/10.1111/age.12905

Decker, J.E., McKay, S.D., Rolf, M.M., Kim, J.W., Alcalá, A.M., Sonstegard, T.S., Hanotte, O. *et al.* 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genetics*, 10: e1004254. https://doi.org/10.1371/journal.pgen.1004254.

Delaneau, O., Marchini, J. & Zagury, J. F. 2012. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9: 179–181. https://doi.org/10.1038/nmeth.1785.

de Meeûs, T. & Goudet, J. 2007. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infection, Genetics Evolution*, 7: 731–735. https://doi.org/10.1016/j.meegid.2007.07.005.

Deng, J., Xie, X-L., Wang, D-F., Zhao, C., Lv, F-H., Li, X., Yang, J., *et al.* 2020. Paternal origins and migratory episodes of domestic sheep. *Current Biology,* 30: 4085–4095. https://doi.org/10.1016/j.cub.2020.07.077.

de Simoni Gouveia J. J., da Silva M. V., Paiva S. R., de Oliveira S. M. 2014. Identification of selection signatures in livestock species. *Genetics and Molecular Biology*, 37: 330–42. https://doi.org/10.1590/S1415-47572014000300004.

DeWan, A., Liu, M., Hartman, S., Zhang, S.S.-M., Liu, D.T.L., Zhao, C., Tam, P.O.S. *et al.* 2006. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*, 314: 989–992. https://doi.org/10.1126/science.1133807.

Doekes, H. P., Veerkamp, R. F., Bijma, P., de Jong, G., Hiemstra, S. J. & Windig, J. J. 2019. Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein-Friesian dairy cattle. *Genetics Selection Evolution*, 51: 514. https://doi.org/10.1186/s12711-019-0497-z.

Dopazo, J., Dress, A. & Von Haeseler, A. 1993. Split decomposition: A technique to analyze viral evolution. *Proceedings of the National Academy of Sciences*, 90: 10320–10324. https://doi.org/10.1073/pnas.90.21.10320.

Dress, A., Huson, D. & Moulton, V. 1996. Analyzing and visualizing sequence and distance data using SplitsTree. *Discrete Applied Mathematics*, 71: 95–105. https://doi.org/10.1016/S0166-218X(96)00059-5.

Druet, T. & Gautier, M. 2017. A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Molecular Ecology*, 26: 5820–5841. https://doi.org/10.1111/mec.14324.

Ellegren, H. 2009. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics*, 25: 278–284. https://doi.org/10.1016/j.tig.2009.04.005

Engelhardt, B. E. & Stephens, M. 2010. Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics*, 6: e1001117. https://doi.org/10.1371/journal.pgen.1001117.

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., Mason, B. A. & Goddard, M. E. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 95: 4114–4129. https://doi.org/10.3168/jds.2011-5019.

European Cattle Genetic Diversity Consortium. 2005. Marker-assisted conservation of European cattle breeds: an evaluation. Animal Genetics, 37: 475-481. https://doi.org/10.1111/j.1365-2052.2006.01511.x

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9: e1003905. https://doi.org/10.1371/journal.pgen.1003905.

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. https://www.genetics.org/content/131/2/479.

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., Constantz, C.M. et al. 2019. Tracking five millennia of horse management with extensive ancient genome time series. *Cell*, 177: 1419–1435.e31. https://doi.org/https://doi.org/10.1016/j.cell.2019.03.049.

Falconer D. S. & MacKay T. F. C. 1996. *Introduction to Quantative Genetics* (4th Edition). Longman, Essex.

Falush, D., Stephens, M. & Pritchard, J. K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164: 1567–1587. https://www.genetics.org/content/164/4/1567.

Fan, R., Gu, Z., Guang, X., Marín, J. C., Varas, V., González, B. A., Wheeler, J. C. *et al.* 2020. Genomic analysis of the domestication and post-Spanish conquest evolution of the llama and alpaca. *Genome Biology*, 21: 159. https://doi.org/10.1186/s13059-020-02080-6.

FAO. 2007. Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration. Rome (available at http://www.fao.org/docrep/010/a1404e/a1404e00.htm).

FAO. 2013. *In vivo Conservation of Animal Genetic Resources*. FAO Animal Production and Health Guidelines. No. 14. Rome (available at http://www.fao.org/3/i3327e/i3327e00.htm).

Fariello, M. I., Boitard, S., Naya, H., San Cristobal, M. & Servin, B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193: 929–941. https://www.genetics.org/content/193/3/929.

Felius, M. 1995. *Cattle Breeds: An Encyclopedia*. C Misset bv, Doetinchem, The Netherlands.

Felius, M., Theunissen, B., & Lenstra, J.A. 2015. Conservation of cattle genetic resources: the role of breeds. *The Journal of Agricultural Science*, 153: 152-162. http://doi.org/10.1017/S0021859614000124.

Felkel, S., Wallner, B., Chuluunbat, B., Yadamsuren, A., Faye, B., Brem, G., Walzer, C. & Burger, P. A. 2019. A first Y-chromosomal haplotype network to investigate male-driven population dynamics in domestic and wild bactrian camels. *Frontiers in Genetics*, 10: 423. https://doi.org/10.3389/fgene.2019.00423.

Fonseca, R. R., Ureña, I., Afonso, S., Pires, A. E. & Jørsboe, E. 2018. Consequences of breed formation on patterns of genomic diversity and differentiation: The case of highly diverse peripheral Iberian cattle. *BMC Genomics,* 20: 334. https://doi.org/10.1186/s12864-019-5685-2.

Forutan, M., Ansari Mahyari, S., Baes, C., Melzer, N., Schramm Schenkel, F. & Sargolzaei, M. 2018. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. *BMC Genomics,* 19: 98. https://doi.org/10.1186/s12864-018-4453-z.

François, O., Ancelet, S. & Guillot, G. 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, 174: 805–816. https://doi.org/10.1534/genetics.106.059923.

François, O. & Jay, F. 2020. Factor analysis of ancient population genomic samples. *Nature Communications*, 11 : 4661. https://doi.org/10.1038/s41467-020-18335-6.

Frankham, R., Ballou, J. D., Briscoe, D. A. & McInnes, K. H. 2002. *Introduction to Conservation Genetics*. Cambridge University Press.

Frantz, L.A.F., Haile, J., Lin, A.T., Scheu, A., Geörg, C., Benecke, N., Alexander, M., *et al.* 2019. Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proceedings of the National Academy of Sciences of the United States of America*, 116: 17231–17238. https://doi.org/10.1073/pnas.1901169116.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. 2014. Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196: 973–983. https://doi.org/10.1534/genetics.113.160572.

Fuller, Z. L., Mocellin, V. J. L., Morris, L. A., Cantin, N., Shepherd, J., Sarre, L., Peng, J. *et al.* 2020. Population genetics of the coral *Acropora millepora*: Toward genomic prediction of bleaching. *Science*, 369: eaba4674. https://doi.org/10.1126/science.aba4674.

Gaspa, G., Veerkamp, R. F., Calus, M. P. L. & Windig, J. P. L. 2015. Assessment of genomic selection for introgression of polledness into Holstein Friesian cattle by simulation. *Livestock Science*, 179: 86–95. https://doi.org/10.1016/j.livsci.2015.05.020.

Gautier, M., Moazami-Goudarzi, K., Levéziel, H., Parinello, H., Grohs, C., Rialle, S., Kowalczyk, R. & Flori, L. 2016. Deciphering the wisent demographic and adaptive histories from individual whole-genome sequences. *Molecular Biology and Evolution*, 33: 2801–2814. https://doi.org/10.1093/molbev/msw144.

Gebrehiwot, N.Z., Strucken, E.M., Marshall, K., Aliloo, H. & Gibson, J. 2021. SNP panels for the estimation of dairy breed proportion and parentage assignment in African crossbred dairy cattle. *Genetics Selection Evolution*, 53: 21. https://doi.org/10.1186/s12711-021-00615-4

Georges, M., Charlier, C. & Hayes, B. 2019. Harnessing genomic information for livestock improvement. *Nature Reviews Genetics*, 20: 135–156. https://doi.org/10.1038/s41576-018-0082-2.

Ginja, C., Gama, L.T., Cortés, O., Burriel, I.M., Vega-Pla, J.L., Penedo, C., Sponenberg, P. *et al.* 2019. The genetic ancestry of American Creole cattle inferred from uniparental and autosomal genetic markers. *Scientific Reports*, 9: 11486. https://doi.org/10.1038/s41598-019-47636-0.

Ginja, C., Gama, L. T., Cortes, O., Delgado, J. V., Dunner, S., García, D., Landi, V. *et al.* 2013. Analysis of conservation priorities of Iberoamerican cattle based on autosomal microsatellite markers. *Genetics Selection Evolution*, 45: 35. https://doi.org/10.1186/1297-9686-45-35.

Ginja, C., Da Gama, L. T. & Penedo, M. C. T. 2010. Analysis of STR markers reveals high genetic structure in Portuguese native cattle. *Journal of Heredity*, 101: 201–210. https://doi.org/10.1093/jhered/esp104.

Goddard, M. E. & Hayes, B. J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Review Genetics*, 10: 381–391. https://doi.org/10.1038/nrg2575,

Griffiths, R. C. & Marjoram, P. 1997. An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution.* Springer. New York.

Griffiths, R. C. & Tavaré, S. 1994. Simulating probability distributions in the coalescent. *Theoretical Population Biology*, 46: 131–159. https://doi.org/10.1006/tpbi.1994.1023.

Groeneveld, L.F., Lenstra, J.A., Eding, H., Toro, M.A., Scherf, B., Pilling, D., Negrini, R. *et al.* 2010. Genetic diversity in farm animals – a review. *Animal Genetics*, 41 Suppl 1: 6–31. https://doi.org/10.1111/j.1365-2052.2010.02038.x.

Grünwald, N. J., Everhart, S. E., Knaus, B. J. & Kamvar, Z. N. 2017. Best practices for population genetic analyses. *Phytopathology*, 107: 1000–1010. https://doi.org/10.1094/PHYTO-12-16-0425-RVW.**Guillot, G., Mortier, F. & Estoup, A.** 2005. GENELAND: A computer package for landscape genetics. *Molecular Ecology Notes*, 5: 712–715. https://doi.org/10.1111/j.1471-8286.2005.01031.x.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5: e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Haller, B. C. & Messer, P. W. 2019. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, 36: 632–637. https://doi.org/10.1093/molbev/msy228.

Han, H., Wallner, B., Rigler, D., MacHugh, D. E., Manglai, D. & Hill, E. W. 2019. Chinese Mongolian horses may retain early domestic male genetic lineages yet to be discovered. *Animal Genetics*, 50: 399–402. https://doi.org/10.1111/age.12780.

Hanghøj, K., Renaud, G., Albrechtsen, A. & Orlando, L. DamMet: Ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage, *GigaScience*, 8: giz025. https://doi.org/10.1093/gigascience/giz025.

Hayes, B. J. & Daetwyler, H.D. 2019. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual Review of Animal Biosciences*, 7: 89-102.

Hayes, B.J., Pryce, J., Chamberlain, A. J., Bowman, P. J. & Goddard, M. E. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics*, 6: e1001139. https://doi.org/10.1371/journal.pgen.1001139.

Hedrick, P. W. 1998. Balancing selection and MHC. *Genetica*, 104: 207–214. https://doi.org/10.1023/A:1026494212540.

Hein, J. Schierup, M. H. & Wiuf, C. 2003. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, Oxford.

Herbots, H. M. J. D. 1994. *Stochastic Models in Population Genetics: Genealogy and Genetic Differentiation in Structured Populations.* University of London, London.

Hivert, V., Sidorenko, J., Rohart, F., Goddard, M. E., Yang, J., Wray, N. R., Yengo, L. & Visscher, P. M. 2021. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *The American Journal of Human Genetics*, 108: 786–798. https://doi.org/10.1016/j.ajhg.2021.02.014.

Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics*, 3: e7. https://doi.org/10.1371/journal.pgen.0030007.

Horscroft C., Ennis S., Pengelly R. J., Sluckin, T. J. & Collins, A. 2019. Sequencing era methods for identifying signatures of selection in the genome. *Briefings in Bioinformatics*, 20: 1997–2008. doi: 10.1093/bib/bby064.

Howrigan, D. P., Simonson, M. A. & Keller, M. C. 2011. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics*, 12: 460. https://doi.org/10.1186/1471-2164-12-460.

Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, 9: 1322–1332. https://doi.org/10.1111/j.1755-0998.2009.02591.x.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7: 1–44.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18: 337–338. https://doi.org/10.1093/bioinformatics/18.2.337.

Huson, D. H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics*, 14: 68–73. https://doi.org/10.1093/bioinformatics/14.1.68.

The International Sheep Genomics Consortium. 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology*, 10: e1001258. https://doi.org/10.1371/journal.pbio.1001258

Jobling, M. A. & Tyler-Smith, C. 1995. Fathers and sons: The Y chromosome and human evolution. *Trends in Genetics*, 11: 449–56. https://doi.org/10.1016/S0168-9525(00)89144-1.

Jobling, M. A. & Tyler-Smith, C. 2017. Human Y-chromosome variation in the genome-sequencing era. *Nature Reviews Genetics*, 18: 485–497. https://doi.org/10.1038/nrg.2017.36.

Joo, J.W., Hormozdiari, F., Han, B. & Eskin, E. 2016. Multiple testing correction in linear mixed models. *Genome Biology*, 17: 62. https://doi.org/10.1186/s13059-016-0903-6

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C. & Eskin, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42: 348–354. https://doi.org/10.1038/ng.548.

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J.A.C., Baurain, D., Cambisano, N. *et al.* 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nature Genetics*, 43: 405–413. https://doi.org/10.1038/ng.814.

Keenan, K., Mcginnity, P., Cross, T. F., Crozier, W. W. & Prodöhl, P. A. 2013. DiveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, 4: 782–788. https://doi.org/10.1111/2041-210X.12067.

Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K. & McVean, G. 2019. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51: 1330–1338. https://doi.org/10.1038/s41588-019-0483-y.

Kijas, J.W., Lenstra, J.A., Hayes, B., Boitard, S., Neto, L.R., Cristobal, M.S., Servin, B. *et al.* 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture

and strong recent selection. *PLoS Biology*, 10: e1001258. https://doi.org/10.1371/journal.pbio.1001258.

Kim, J., Hanotte, O., Mwai, O.A., Dessie, T., Salim, B., Diallo, B., Agaba, M. *et al.* 2017. The genome landscape of indigenous African cattle. *Genome Biology*, 18: 34. https://doi.org/10.1186/s13059-017-1153-y.

Kingman J. F. C. 1982. The coalescent. *Stochastic Processes and Their Applications*, 13: 235–248. https://doi.org/10.1016/0304-4149(82)90011-4.

Kong, A., Frigge, M.L., Magnusson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A. *et al.* 2012. Rate of de novo mutations, father's age, and disease risk. *Nature*, 488: 471–475. doi: 10.1038/nature11396.

Kopelman, N.M., Stone, L., Gascuel, O. & Rosenberg, N.A. 2013. The behavior of admixed populations in neighbor-joining inference of population trees. *Pacific Symposium on Biocomputing*, 2013: 273–284. https://pubmed.ncbi.nlm.nih.gov/23424132/..

Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. 2014. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15: 356. https://doi.org/10.1186/s12859-014-0356-4.

Kristensen, T.N., Hoffmann, A.A., Pertoldi, C. & Stronen, A. V. 2015. What can livestock breeders learn from conservation genetics and vice versa? *Frontiers in Genetics*, 6: 38. https://doi.org/10.3389/fgene.2015.00038.

Larson, G. & Fuller, D.Q. 2014. The evolution of animal domestication. *Annual Review of Ecology, Evolution, and Systematics*, 45: 115-136. http://doi.org/10.1146/annurev-ecolsys-110512-135813

Laval, G., SanCristobal, M. & Chevalet, C. 2002. Measuring genetic distances between breeds: Use of some distances in various short term evolution models. *Genetics Selection Evolution*, 34: 481–507. http://doi.org/10.1186/1297-9686-34-4-481.

Lawal, R.A., Martin, S.H., Vanmechelen, K., Vereijken, A., Silva, P., Al-Atiyat, R.M., Aljumaah, R.S. *et al.* 2020. The wild species genome ancestry of domestic chickens. *BMC Biology*, 18: 13. https://doi.org/10.1186/s12915-020-0738-1.

Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. 2012. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8: e1002453. https://doi.org/10.1371/journal.pgen.1002453.

Lenstra, J.A., Ajmone-Marsan, P., Beja-Pereira, A., Bollongino, R., Bradley, D. G., Colli, L., De Gaetano, A., *et al.* 2014. Meta-analysis of mitochondrial DNA reveals several population bottlenecks during worldwide migrations of cattle. *Diversity*, 6: 178–187. https://doi.org/10.3390/d6010178.

Lenstra, J.A., Groeneveld, L.F., Eding, H., Kantanen, J., Williams, J.L., Taberlet, P., Nicolazzi, E.L. *et al.* 2012. Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. *Animal Genetics*, 43: 483–502. https://doi.org/10.1111/j.1365-2052.2011.02309.x.

Leppälä, K., Nielsen, S.V. & Mailund, T. 2017. Admixturegraph: An R package for admixture graph manipulation and fitting. *Bioinformatics*, 33: 1738–1740. https://doi.org/10.1093/bioinformatics/btx048.

Leroy, G., Boettcher, P., Besbes, B., Peña, C.R., Jaffrezic, F. & Baumung, R. Food securers or invasive aliens? Trends and consequences of non-native livestock introgression in developing countries. *Global Food Security*, 26: 100420. http://10.1016/j.gfs.2020.100420.

Li, H. & Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475: 493–496. https://doi.org/10.1038/nature10231.

Li, N. & Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165: 2213–2233. https://doi.org/10.1093/genetics/165.4.2213.

Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. 2021. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Research,* 31: 529-537. doi: 10.1101/gr.266486.120.

Librado P., Der Sarkissiana C., Erminia L., Schubert, M., Jonsson, H., Albrechtsen, A., Fumagalli, M. *et al.* 2015. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proceedings of the National Academy of Sciences of the United States of America*, 112: E6889-E6897. https://doi.org/10.1073/pnas.1513696112.

Lipson, M., Loh, P. R., Levin, A., Reich, D., Patterson, N. & Berger, B. 2013. Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular Biology and Evolution*, 30: 1788–802. https://doi.org/10.1093/molbev/mst099.

Lu, Y., Vandehaar, M.J., Spurlock, D.M., Weigel, K.A., Armentano, L.E., Connor, E.E., Coffey, M., *et al.* 2018. Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency. *Journal of Dairy Science*, 101: 3140–3154. https://doi.org/10.3168/jds.2017-13364

Luo, W., Luo, C., Wang, M., Guo, L., Chen, X., Li, Z., Zheng, M. *et al.* 2020. Genome diversity of Chinese indigenous chicken and the selective signatures in Chinese gamecock chicken. *Scientific Reports*, 10: 14532. https://doi.org/10.1038/s41598-020-71421-z

MacLeod, I. M., Larkin, D. M., Lewin, H. A., Hayes, B. J. & Goddard, M. E. 2013. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Molecular Biology and Evolution*, 30: 2209–2223. https://doi.org/10.1093/molbev/mst125.

Malomane, D. K., Weigend, S., Schmitt, A. O., Weigend, A., Reimer, C. & Simianer, H. 2021. Genetic diversity in global chicken breeds in relation to their genetic distances to wild populations. *Genetics Selection Evolution*, 53: 36. https://doi.org/10.1186/s12711-021-00628-z

Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storfer, A., Rosenberg, M. S., Scribner, K. T., Bonin, A. & Fortin, M. J. 2010. Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, 19: 3760–3772. doi: 10.1111/j.1365-294X.2010.04717.x..

Manel, S., Schwartz, M., Luikart, G. & Taberlet, P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18: 189–197. https://doi.org/10.1016/S0169-5347(03)00008-9.

Manirakiza, J. Hatungumukama, G. Thévenon, S. Gautier, M. Besbes, B. Flori, L. & Detilleux, J. 2017. Effect of genetic European taurine ancestry on milk yield of Ankole-Holstein crossbred dairy cattle in mixed smallholders system of Burundi highlands. *Animal Genetics*, 48: 544-550. https://doi.org/10.1111/age.12578

Martín-Burriel, I., Rodellar, C., Cañón, J., Cortés, O., Dunner, S., Landi, V., Martínez-Martínez, A., Gama, L. T., Ginja, C., Penedo, M. C. T., Sanz, A., Zaragoza, P. & Delgado, J. V. 2011. Genetic diversity, structure, and breed relationships in Iberian cattle. *Journal of Animal Science*, 89: 893–906. https://doi.org/10.2527/jas.2010-3338

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. S. & Tassell, C. P. V. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE*, 4: e5350. https://doi.org/10.1371/journal.pone.0005350.

Maynard-Smith, J.M. & Haigh, J. 1974. The hitchhiking effect of a favourable gene. *Genetics Research*, 23: 23–35. doi: https://doi.org/10.1017/S0016672300014634.

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. 2016. On the importance of being structured: Instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity*, 116: 362–371. https://doi.org/10.1038/hdy.2015.104.

McManus, C., Hermuche, P., Paiva, S.R., Melo, C.B. & Mendes, C.Q. Geographical distribution of sheep breeds in Brazil and their relationship with climatic and environmental factors as risk classification for conservation. *Brazilian Journal of Science and Technology*, 1: 3. https://doi.org/10.1186/2196-288X-1-3

McVean, G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5: e1000686. https://doi.org/10.1371/journal.pgen.1000686.

Mdladla, K., Dzomba, E. F. & Muchadeyi, F. C. 2018. Landscape genomics and pathway analysis to understand genetic adaptation of South African indigenous goat populations. *Heredity,* 120: 369–378. https://doi.org/10.1038/s41437-017-0044-z.

Mei, C., Wang, H., Liao, Q., Wang, L., Cheng, G., Wang, H., Zhao, C. *et al.* 2018. Genetic architecture and selection of Chinese cattle revealed by whole genome resequencing. *Molecular Biology and Evolution*, 35: 688–699. https://doi.org/10.1093/molbev/msx322.

Meiklejohn, C.D., Landeen, E.L., Gordon, K.E., Rzatkiewicz, T., Kingan, S.B., Geneva, A.J., Vedanayagam, J.P., Muirhead, C. A., Garrigan, D., Stern, D. L. & Presgraves, D. C. 2018. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. *eLife*, 7: e35468. https://doi.org/10.7554/eLife.35468.

Menozzi, P., Piazza, A. & Cavalli-Sforza, L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science*, 201: 786–792. https://doi.org/10.1126/science.356262.

Mészáros, G., Boison, S. A., Pérez O'Brien, A. M., Ferencakovic, M., Curik, I., Da Silva M. V. B., Utsunomiya, Y. T., Garcia, J. F. & Sölkner, J. 2015. Genomic analysis for managing small andendangered populations: A case studyin Tyrol Grey cattle. *Frontiers in Genetics*, 6: 173. doi: 10.3389/fgene.2015.00173.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157: 1819–1829.https://www.genetics.org/content/157/4/1819.

Meyermans, R., Gorssen, W., Buys, N. & Janssens, S. 2020. How to study runs of homozygosity using plink? A guide for analyzing medium density snp data in livestock and pet species. *BMC Genomics*, 21: 94. https://doi.org/10.1186/s12864-020-6463-x.

Michalakis, Y. & Excoffier, L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, 142: 1061–1064. https://www.genetics.org/content/142/3/1061.

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R. & Visscher, P. M. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genetics*, 11: e1004969. https://doi.org/10.1371/journal.pgen.1004969.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70: 3321–3323. https://doi.org/10.1073/pnas.70.12.3321.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89(3): 583–590. https://www.genetics.org/content/89/3/583.

Nei, M. & Li, W. H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76: 5269–5273. https://doi.org/10.1073/pnas.76.10.5269.

Neuditschko, M., Khatkar, M. S. & Raadsma, H. W. 2012. NetView: A high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. *PLoS ONE*, 7: e48375. https://doi.org/10.1371/journal.pone.0048375.

Neuditschko, M., Raadsma, H. W., Khatkar, M. S., Jonas, E., Steinig, E. J., Flury, C., Signer-Hasler, H., Frischknecht, M., Von Niederhäusern, R., Leeb, T. & Rieder, S. 2017.

Identification of key contributors in complex population structures. *PLoS ONE*, 12: e0177638. https://doi.org/10.1371/journal.pone.0177638.

Nielsen, R. & Wakeley, J. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics*, 158: 885–896. https://www.genetics.org/content/158/2/885.

Nosková, A., Hiltpold, M., Janett, F., Echtermann, T., Fang, Z.-H., Sidler, X., Selige, C., Hofer, A., Neuenschwander, S. & Pausch, H. 2020. Infertility due to defective sperm flagella caused by an intronic deletion in DNAH17 that perturbs splicing. *Genetics*, 217: iyaa033. https://doi.org/10.1093/genetics/iyaa033.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M. & Bustamante, C. D. 2008. Genes mirror geography within Europe. *Nature*, 456: 98–101. https://doi.org/10.1038/nature07331.

Novembre, J. & Stephens, M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40: 646–649. https://doi.org/10.1038/ng.139.

Obšteter, J., Jenko, J., Hickey, J. M. & Gorjanc, G. 2019. Efficient use of genomic information for sustainable genetic improvement in small cattle populations. *Journal of Dairy Science*, 102: 9971–9982. https://doi.org/10.3168/jds.2019-16853.

Oleksyk T. K., Smith M. W. & O'Brien S. J. 2010. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B Biological Sciences*, 365: 185–205. doi: 10.1098/rstb.2009.0219.

Oldenbroek, K. 2017. *Genomic Management of Animal Genetic Diversity.* Wageningen, Wageningen Academic Publishers.

OMIA. 2021. *Online Mendelian Inheritance in Animals, OMIA* [online]. Sydney. [Cited 10 March 2021]. https://www.omia.org/home.

Orozco-terWengel, P., Barbato, M., Nicolazzi, E., Biscarini, F., Milanesi, M., Davies, W., Williams, D., Stella, A., Ajmone-Marsan, P. & Bruford, M. W. 2015. Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Frontiers in Genetics*, 6: 191. https://doi.org/10.3389/fgene.2015.00191.

Paim, T.P., Paiva, S.R., de Toledo, N.M., Yamaghishi, M.B., Carneiro, P.L.S., Facó, O., de Araújo, A.M., Azevedo, H.C., Caetano, A.R., Braga, R.M. & McManus, C. 2021. Origin and population structure of Brazilian hair sheep breeds. *Animal Genetics*, 52: 492-504. https://doi.org/10.1111/age.13093

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. & Reich, D. 2012. Ancient admixture in human history. *Genetics*, 192: 1065–1093. https://doi.org/10.1534/genetics.112.145037.

Patterson, N., Price, A. L. & Reich, D. 2006. Population structure and eigen analysis. *PLoS Genetics*, 2: e190. https://doi.org/10.1371/journal.pgen.0020190.

Pausch, H., Flisikowski, K., Jung, S., Emmerling, R., Edel, C., Götz, K.-U. & Fries, R. 2011. Genome-wide association study identifies two major loci affecting calving ease and growth-related traits in cattle. *Genetics*, 187: 289–297. https://doi.org/10.1534/genetics.110.124057.

Pausch, H., MacLeod, I. M., Fries, R., Emmerling, R., Bowman, P. J., Daetwyler, H. D. & Goddard, M. E. 2017. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution*, 49: 24. https://doi.org/10.1186/s12711-017-0301-x.

Peripolli, E., Munari, D. P., Silva, M. V. G. B., Lima, A. L. F., Irgang, R. & Baldi, F. 2017. Runs of homozygosity: current knowledge and applications in livestock. *Animal Genetics*, 48: 255–271. https://doi.org/10.1111/age.12526.

Peter, B. M. 2016. Admixture, population structure, and f-statistics. *Genetics*, 202: 1485–1501. https://doi.org/10.1534/genetics.115.183913.

Pickrell, J. K. & Pritchard, J. K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, 8: e1002967. https://doi.org/10.1371/journal.pgen.1002967.

Piry, S., Alapetite, A., Cornuet, J. M., Paetkau, D., Baudouin, L. & Estoup, A. 2004. GENECLASS2: A software for genetic assignment and first-generation migrant detection. *Journal of Heredity*, 95: 536–539. https://doi.org/10.1093/jhered/esh074.

Pitt, D., Bruford, M. W., Barbato, M., Orozco-terWengel, P., Martínez, R. & Sevane, N. 2019. Demography and rapid local adaptation shape Creole cattle genome diversity in the tropics. *Evolutionary Applications*, 12: 105–122. https://doi.org/10.1111/eva.12641.

Plassais, J., Kim, J., Davis, B. W., Karyadi, D. M., Hogan, A. N., Harris, A. C., Decker, B., Parker, H. G. & Ostrander, E. A. 2019. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature Communications*, 10: 1489. https://doi.org/10.1038/s41467-019-09373-w.

Pool, J. E. & Nielsen, R. 2007. Population size changes reshape genomic patterns of diversity. *Evolution*, 61: 3001-3006. https://doi.org/10.1111/j.1558-5646.2007.00238.x.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38: 904–909. https://doi.org/10.1038/ng1847.

Pryce. J. E., Haile-Mariam, M., Goddard, M. E. & Hayes, B. J. 2014. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. *Genetics Selection Evolution*, 46: 71. https://doi.org/10.1186/s12711-014-0071-7.

Pritchard, J. K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155: 945–959. https://www.genetics.org/content/genetics/155/2/945.full.pdf.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J. & Sham, P. C. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81: 559-575. https://doi.org/10.1086/519795.

Purfield, D. C., Berry, D. P., McParland, S. & Bradley, D. G. 2012. Runs of homozygosity and population history in cattle. *BMC Genetics*, 13: 70. https://doi.org/10.1186/1471-2156-13-70.

Ramljak, J., Bunevski, G., Bytyqi, H., Marković, B., Brka, M., Ivanković, A., Kume, K. *et al.* 2018. Conservation of a domestic metapopulation structured into related and partly admixed strains. *Molecular Ecology*, 27: 1633–1650. https://doi.org/10.1111/mec.14555.

Raudsepp, T., Finno, C. J., Bellone, R. R. & Petersen, J. L. 2019. Ten years of the horse reference genome: Insights into equine biology, domestication and population dynamics in the post-genome era. *Animal Genetics*, 50: 569–597. doi: 10.1111/age.12857.

Raymond, B., Yengo, L., Costilla, R., Schrooten, C., Bouwman, A.C., Hayes, B.J., Veerkamp, R.F. and Visscher, P. 2020. Using prior information from humans to prioritize genes and gene-associated variants for complex traits in livestock. *PLoS Genet*ics, 16: e1008780. https://doi.org/10.1371/journal.pgen.1008780

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P. & Estoup, A. 2019. ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35: 1720–1728. https://doi.org/10.1093/bioinformatics/bty867.

Rice, T.K., Schork, N.J. & Rao, D.C. 2008. Methods for handling multiple testing. *Advances in Genetics*, 60: 293-308. http://doi.org/10.1016/S0065-2660(07)00412-9

Ruane, J. 2000. A framework for prioritizing domestic animal breeds for conservation purposes at the national level: A Norwegian case study. *Conservation Biology*, 14: 1385–1393. https://www.jstor.org/stable/2641791.

Rubinacci, S., Ribeiro, D. M., Hofmeister, R. & Delaneau, O. 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics,* 53: 120–126. https://doi.org/10.1038/s41588-020-00756-0.

Saitou, N. & Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4: 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Sánchez-Molano, E., Pong-Wong. R. & Banos G 2016. Genomic-based optimum contribution in conservation and genetic improvement programs with antagonistic fitness and productivity traits. *Frontiers in Genetics*, 7: 25. https://doi.org/10.3389/fgene.2016.00025.

Saravanan, K. A., Panigrahi, M., Kumar, H. & Bhushan, B. 2020. Selection signatures in livestock genome: A review of concepts, approaches and applications. *Livestock Science*, 241: 104257. https://doi.org/10.1016/j.livsci.2020.104257.

Schiffels, S. & Durbin, R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46: 919–925. https://doi.org/10.1038/ng.3015.

Schoenebeck, J. J. & Ostrander, E. A. 2013. The genetics of canine skull shape variation. *Genetics*, 193: 317–325. https://doi.org/10.1534/genetics.112.145284.

Signer-Hasler, H., Flury, C., Haase, B., Burger, D., Simianer, H., Leeb, T. & Rieder, S. 2012. A Genome-wide association study reveals loci influencing height and other conformation traits in horses. *PLoS ONE*, 7: e37282. https://doi.org/10.1371/journal.pone.0037282.

Silva, B.D., Castro, E.A., Souza, C.J., Paiva, S.R., Sartori, R., Franco, M.M., Azevedo, H.C., Silva, T.A., Vieira, A.M., Neves, J.P. & Melo, E.O. 2011. A new polymorphism in the Growth and Differentiation Factor 9 (GDF9) gene is associated with increased ovulation rate and prolificacy in homozygous sheep. *Animal Genetics*, 42: 89-92. http://doi.org/10.1111/j.1365-2052.2010.02078.x

Speidel, L., Forest, M., Shi, S. & Myers, S. R. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51: 1321–1329. https://doi.org/10.1038/s41588-019-0484-x.

Stephan, W. 2019. Selective sweeps. *Genetics*, 211: 5–13. https://doi.org/10.1534/genetics.118.301319.

Strucken, E.M., Gebrehiwot, N.Z., Swaminathan, M., Joshi, S., M. Al Kalaldeh, M. & Gibson, J. 2021. Genetic diversity and effective population sizes of thirteen Indian cattle breeds. *Genetics Selection Evolution*, 53: 47. https://doi.org/10.1186/s12711-021-00640-3

Taberlet, P., Valentini, A., Rezaei, H. R., Naderi, S., Pompanon, F., Negrini, R. & Ajmone-Marsan, P. 2008. Are cattle, sheep, and goats endangered species? *Molecular Ecology*, 17: 275–284. https://doi.org/10.1111/j.1365-294x.2007.03475.x.

Tang, H., Coram, M., Wang, P., Zhu, X. & Risch, N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *American Journal of Human Genetics*, 79: 1–12. https://doi.org/10.1086/504302.

Toro, M. A. & Caballero, A. 2005. Characterization and conservation of genetic diversity in subdivided populations. *Philosophical Transactions of The Royal Society B Biological Sciences*, 360: 1367–1378. https://doi.org/10.1098/rstb.2005.1680.

Toro, M. A., Fernandez, J. & Caballero, A. 2009. Molecular characterization of breeds and its use in conservation. *Livestock Science*, 120: 174–195. https://doi.org/10.1016/j.livsci.2008.07.003.

Toro, M., Villanueva, B. & Fernandez, J. 2014. Genomics applied to management strategies in conservation programmes. *Livestock Science*, 166: 48–53. https://doi.org/10.1016/j.livsci.2014.04.020.

Tzeng, J., Lu, H. & Li, W. H. 2008. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, 9: 179. https://doi.org/10.1186/1471-2105-9-179.

Upadhyay, M., Bortoluzzi, C., Barbato, M., Marsan, P.A., Colli, L., Ginja, C., Sonstegard, T. S., Bosse, M., Lenstra, J. A., Groenen, M. A. M. & Crooijmans, R. P. M. A. 2019. Deciphering the patterns of genetic admixture and diversity in southern European cattle using genome-wide SNPs. *Evolutionary Applications*, 12: 951–963. https://doi.org/10.1111/eva.12770.

Utsunomiya, Y.T., Pérez O'Brien, A.M., Sonstegard, T.S., Sölkner, J., & Garcia, J.F. 2015. Genomic data as the "hitchhiker's guide" to cattle adaptation: tracking the milestones of past selection in the bovine genome. *Frontiers in Genetics*, *6*: 36. https://doi.org/10.3389/fgene.2015.00036

Vajana, E., Barbato, M., Colli, L., Milanesi, M., Rochat, E., Fabrizi, E., Mukasa, C., *et al.* 2018**.** Combining landscape genomics and ecological modelling to investigate local adaptation of indigenous Ugandan cattle to East Coast Fever. *Frontiers in Genetics*, 9: 385. https://doi.org/10.3389/fgene.2018.00385.

Van Belleghem, S. M., Baquero, M., Papa, R., Salazar, C., McMillan, W. O., Counterman, B. A., Jiggins, C. D. & Martin, S. H. 2018. Patterns of Z chromosome divergence among *Heliconius* species highlight the importance of historical demography. *Molecular Ecology*, 27: 3852–3872. https://doi.org/10.1111/mec.14560.

van den Berg, I., Xiang, R., Jenko, J., Pausch, H., Boussaha, M., Schrooten, C., Tribout, T., Gjuvsland, A. B., Boichard, D., Nordbø, Ø., Sanchez, M.-P. & Goddard, M. E. 2020. Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. *Genetics Selection Evolution*, 52: 37. https://doi.org/10.1186/s12711-020-00556-4.

Van Oven, M., Van Geystelen, A., Kayser, M., Decorte, R. & Larmuseau, M. H. 2014. Seeing the wood for the trees: A minimal reference phylogeny for the human Y chromosome. *Human Mutation*, 35: 187–191. https://doi.org/10.1002/humu.22468.

Venn, O., Turner, I., Mathieson, I., De Groot, N., Bontrop, R. & McVean, G. 2014. Strong male bias drives germline mutation in chimpanzees. *Science*, 344: 1272–1275. https://doi.org/10.1126/science.344.6189.1272.

Verdugo, M.P., Mullin, V.E., Scheu, A., Mattiangeli, V., Daly, K.G., Maisano Delser, P., Hare. A.J., *et al.* 2019. Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science,* 365: 173–176. https://doi.org/10.1126/science.aav1002.

Wakeley, J. 1999. Nonequilibrium migration in human history. *Genetics*, 153: 1863–1871. https://www.genetics.org/content/153/4/1863.

Wallner, B., Palmieri, N., Vogl, C., Rigler, D., Bozlak, E., Druml, T., Jagannathan, V,*et al.* 2017. Y chromosome uncovers the recent oriental origin of modern stallions. *Current Biology*, 27: 2029–2035.e5. https://doi.org/10.1016/j.cub.2017.05.086.

Wallner, B., Vogl, C., Shukla, P., Burgstaller, J. P., Druml, T. & Brem, G. 2013. Identification of genetic variation on the horse Y chromosome and the tracing of male founder lineages in modern breeds. *PLoS ONE*, 8: e60015. https://doi.org/10.1371/journal.pone.0060015.

Wang, K., Lenstra, J. A., Liu, L., Hu, Q., Ma, T., Qiu, Q. & Liu, J. 2018. Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. *Communications Biology*, 1: 169. https://doi.org/10.1038/s42003-018-0176-6.

Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffieer, L. 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, 11:116. https://doi.org/10.1186/1471-2105-11-116

Weir, B. S. & Cockerham, C. C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution*, 38: 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x.

Weldenegodguad, M., Popov, R., Pokharel, K., Ammosov, I., Ivanova, Z. & Kantanen, J. 2019. Whole-genome sequencing of three native cattle breeds originating from the northernmost cattle farming regions. *Frontiers in Genetics*, 9: 728. doi: 10.3389/fgene.2018.00728.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447: 661–678. https://doi.org/10.1038/nature05911.

Wellmann, R. Hartwig, S. & J. Bennewitz, J. 2012. Optimum contribution selection for conserved populations with historic migration. *Genetics Selection Evolution*, 44: 34. https://doi.org/10.1186/1297-9686-44-34.

Wilson Sayres, M. A. 2018. Genetic diversity on the sex chromosomes. *Genome Biology and Evolution*, 10: 1064–1078. https://doi.org/10.1093/gbe/evy039.

WorldClim. 2020. *WorldClim.* [online]. [Cited 20 March 2021]. https://www.worldclim.org/.

Wright, S. 1951. The genetical structure of populations. *Annals of Eugenics*, 15: 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.

Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. 2011. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88: 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Yurchenko, A. A., Daetwyler, H. D., Yudin, N., Schnabel R. D., Vander Jagt, C. J., Soloshenko, V., Lhasaranov, B., Popov, R., Taylor, J. F. & Larkin, D. M. 2018. Scans for signatures of selection in Russian cattle breed genomes reveal new candidate genes for environmental adaptation and acclimation. *Scientific Reports*, 8: 12984. doi: 10.1038/s41598-018-31304-w.

Xiang, R., Berg, I. van den, MacLeod, I.M., Hayes, B.J., Prowse-Wilkins, C.P., Wang, M., Bolormaa, S. *et al.* 2019. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences of the United States of America*, 116: 19398–19408. https://doi.org/10.1073/pnas.1904159116.

Zeder, M.A. 2015. Core questions in domestication research. *Proceedings of the National Academy of Sciences of the United States of America*, 112: 3191-3196. https://doi.org/10.1073/pnas.1501711112

Zhang, M., Han, W., Tang, H., Li, G., Zhang, M., Xu, R., Liu, Y., Yang, T., Li, W., Zou, J. & Wu, K. 2018. Genomic diversity dynamics in conserved chicken populations are revealed by genome-wide SNPs. *BMC Genomics*, 19: 598. https://doi.org/10.1186/s12864-018-4973-6

Zhang, Q., Guldbrandtsen, B., Bosse, M., Lund, M. S. & Sahana, G. 2015. Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics*, 16: 524. https://doi.org/10.1186/s12864-015-1715-x.

SECTION 5

Conclusions and recommendations

CONCLUSIONS AND RECOMMENDATIONS

The FAO and the ISAG/FAO Advisory Group on Animal Genetic Diversity recommend that:

1. Countries continue to characterize their animal genetic resources (AnGR) to provide critical information for their improved management and to contribute to implementation of the Global Plan of Action for Animal Genetic Resources. Genetic characterization should apply the most advanced methods feasible (currently single nucleotide polymorphisms (SNP) and whole-genome sequencing (WGS)). For most current applications in genomic characterization, medium density SNP panels will provide sufficient information. Genomic data should be complemented by phenotypic characterization and description of the production environment and recording of geographic coordinates.

2. National genetic resources are investigated by or in close collaboration with researchers from the same country while: (i) respecting the wishes and interests of the livestock keepers; and (ii) complying with legislative, administrative and policy measures addressing access to genetic resources and the sharing of benefits derived from their utilization, as applicable; (iii) complying with measures addressing access to traditional knowledge associated with genetic resources and the sharing of benefits arising from the utilization of such knowledge with indigenous peoples and local communities; and (iv) complying with measures addressing access to genetic resources that are held by indigenous peoples and local communities and the sharing of benefits arising from the utilization of such resources with the peoples and communities concerned.

3. FAO, National Coordinators and any National Advisory Committees for AnGR are made aware of all diversity projects at any geographic level, so that results can contribute to the planning and development of national conservation and sustainable use activities and so that FAO can help facilitate coordination among projects, exchange information and promote funding. Contact information for National Coordinators can be found at http://www.fao.org/dad-is/national-coordinators/en/.

4. Locally adapted genomic resources are considered within an international context, which for marker-based molecular genetic studies imply the use of the same or overlapping sets of genetic markers employed in previous studies, and that, if appropriate, data collected from characterization studies are placed in open-access repositories, including raw data, associated metadata and analysis pipelines and procedures.

5. Studies are guided by one or more specific questions to be answered, recognizing that these may be expanded during data analysis. In addition to establishing the genetic relationships of populations on the basis of neutral genetic markers, the molecular variation involved in phenotypic variation should be searched for, with special emphasis on environmental adaptation and disease susceptibility or resistance.

6. Studies of autosomal markers are complemented by the characterization of mitochondrial DNA sequences as markers representing the maternal lineages, and of Y-chromosomal variation as markers of paternal lineages.

7. Studies of the present patterns of genetic diversity of livestock are complemented, where relevant, by the analysis of ancient DNA samples in order to provide a historic context.

8. Genotype data are subjected to a thorough quality check and filtering in order to identify and remove unreliable data, including duplicates, mislabelled samples and outliers associated with breed-level differentiation or undocumented crossbreeding.

9. For all breeds studied, heterozygosities or other indicators of overall genetic variability are calculated to identify breeds with low diversity and anticipate the potential for bias towards these breeds in multi-breed analyses.

10. Researchers understand the statistical methods being applied in data analysis and critically evaluate the results obtained from computer software.

11. Parameters related to within-breed genetic diversity are not simply estimated, but that this information be complemented with that of clustering methods and demographic modelling to better

understand evolutionary processes. By combining different approaches, genomics can provide more informed management of populations by monitoring of within-breed genetic variation and crossbreeding, avoiding high levels of inbreeding and ultimately preserving adaptive variation.

12. Publications of molecular diversity studies are freely accessible.

13. The scientific progress is communicated to breeders, livestock industry, the relevant government agencies and the general public.

14. Results from genomic characterization studies are actively utilized in the monitoring and management of AnGR. Standardized estimates of effective population sizes or other measures of genetic variation could complement data on population size for informing countries on the genetic status and risk of extinction of their national breed populations. More research is, however, required to determine the most appropriate measure of genetic variation. Management of AnGR may include purebred genetic improvement and conservation measures, as well as controlled crossbreeding that preserves the original populations and maintains environmental adaptation.

15. The advances in molecular technology and bioinformatics continue to be closely monitored and relevant innovations be communicated to the AnGR community.

Appendices

Appendix 1

Glossary of technical terms

Approximate Bayesian Calculation (ABC) Modelling approach based on optimizing the agreement between simulated and measured values of summary statistics.

Admixture The presence of genetic influence of multiple breeds in the genome of an individual or a population, due to historical interbreeding.

Allelic richness Measure of genetic diversity obtained by simply counting the number of different alleles within a breed or other population of interest. Allelic richness among populations is only directly comparable if the number of animals genotyped per population is equal.

AMOVA Analysis of molecular variance, estimation of the portioning of diversity over different hierarchical levels: within breeds, among breeds within regions, between regions, etc.

ARMS Amplification refractory mutation system, a polymerase chain reaction (PCR) during which one of the primer covers a single nucleotide polymorphism (SNP) site in such a way that the amplification depend on the SNP allele in the template DNA.

Ascertainment bias (AB) Systematic distortion in estimates of molecular genetic parameters (such as allelic frequencies) due to irregularities in the process used to identify the markers. For instance, many SNP in large panels were selected according to their high minor allele frequency in international transboundary breeds and can thus underestimate the relative diversity in other breeds.

Assembly The process of arranging overlapping individual DNA sequence reads into a contiguous whole-genome sequence.

Autosomes All chromosomes with the exception of the mammalian X or Y or the avian W and Z sex chromosomes.

Bayesian analysis Estimation of a likelihood distribution of model parameters on the basis of the likelihoods of parameter values in the absence of data (the prior) and the likelihoods of the observed data given different values of the model parameters. These estimations depend on a specific model and are often achieved by a strategy (like the Multiple Chain Monte Carlo simulations) to explore different plausible values of the parameters (the "parameter space").

Bead array Silica bead-based microarray for high-density SNP genotyping.

Coalescence analysis Estimation of the divergence times of individual DNA sequences since their descendance from a hypothetical most recent common ancestor, often used to infer present and past effective population sizes.

CNV Copy number variation, a type of structural variation in the genome resulting from differences in the copy number of chromosomal fragments of up to several megabases in length. CNV can be used as a genetic marker and has been associated with differences in human phenotypes.

Contig Uninterrupted ("contiguous") DNA sequence assembled by combining separate experimental sequence reads.

CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeat/ CRISPR-associated (Cas) endonuclease 9) Method for highly specific and rapid modification of DNA in a genome, leading to a new biotechnology revolution called "genome editing".

Crossing over Mechanism of creating new genetic variability through recombination of chromosomes during meiosis. The newly formed oocyte or sperm chromosome combines segments of the two homologous chromosomes originating from the mother and the father of the individual in which crossing-over takes place.

DGGE Denaturing gradient gel electrophoresis, which generates for a PCR product containing a SNP site a pattern that depends on the SNP alleles.

Depth the approximate number of times that a chromosomal region will be sequenced through a whole-genome procedure that involves sequencing and subsequent assembly of segments.

Diploid Characteristic of a species defined by its genome containing of two copies of all autosomes, the maternal and the paternal copies.

Effective population size (N_e) Hypothetical population size that would generate observed values of diversity parameters for a given population if mated randomly and not subject to forces such as selection and migration. The N_e corresponds to the number of breeding animals per generation and is usually smaller than the actual population count. It may be calculated separately for males and females.

Fixation Retainment within a populations of only one allele for a give genetic marker which thus becomes homozygous, typically the effect of inbreeding reducing the genetic diversity.

Fixation index (F_{ST}) The proportional increase in homozygosity that occurs through population subdivision (for example, the creation of breeds).

Genetic distance A measure of the genetic differences between two populations (or species) calculated on the basis of allelic frequencies in both populations/species.

Genetic marker Site in the genome that is variable (polymorphic) within a species. The different variants are called alleles, such as the two (usually) different nucleotides of a SNP.

Genome-wide association study (GWAS) Statistical analysis to detect a correlation between the phenotypes of a group of individuals and their genotypes of a given SNP, calculated for 50 000 to 1 000 000 SNPs spanning the whole genome; values exceeding a significance threshold may indicate that the SNP is near the causative gene.

Genotyping by sequencing (GBS) Sequencing of a subset (reduced representation) of the genome, either by a given class of restriction-enzyme fragments or fragments captured by hybridization to a set of DNA fragments. Sequencing of more or less the same fragments in a panel of animals yields SNPs and at the same time their genotypes. This allows genome-wide SNP genotyping for species for which no bead-array has been developed, such as wild species.

Haplotype Combination of alleles of two or more genetic markers on the same DNA segment. Haplotypes defined by alleles of markers on mitochondrial DNA and the non-recombining part of the Y-chromosome are transmitted to offspring. Autosomal and X-chromosomal haplotypes, typically defined by markers in or near the same gene (in which case a haplotype represents a gene variant), are transmitted to the offspring only if there is no recombination between the markers.

Haplogroup Group of closely related haplotypes.

Haploid Genetic material containing a single copy of the genome, such as sperm cells and oocytes.

Hardy-Weinberg equilibrium (HWE) Ratio for a given marker and population of the numbers of homozygote and heterozygote genotypes as predicted by random mating in a large population in the absence of selection, migration and mutation.

Heterogametic sex The gender that carries two different sex chromosomes, such as XY mammalian males and WZ avian females.

Heterozygosity State of a genetic marker (or any locus) at which both alleles are different.

HDP High-density panel of SNPs on a bead array, typically containing > 500 000 SNPs.

Homogametic sex The gender that carries two identical sex chromosomes, such as XX mammalian females and ZZ avian males.

Homozygosity State of a marker (or any locus) at which both alleles are identical.

Identity by descent (IBD) Identical genotypes of a given marker in two or more individuals because the individuals descend from the same ancestor.

Identity by state (IBS) Identical genotypes of a given marker in two or more individuals of a population.

Imputation Interpolation of greater-density genotypes (or full sequences) for a given (group of) individual(s) on the basis of low-density data for the same individuals and greater-density data for a panel of related individuals.

Indel Acronym for "inserts and/or deletions" of small amounts of sequence (<50 basepairs) in the genome. Larger insertions and deletions are usually considered to be "structural variation".

Introgression Movement of a particular allele or set of alleles from one population (i.e. breed) to another, usually by either deliberate crossbreeding or casual contact between neighbouring populations.

KASP Competitive allele specific polymerase chain reaction, a unique competitive allele-specific PCR combined with a fluorescence-based reporting system for the identification of genetic variation at the nucleotide level, including SNPs or insertions/deletions.

Kinship The probability that a randomly selected allele from two individuals (at the same locus) is identical by descent from a common ancestor (also known as the "coancestry" or "coefficient of coancestry").

Linkage disequilibrium (LD) Distribution of multilocus genotype combinations in a population for a given pair of markers that is incompatible with independent inheritance, thus indicating genetic linkage of the loci.

Linkage disequilibrium (LD) pruning Removal of SNP genotypes from a genome-wide SNP dataset in such a way that there is no LD between any pair of SNPs that surpasses a user-give threshold.

Linkage disequilibrium (LD) block region of the genome with high linkage disequilibrium (also known as a "haplotype block".

Locus A distinct region of DNA (often a gene) in the genome.

Meiosis Cell division in the germline leading to haploid sperm or oocyte cells.

Microsatellite Tandem DNA repeat of a 2 to 5 bp unit. In most cases, the repeat unit is the dinucleotide CA. The number of repeats of a given microsatellite locus is often polymorphic within populations, in which case the microsatellite may serve as a genetic marker. Also known as STR (simple tandem repeat) or SSR (simple sequence repeat).

Minor allele frequency (MAF) A metric primarily used to evaluate SNPs, corresponding to the frequency of the less common of the two alleles (SNPs are usually biallelic). A threshold of MAF $\geq$ 0.01 is sometimes considered to define a SNP. Genetic variability and information content of a SNP increases as MAF approaches 0.50 (i.e. the maximum value) and MAF is typically among the criteria for the selection of SNPs in commercial panels.

Mitochondrial DNA (mtDNA) The DNA contained in mitochondria, which is widely used in phylogenetic studies because of its variability, lack of recombination and maternal inheritance.

Neigbour joining (NJ) tree Phylogenetic tree constructed on the basis of a genetic-distance matrixfollowing the principle of identifying pairs of operational taxonomic units (OTUs = neighbours) to minimize the total branch length at each stage of clustering of the OTUs starting with a starlike tree.

Nucleotide Any of the four types of molecules that make up the structural units of DNA (and RNA). For DNA, these molecules are adenine, cytosine, guanine and thymine and are often denoted by their first letter (i.e. A, C, G, and T, respectively).

Phylogeny Evolutionary history of a taxonomic group in terms of successive divergence events and mutations or genetic drift.

PLINK A popular and powerful software package (an open-source C/C++ WGAS tool set) for handling and analyzing large SNP datasets.

Polymorphism The presence of at least two different genetic variants or alleles at a given locus.

Primer An oligonucleotide that serves as a starting point for DNA synthesis in PCR. The sequences of the primers define the start and the end of the DNA segment to be amplified (based on complementarity of the base-pair sequences) and during the PCR reaction bind on the template DNA (often genomic DNA).

Principal component analysis (PCA) A method for analysis of a set of variables, such as allele frequencies, by calculation of a new set of statistically independent coordinates that each corresponds to a weighted combination of the original variables in such a way that each coordinate captures as much variation in the original variables as possible. In many datasets, a small number of coordinates may explain a large proportion of the initial variability, thus increasing efficiency. Plotting the distribution of individuals or breeds in a graph of the first two or three coordinates allows for simple visualization of the pattern of diversity.

Quantitative trait locus (QTL) Locus contributing to the variation in a multigenic quantitative trait (such as milk production or carcass weight).

Read The sequence of an individual segment of chromosomal DNA that is obtained through a next generation sequencing procedure.

Resequencing Whole-genome sequencing approach assembled by mapping the reads to the genome sequence of a reference genome of the same species.

Runs of homozygosity (ROH) Stretches of DNA, typically of 1 to15 Mb, that are (almost) completely homozygous.

Scaffold Non-contiguous combination of contigs that are positioned related to each other on the basis of paired ends, mate pairs, optical mapping or other long-range assembly method.

Selection signature Locus with a deviating pattern of diversity, such as a local high fixation index value in comparison of breeds or a high frequency of a homozygous haplotype within a breed, that can be explained by selection acting on the locus.

Sex chromosomes Chromosomes for which the number of copies per cell (in diploid cells normally two) depends on the sex, such as the mammalian X chromosome (two in females, one on males), the mammaliam Y chromosome (one in males, zero in females), the avian W chromosome (zero in males, one in females) and the avian Z chromosome (two in males, one in females).

Single nucleotide polymorphism (SNP) Variation resulting from a point mutation and most often corresponding to a biallelic (having two different alleles) marker.

Sliding window a concept associated with the application of a given statistical test to a fixed-length segment of the genome (typically measured in either number of SNP or basepairs), which is repeatedly applied across the entire genome. The location of the "window" is progressively advanced by a chosen number of SNP or basepairs, and the location yielding the highest test of significance is assumed to contain the genes responsible the effects for which the test is being applied.

Structural variation DNA sequence variation based on copy number variations (CNV: deletions, duplications and large-scale copy number variants) and on insertions, inversions and translocations.

TaqMan assay A PCR-based SNP genotyping assay that uses TaqMan® 5′-nuclease chemistry for amplifying and detecting specific polymorphisms in purified genomic DNA samples. Each assay allows genotyping of individuals for a SNP. The PCR includes two probes, each specific for one of the two SNP alleles and each carrying a different fluorescent reporter(e.g., VIC or 6-FAM dye), bind to one of the strands of the PCR product depending on the alleles during amplification. A bound probe is degraded by the exonuclease activity of *Taq* polymerase to generate an allele-dependent fluorescent signal, allowing the discrimination of two different homozygous alleles and also the differentiation of homozygous from heterozygous alleles.

Transcriptomics the study of all RNA molecules present in a given sample. This may include only the messenger RNA (mRNA) which are translated to proteins and non-coding RNA.

Appendix 2

Example[11] Material Transfer Agreement

MATERIAL TRANSFER AGREEMENT (MTA)
for genetic material for genotyping

This Material Transfer Agreement is made by and between,

_____

Name of provider of genetic material ("Provider")

Mailing Address _____

Other contact information – i.e. telephone and fax numbers, email address

_____

and

_____

Name of recipient of genetic material ("Recipient").

Mailing Address _____

Other contact information – i.e. telephone and fax numbers, email address

_____

The parties have agreed as follows:

1. Provider agrees to transfer to Recipient the following (biological) material ("Material"):

Description of the genetic material including type (e.g. DNA, blood, tissue) amount (i.e. number of samples) and other information (e.g. means of preservation)

2. This Material will be used by Recipient solely in connection with the project described as follows:

Description of the project ("Research Project")*, including assays to be performed (e.g. genomic characterization using a panel of SNPs), use of the data, context in a larger project and project sponsors*.

3. This Material will only be used for research purposes by the Recipient in its laboratory. By requesting the material and signing this agreement, the Recipient is considered responsible for appropriate handling of the material and guarantees that suitable containment conditions are available and will be applied in the Recipient's laboratory. This Material will not be used for commercial purposes, such as the production or sale of products or services. Recipient will promptly, after termination of the Research Project, inform Provider of the results of the Research Project.

4. To the extent permitted by law, Recipient agrees to treat in confidence, for a period of *XXXX* years from the date of its disclosure, any of the Provider's written information about this

[11] Example kindly provided by the International Livestock Research Institute, Kenya, Nairobi.

Material that is stamped "CONFIDENTIAL" (hereinafter "Confidential Information"), except for information that was previously known to Recipient or that is or becomes publicly available through no fault of Recipient or which is lawfully disclosed to Recipient without a confidentiality obligation or that is independently developed by Recipient or its affiliated entities without the benefit of any disclosure by Provider. Recipient may publish or otherwise publicly disclose the results of the Research Project, provided that in all such oral presentations or written publications concerning the Research Project, Recipient will acknowledge Provider's contribution of this Material unless otherwise requested by Provider.

5. This Material is considered proprietary to Provider. Recipient therefore agrees to retain control over this Material, and further agrees not to transfer the Material to other parties not under its supervision without prior written consent of Provider. Provider reserves the right to distribute the Material to others and to use it for its own purposes. When the Research Project is completed, the Material will be disposed of as mutually agreed upon by Provider and Recipient.

6. This Material IS BEING SUPPLIED TO RECIPIENT WITH NO WARRANTIES, EXPRESS OR IMPLIED, INCLUDING ANY WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Provider makes no representations that the use of the Material will not infringe any patent or proprietary rights of third parties. Recipient agrees to hold harmless and indemnify Provider for all liabilities, demands, damages, expenses and losses arising out of or as a result of Recipient's use of the Material for any purpose.

7. Nothing in this Material Transfer Agreement shall or may be construed as granting Recipient any right or license to the Material for any use other or further than the evaluation described here above.

8. This Agreement shall be governed and construed in accordance with the laws of *(the country where the 'Research Project' was conducted*). All disputes arising out of or in connection with this Agreement shall be settled in first instance by the relevant court of *(the country where the 'Research Project' was conducted)*.

RECIPIENT	**PROVIDER**
Place:	Place:
Date:	Date:
By:	By:
Title:	Title:

Appendix 3

Sampling of blood for DNA

Description

This protocol describes how the collection of blood from the jugular vein into a vacutainer EDTA containing tubes should be organized. The actual sampling should be done by a veterinarian or someone with comparable qualifications valid in the country where the sampling is done.

Personnel

Qualified sampler (such as a veterinarian), and ideally at least two assistants (one to hold the animal and another to assist with blood sampling and recording)

Non-veterinary equipment

Labelled tubes for blood samples

Marker pen with permanent ink

Box or other container to store and transport the samples

Notebook, a pen and a permanent marker for documentation

Digital camera

To be done by the assistance

- Obey strictly instructions of the owner or his representative regarding hygienic precautions and allowable distance to the animals.
- Photograph the animal, especially relevant morphological traits and, if present, the labels with registration number.
- After receiving the tube, gently invert the tube 4-5 times to mix blood with EDTA and use the marker pen to write the animal ID, breed and date on the tube. Keep the tube at ambient temperature, but do not expose to the sun or other strong light. Preferably, extract DNA within a few days.

Remember to collect as much information about the animal as possible. Owner of the animal, Animal ID, breed, sampling site and date, age of the animal, animal origin and pedigree information as well as known, short description of appearance, major diseases (or lack of them) or other observations by the owner (see Appendix 4).

Appendix 4

Example questionnaire to be filled during sampling

Questions in **bold** are generally considered to be mandatory)

Animal code _____

Farm ID _____

Species code _____

AA = Anser anser domesitcus (greylag goose)

AC = Anser cygnoides (swan goose)

AP = Anas platyrhynchos (mallard duck)

BB = Bubalus bubalis (water buffalo)

BF = Bos frontalis (gayal)

BG = Bos grunniens (yak)

BI = Bos indicus (zebu)

BJ = Bos javanicus (banteng, Bali cattle)

BT = Bos taurus (taurine cattle)

CB = Camelus bactrianus (Bactrian camel, two-humped)

CD = Camelus dromedarius (dromedary, one-humped camel)

CH = Capra hircus (goat)

CM = Cairina moschata (muscovy duck)

EA = Equus asinus (donkey)

EC = Equus caballus (horse)

GG = Gallus gallus (chicken)

LG = Lama glama (Llama)

MG = Melea gallopavo (turkey)

OA = Ovis aries (sheep)

SS = Sus Scrofa (pig)

VP = Vicugna pacos (alpaca)

VV = Vicugna vicugna (vicuña)

Species name _____

Country _____

Number of the sample _____

Official Animal ID (if available) _____

Animal and sampling information

Sex of animal: female □ male □

Year of birth of the animal: _____(YYYY)

Place (locality) of birth of the animal: _____

Date of collection: _____ (DD.MM.YYYY)

Breed's full name: _____

Collector's name: _____

Collector's institution:_____

Address of the farm and telephone number (if available)

Country of the farm: _____

Province/county of the farm: _____

Region of the farm: _____

Closest town to the farm: _____

International phone code: _____(4 digits - ex: 0033, 0041, ...)

Area phone code: _____

Phone number: _____

Type of biological material: blood □ tissue □ hair □ other (specify) □

GPS coordinates _____

Appendix 5

Breed questionnaire

(To be completed once per breed – complete all questions that are relevant and for which information is available)

COMMON NAME OF BREED: _____

NAME OF SPECIES: _____

Transboundary or brand name				
Local breed name				
Main location				
Breed society?	Circle: Yes No Year established:			
Description of origin and development				
Population size	Year: N° of animals:			
N° of reproductive animals	Males in natural service			
	Males used for AI			
	Breeding females			
	Trend in breeding females	Increase □	Steady □	Decrease □
Females mated pure (%)				
Adult size (male/female)	Withers height (cm)		M	F
	Live weight (kg)			
N° of farmer/breeders				
Main uses (e.g. meat, milk)				
Typical management conditions	Type			
	Housing			
	Feeding			
Conservation activities	In situ: Y / N Ex situ: Y / N Cryo: Y / N			

Performance comparison	Relative to which breed:
much higher in:	(e.g. milk yield)
higher in:	
equal in:	
lower in:	
much lower in:	

Appendix 6

Software for genomic analysis

This is a (incomplete) list of genetic analysis programs with a short specification, grouped according to purpose. Although the software are grouped, many of them can be applied to multiple purposes. Most programs can be downloaded freely from the internet, along with detailed instruction manuals. If not given in this list, details, literature references and URL of programs can be found by search engines. Appearance on this list does not constitute an endorsement of the software by the Food and Agriculture Organization of the United Nations.

SOFTWARE FOR MANAGING AND/OR ANALYZING LARGE SNP DATASETS

AlphaPeel software package for calling, phasing, and imputing genotype and sequence data in pedigree populations. https://github.com/AlphaGenes/alphapeel

AlphaFamImpute genotype calling, phasing, and imputation algorithm for large full-sib families in diploid plants and animals which supports individuals genotyped with SNP array or GBS data https://github.com/AlphaGenes/alphafamimpute

ANGSD utility program for low to medium coverage WGS data

http://www.popgen.dk/angsd/index.php/ANGSD

diveRsity R package (http://www.r-project.org) that calculates a range of genetic diversity statistics. https://cran.r-project.org/web/packages/diveRsity/index.html

KING v 2.2.6 toolset to explore family relationship inference and pedigree error checking, quality control, and population substructure identification in genome-wide data. https://people.virginia.edu/~wc9c/KING/

PLINK 1.8 & 2.0 Clearly documented high-density SNP handling and analysis program. Requires special format. Outputs to formats compatible with many other applications. Performs a variety of data handling operations and calculations, such as allele-sharing between individuals and coordination analysis. https://www.cog-genomics.org/plink

SOFTWARE FOR HANDLING GENOMIC DATA

Samtools suite of programs for working with high-throughput sequencing data, particularly data stored in SAM format files. http://www.htslib.org

VCFtools utility program for working with VCF files (Variant call format, a common system for storing of genetic variation information)

https://vcftools.github.io/index.html

GENOTYPE BY SEQUENCING TOOLS

DECONVQC repository that contains scripts used to process and manage output sequence data and metadata related to Illumina hiseq and miseq machines, including running a number of GBS-specific Q/C steps for predominantlypredominately GBS-related sequencing output. This project is focused on immediate upstream Q/C and sequence delivery, rather than custom downstream analyses. https://github.com/AgResearch/DECONVQC

KGD kinship (genetic relatedness) using GBS (genotyping-by-sequencing) with depth adjustment. https://github.com/AgResearch/KGD

TASSEL software package used to evaluate traits associations, evolutionary patterns, and linkage disequilibrium. – it can work with genotyping-by-sequencing data. https://www.maizegenetics.net/tassel

See also a range of software described in tutorials for analysis of low-coverage whole genome sequencing data. https://github.com/nt246/lcwgs-guide-tutorial

PHASING AND IMPUTATION

AlphaImpute software that combines heuristics and HMM to perform both family-based and population-based imputation. Requires phased data, but presents the AlphaPhase algorithm embedded into the software for phasing. https://alphagenes.roslin.ed.ac.uk/wp/software-2/alphaimpute/

AlphaPhase software that performs genotype phasing and minor imputation through a heuristic algorithm that employs the concepts of surrogate parents and Erdös distance. https://alphagenes.roslin.ed.ac.uk/wp/software-2/alphaphase/

Beagle v4.1 (current v5.1 does not handle genotype likelihood input from sequencing) Pedigree-free imputation of genotyping-by-sequencing data. https://faculty.washington.edu/browning/beagle/b4_1.html

Eagle highly efficient and fast genotype phasing and minor imputation software that uses an HMM-based algorithm combined with a positional Burrows-Wheeler transform. https://alkesgroup.broadinstitute.org/Eagle/

fastPhase performs genotype phasing and minor imputation using HMM fitted via an EM algorithm. http://scheet.org/software.html

FImpute implements a deterministic algorithm that performs both family-based and population-based imputation. Does not require phased data. https://animalbiosciences.uoguelph.ca/~msargol/fimpute/

GLIMPSE phasing and imputation method for large-scale low-coverage sequencing studies https://odelaneau.github.io/GLIMPSE.

IMPUTE fast population-based imputation through HMM and positional Burrows-Wheeler transform. Requires phased data. https://jmarchini.org/impute5/

Minimac population-based imputation using HMM with state space reduction. Requires phased data. https://genome.sph.umich.edu/wiki/Minimac4

SHAPEIT fast genotype phasing and minor imputation via HMM and positional Burrows-Wheeler transform. https://odelaneau.github.io/shapeit4/

STITCH R package for Sequencing To Imputation Through Constructing Haplotypes https://github.com/rwdavies/STITCH

GENETIC DISTANCES, TREES AND PLOTS

Beast performs Bayesian MCMC analysis of molecular sequences, inferring rooted, time-measured phylogenies using strict or relaxed molecular clock models. Provides a framework for testing evolutionary hypotheses without conditioning on a single tree topology. http://beast.bio.ed.ac.uk/Main_Page

Mega calculation of a wide variety of population genetics statistics and convenient tree reconstruction program by the most common algorithms except the Bayesian method. http://www.megasoftware.net/

MrBayes command-line operated for handling nexus sequence files for Bayesian tree reconstructions. http://mrbayes.csit.fsu.edu/

neigborNet R package for construction of phylogenentic networks based on the Neighbor-Joining algorithm. https://www.rdocumentation.org/packages/phangorn/versions/2.5.5/topics/neighborNet

Network constructs Median-joining networks of haplotype data. Generates evolutionary trees and networks from genetic and other data. http://www.fluxus-engineering.com/sharenet.htm

Netview R package for network visualization of individuals of breeds using maximally k nearest neighbors https://github.com/esteinig/netview

Phylip classical command-line comprehensive package requiring its own file format for tree reconstruction according to the most common algorithms, but offering fewer options than PAUP. http://phylip.com

QPTools package within **AdmixTools** https://github.com/DReichLab/AdmixTools (or https://github.com/uqrmaie1/admixtools) and **ADMIXTUREGRAPH** an R package https://github.com/mailund/admixture_graph optimize the branch lengths and admixture proportions from Admixture graphs.

Relate estimates genome-wide genealogies in the form of trees, based on inferred haplotypes.

https://myersgroup.github.io/relate/

SNAPP uses a likelihood-based approach to estimate the "true" phylogenetic tree of populations based on the genetic variation of markers across the genome. Built upon the Beast software.

https://www.beast2.org/snapp

SplitsTree constructs neighbor-joining tree, SplitsTree graphs and NeighborNet graphs. Aceepts Nexus files. Many graphical output options. http://www-ab.informatik.uni-tuebingen.de/software/splitstree4/welcome.html

Treemix constructs maximum-likelihood trees based on the covariance of allelic frequencies. Attempt to infer the number of admixture events among the populations.

https://speciationgenomics.github.io/Treemix

Tsinfer develops a tree sequence based on inferred haplotypes.

https://github.com/tskit-dev/tsinfer

PRINCIPAL COMPONENT ANALYSIS

Eigensoft Analyses population structure by combining statistical genetics with principal components analysis (**Eigenstrat**) to explicitly model ancestry differences between cases and controls along continuous axes of variation http://genepath.med.harvard.edu/~reich/Software.htm

POPULATION ASSIGNMENT AND CLUSTER ANALYSIS

Admixture Produces results for unsupervised clustering (i.e., without prior population information) comparable or identical to **Structure** but faster by means of a more efficient algorithm. http://www.genetics.ucla.edu/software/admixture/

Baps comparable to Structure but with increased flexibility in the definitions of levels at which genetic structure may exist. http://web.abo.fi/fak/mnf/mate/jc/software/baps.html

Clumpp accepts the output of **Structure** or other clustering programs in order to align the output of different runs. http://rosenberglab.bioinformatics.med.umich.edu/clumpp.html

Distruct reads in tables of genomic components from the Structure output and files of options set by the user in order to provide graphical output of the Structure clustering. http://rosenberglab.bioinformatics.med.umich.edu/clumpp.html

fineStructure clustering program on the basis of ChromoPaint-based haplotype sharing

https://people.maths.bris.ac.uk/~madjl/finestructure/finestructure_info.html

Frappe estimates individual ancestry proportions - comparable to Admixture https://med.stanford.edu/tanglab/software/frappe.html

Geneland clustering program that can make use of both geographic and genetic informations to estimate the number of populations in a dataset and delineate their spatial organisation. http://www2.imm.dtu.dk/~gigu/Geneland/

Instruct joint inference of population structure and inbreeding rates, eliminating the assumption of Hardy-Weinberg equlilibrium and especially applicable in cases of self-fertilization or inbreeding. http://cbsuapps.tc.cornell.edu/InStruct.aspx

LEA R package - estimates admixture proportions similar to Structure or Admixture software, but with reduced computation time. Less affected by the inbreeding of the analysed populations. http://membres-timc.imag.fr/Olivier.Francois/LEA/software.htm

Spaida and Spaign assign individual animals to genetic clusters based on spatial autocorrelations. http://notendur.hi.is/~snaebj/programs.html

Structure popular and user-friendly program for an informative visualization of patterns of diversity. Reconstructs model-based subdivision of individual genotypes into a user-specified number of clusters (k) by optimizing of Hardy-Weinberg equilibrium and minimizing linkage disequilibrium within clusters; estimates proportion of individual genomes derived from the inferred clusters. Clusters may correspond to ancestral components, but also to a relatively homogeneous breed or group of breeds. Optionally allows for admixture, linkage between markers, codominant markers, incorporation of prior population information and incorporation of spatial information. It is substantially slower than Admixture but has more options. http://pritch.bsd.uchicago.edu/structure.html

Whichrun assigns individuals to populations based upon maximum likelihood theory http://www-bml.ucdavis.edu/whichrun.htm

IDENTIFICATION OF SELECTION SIGNATURES

cgaTOH detection of intra-population selection signatures based on runs of homozygosity.

https://github.com/hernanmd/cgaTOH.2018

HapFLK implements the the Lewontin and Krankakauer tests to identify inter-population signatures based on individual loci (FLK) or haplotypes (hapFLK). https://pypi.org/project/hapflk/

HierFstat R package for detection of inter-population selection signatures based on fixation indexes (F_{ST}). https://cran.r-project.org/web/packages/hierfstat/index.html

PCAdmix estimation of local ancestry by applying PCA based upon phased haplotypes. https://bio.tools/pcadmix

rehh performs tests based on extended hapotype homozyosity to detect genomic regions subject to selection within and across populations. http://cran.r-project.org/web/packages/rehh/index.html

SweeD identification of intra-population selection signatures based on the site frequency spectrum

https://cme.h-its.org/exelixis/web/software/sweed/

Sweep large-scale intra-population analysis of haplotype structure in genomes for the primary purpose of detecting evidence of natural selection. https://software.broadinstitute.org/mpg/sweep/

Selscan detection of intra-population selection based on extended haplotype homozygosity (EHH) and integrated haplotype score (iHS) methods. Detection of inter-population selection based on cross population extended haplotype homozygosity (XP-EHH) https://github.com/szpiech/selscan

LANDSCAPE GENOMICS

LFMM2 applies latent factor mixed models to test for association between genotypes and explanatory variables that can include environmental or geographical measures.

https://bcm-uga.github.io/lfmm/

BAYENV applies Bayesian method to estimate the pattern of covariance in allele frequencies between populations and then uses this as a null model for a test of environmental variables at individual SNPs.

https://bitbucket.org/tguenther/bayenv2_public/

BAYPASS identifies genetic markers subjected to selection or associated with population-specific covariates, including geographical and environmental variables.
http://www1.montpellier.inra.fr/CBGP/software/baypass/index.html

BAYESCENV performs F_{ST}-based genome scans to detect local adaptation.

https://github.com/devillemereuil/bayescenv

R.SamBada R-package providing a pipeline for landscape genomic analysis, spanning from the retrieval of environmental variables at sampling locations to gene annotation using the Ensembl genome browser. This application standardizes the landscape genomics pipeline, eases the search for candidate genes involved in adaptation processes, and enhances reproducibility of the studies.

https://github.com/SolangeD/R.SamBada

GENOME-WIDE ASSOCIATION STUDIES

bayesR Bayesian hierarchical model for complex trait analysis, including GWAS.

https://github.com/syntheke/bayesR

BOLT applies linear mixed-models for genetic association testing. Includes an application that applies an algorithm for restricted maximum-likelihood estimation of variance components.

https://alkesgroup.broadinstitute.org/BOLT-LMM/

EMMAX performs large-scale GWAS while correcting for population sampling structure.

https://genome.sph.umich.edu/wiki/EMMAX

FaST-LMM genome-wide association testing, genomic prediction, heritability analysis.

https://github.com/fastlmm/FaST-LMM/

GCTA estimates genomic relationships, performs association testing, heritability analysis and variance partitioning. https://cnsgenomics.com/software/gcta

GCTB a collection of Bayesian linear mixed models for complex trait analyses using genome-wide SNP data. Simultaneously estimates joint effects of the SNP and the background genetic architecture. https://cnsgenomics.com/software/gctb

GEMMA applies linear mixed models for association testing, multi-trait analyses, heritability analysis. https://github.com/genetics-statistics/GEMMA

JWAS performs routine single-trait and multi-trait genomic prediction and GWAS using Bayesian mixed effects models and either complete or incomplete genomic data.
https://reworkhow.github.io/JWAS.jl/latest/

LFMM2 fits latent factor mixed models, estimating factors based on an exact least squares approach.
https://rdrr.io/bioc/LEA/man/lfmm2.html (also suitable for landscape genomics)

SPECIAL PURPOSE PROGRAMS

AlphaAssign parentage assignment algorithm that works with SNP array and GBS data
https://github.com/AlphaGenes/alphaassign

Rannala software multiple packages that perform various specific functions, including LD mapping, data simulation, and detecting migration by using multilocus genotypes. http://www.rannala.org

SNeP performs estimations of effective population size trajectories over time by using genome-wide SNP data. https://sourceforge.net/projects/snepnetrends

SPAGeDi characterizes the spatial genetic structure of individuals or populations based on genetic marker data. Estimates genetic distance and other basic statistics.
http://ebe.ulb.ac.be/ebe/Software.html

SIMULATION, MODELLING AND PARAMETER ESTIMATION

ABC and **ABCRF** R packages that perform approximate Bayesian computation (ABC) for model selection and parameter inference, the latter by utilizing random forests
https://cran.r-project.org/package=abc and https://rdrr.io/cran/abcrf

ABCtoolbox a series of open-source programs that peform all of the steps of a standard ABC analysis. Offers the user the opportunity to choose among several algorithms.
https://bitbucket.org/wegmannlab/abctoolbox/wiki/Home

dadi simulates the joint frequency spectrum of genetic variation among several populations.

https://dadi.readthedocs.io/en/latest/

FastSimcoal2 flexible modelling and simulation software that allows comparison of alternative tree construction models.

http://cmpg.unibe.ch/software/fastsimcoal2/

PSMC https://github.com/lh3/psmc and **MSMC** https://github.com/stschiff/msmc and R packages for implementing the pairwise sequentially Markov coalescent and multiple sequentially Markov coalescent methods for modelling population histories.

MULTIPURPOSE PROGRAMS ORIGINALLY DEVELOPED FOR MICROSATELLITE DATA

Convert easy program that converts data from an Excel format to files suitable for other software (may not be able to handle large SNP files)

http://www.agriculture.purdue.edu/fnr/html/faculty/Rhodes/Students%20and%20Staff/glaubitz/software.htm

GenAlEx estimates of variability based on allele and genotypic frequencies, genetic distances, Principal Component Analysis, Formatting of data for other software. Runs as a Microsoft Excel addin. http://www.anu.edu.au/BoZo/GenAlEx/

Microsatellite Toolkit convenient Excel microsatellite data handling tool. Format requires sample names in which letters indicate breed and numerals the individual; easily transformed to Structure format. Exports to Microsat, Arlequin, GenePop and Fstat formats. Check errors in the dataset (missing figures, large gaps between alleles, non-unique sample labels, duplicate samples), converts two column per marker and one line per sample to one column per marker and two lines per

individual). Provides summary statistics (observed and expected heterozygosity, number of alleles) and allele frequencies. http://www.animalgenomics.ucd.ie/sdepark/ms-toolkit/

Appendix 7

Genotyping arrays available by species across different genotyping platforms[12]

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Buffalo	Affymetrix Axiom	Buffalo Genotyping Array	Medium	89 988	4 breeds (Mediterranean buffalo, Murrah, Jaffarabadi and Nili-Ravi)	Useful for biodiversity research, GWAS
Camelid - Bactrian	Affymetrix Axiom	Camelid-Multispecies Array	Medium	59 938	Bactrian populations from Mongolia and China	Useful for biodiversity research, GWAS, mapping studies
Camelid - Dromedary				59 958	Dromedary populations from Algeria, Ethiopia, Mauritania, Morocco, Sudan, UAE	Useful for biodiversity research, GWAS, mapping studies
Camelid - New World				60 000	Alpaca (Huacaya and Suri), Llama and Vicugna populations from Peru, Chile, Bolivia, Ecuador	Useful for biodiversity research, GWAS, mapping studies
Cattle	Illumina Infinium	Bovine LD v2.0-Genotyping Beadchip	Low	7 931	26 breeds (Angus, Ayrshire, Beefmaster, Blonde d' Aquitaine, Brahman, Brown Swiss, Charolais, Fleckvieh, Friesian, Gelbvieh, Guernsey, Hereford, Holstein, Jersey (US & Denmark), Limousin, Montbeliard, N'dama, Normande, Norwegian Red, Red Angus, Red Dairy (Angier), Red Danish (Denmark, Finland, Sweden), Santa Gertrudis)	Useful for genomic selection, Genotype Imputation, Parentage verification, biodiversity research; Includes 121 parentage markers

[12] Appearance on this list does not constitute an endorsement of the platform or array by the Food and Agriculture Organization of the United Nations.

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Cattle	Illumina Infinium	Bovine SNP50v3-Beadchip	Medium	53 218	17 breeds (Angus, Beefmaster, Gir, Nelore, Brahman, Charolais, Guernsey, Hereford, Holstein, Jersey, Limousin, N'dama, Piedontese, Red Angus, Romagnola, Santa Getrudis, Sheko)	Useful for genomic selection, GWAS, Parentage verification, biodiversity research; Includes 116 parentage markers
	Affymetrix Axiom	Bovine	Medium	54 560	10 breeds (Holstein, Angus, Jersey, Fleckvieh, Hereford, Limousin, Romagnola, Brahman, Nelore, Gir)	Useful for biodiversity research, GWAS, parentage verification; Consists of 191 ISAG core markers for bovine parentage
	Affymetrix Axiom	BovMDv3 Array	Medium	63 648	20 breeds; Bos taurus (Afrikander, Angus, Ayrshire, Boran, Blonde d'Aquitaine, Brown Swiss, Simmental, Hanwoo, Hereford, Holstein, Japanese Black, Jersey, Limousine, Norwegian Red, Rouge des Pres, Romagnola, Tuli); Bos indicus (Brahman, Nelore, Gir).	Useful for dairy evaluation (genomic selection), biodiversity research, copy number variations, parentage verification, markers for deleterious recessive traits, fertility traits and traceability; Consists of 13K Bos indicus SNPs; Includes several hundred SNPs that are optimized for short tandem repeat (STR) imputation

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Cattle	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Angus-GS-Beadchip	Medium	75 000	Angus	Breed specific SNP array for breeding and improvement of Angus cattle. Includes 22000 new SNP markers with 7800 from Angus sequences that are novel to all available arrays. Also includes markers associated with fertility, feed efficiency, marbling, calving ease and vaccine response. Has markers for detection of genetic conditions like arthrogryposis multiplex, neuropathic hydrocephalus contractural arachnodactyly, osteoporosis, oculocataneous hypopigmentation, developmental duplication, BVDV, M1, D2.
	Illumina Infinium	GGP Bovine 100K	Medium	100 000	Derived from the original Illumina BovineHD-Genotyping and Bovine SNP50 Beadchip	Useful for GWAS, genetic evaluations, biodiversity research and parentage verification; Has 85% overlap with the Council of Dairy Cattle Breeding (CDCB) virtual evaluation and 35000 SNPs overlap with the Canadian Dairy Network (CDN) Evaluation; Average imputation accuracy to the Illumina Bovine HD is >99.5% in Angus and Holstein populations

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Cattle	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Bovine 150K	Medium	150 000	NA	Useful for GWAS, genetic evaluations, biodiversity research, parentage, detection of markers for recessive disorders including fetal death and abnormalities that interfere with growth rate in cattle
	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Bovine-F250	High	235 000	NA	Useful for functional genomic studies in cattle, genomic selection, genotype imputation. Includes SNPs enriched for functional variants such as non-synonymous, frameshift and premature stop codons.
	Affymetrix Axiom	Bos1 Array	High	648 855	20 breeds; Bos taurus (Afrikander, Angus, Ayrshire, Boran, Blonde d'Aquitaine, Brown Swiss, Simmental, Hanwoo, Hereford, Holstein, Japanese Black, Jersey, Limousine, Norwegian Red, Rouge des Pres, Romagnola, Tuli); Bos indicus (Brahman, Nelore, Gir).	Useful for GWAS, genetic evaluations, biodiversity research and linkage disequilibrium studies

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Cattle	Illumina Infinium	BovineHD-Genotyping-Beadchip	High	777 962	20 breeds; Bos taurus (Angus, Bonde d'Aquitaine, Brown Swiss, Charolais, Guernsey, Hereford, Holstein, Jersey, Lagunaire, Limousin, Montbeliad, N'dama, Norwegian Red, Piedontese, Red Angus, Romagnola, Senepol, Simental, Wagyu); Bos indicus (Brahma, Gir, Nelore); Crossbreds (Beefmaster, Brangus, Santa Gertrudis, Sheko)	Useful for genomic selection, GWAS, QTL identification, biodiversity research, Mapping
Cattle - Zebu	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Indicus	Medium	35 090	Brahman, Guzera, Gyr, Nelore, Droughtmaster, Santa Gertrudis, Tropical Composite, etc.	Useful for biodiversity research, GWAS, QTL identification
Chicken	Affymetrix Axiom	ChickenHD Genotyping Array	High	580 961	Commercial broiler (Aviagen), commercial layers (Hyline, Synbreed), Experimental inbred (IAH line), Non-selected (J. Line Roslin)	Useful for genetic evaluation of layers and broilers, GWAS, mapping and biodiversity research; Includes markers associated with wild outbred lines
Goat	Illumina Infinium	Goat SNP50-Genotyping-Beadchip	Medium	50 000	Saanen, Alpine, Creole, Boer, Kacang, and Savanna	Useful for biodiversity research, GWAS, genome mapping
	Affymetrix Axiom	Ovicap Mutispecies Array		60 034	Milk and mixed breeds (Alpine, Saanen and Creole); Meat breeds (Boer, Katjang and Savanna); Norwegian dairy goat	Useful for biodiversity research, GWAS, parentage verification, mapping studies; Includes 583 markers for parentage testing and traceability

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Horse	Illumina Infinium	Equine SNP50	Medium	54 602	15 breeds (Anadalusian, Arabian, Belgian, Franches Montagnes, French Trotter, Hanoverians, Icelandic, Mongolian, Norwegian Fjord, Quarter Horse, Saddlebred, Standardbred, Swiss Warmblood, Thoroughbred, Hokkaido)	Useful for biodiversity research, GWAS, genomic selection, identification of QTL, mapping.
	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Equine		71 947	Derived from Illumina's Equine SNP50 Beadchip	Useful for biodiversity research, GWAS, genomic selection, identification of QTL, general mapping
	EquineHD Genotyping Array	Affymetrix Axiom	High	670 796	32 breeds (Including Arabian, Belgian, Black Forest, Duelmener, Edlblu-Haflinger, French Trotter, Haflinger, Hanoverian, Icelandic, Lusitano, Marremanno, Mongolian, Morgan, Old-Oldenburger, Quarter Horse, Sorraia, Standardbred, SueddeutschesKatblut, Thoroughbred and Welsh)	Useful for biodiversity research, GWAS, association mapping, genomic prediction of disease risk
Multiple Cattle, pig, chicken, horse, goat and sheep	Affymetrix Axiom	IMAGE001 Multispecies Array	Low	60 000 total 10 000 per species	Many breeds/per species (from publicly avaialvable arrays and project data)	Designed for assessment of genebank collections, but also useful for biodiversity research, causal mutation known traits, parentage study, ancestral information

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Multiple Buffalo, rabbit, duck, quail, pigeon and honey bee	Affymetrix Axiom	IMAGE002 Multispecies Array	Low	60 000 total 10 000 per species	Many breeds/per species (from publicly avaialvable arrays and project data)	Designed for assessment of genebank collections, but also useful for biodiversity research, causal mutation known traits, parentage study, ancestral information
Pig	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Porcine50K	Medium	51 000	NA	Useful for biodiversity research, GWAS, parentage verification, evaluation of pure line, identification of multi-line reference populations, genetic evaluation; Includes markers for detection of Porcine Stress syndrome and Rendement Napole (RN).
	Affymetrix Axiom	Porcine-Breeder Array		55 150	Same as Axiom-PorcineHD Genotyping Array, but SNP selection was targeted towards commercial pig breeds	Useful for genomic and trait selection, Parentage testing; Includes 64 ISAG core parentage markers, 42 trait-specific markers selected by USDA and targeted for commercial breeds including Large White, Landrace, Duroc and Pietrain
	Illumina Infinium	Porcine SNP60 v2 Genotyping Beadchip		64 232	Berkshire, Duroc, Hampshire, Landrace, Large White, Meishan, Pietrain, Synthetic (Large White and Pietrain), Wild boar	Useful for biodiversity research, GWAS, genomic selection, identification of QTL, mapping, genetic evaluation, linkage disequilibrium studies
	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Porcine80K		~80 000	NA	Useful for biodiversity research, GWAS, parentage verification, evaluation of pure line, identification of multi-line reference populations, genetic evaluation

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Pig	Affymetrix Axiom	PorcineHD Genotyping Array	High	658 692	27 domestic and 16 wild boar populations; Commercial domestic European (Large White, Landrace, Pietrain, Duroc, Hampshire); Traditional/Rare domestic European (Angler Sattelschwein, British Saddleback, Bunte Bentheimer, Casertana, Cinta Senese, Gloucester Old Spot, Large Black, Linderodssvin, Mangalica, Middle Wite, Negro Iberico, Retinto, Tamoworth); European wild boar; Asian domestic (Jiangquhai, Jinhua,Leping Spotted, Meihan, Thai Native, Wannan Spotted, Wuzishan, Xiang, Zang); Asian wild boar (China, Japan, Thailand)	Useful for biodiversity research, GWAS, high resolution genetic mapping, QTL analysis, genomic prediction
Sheep	Affymetrix Axiom	OvineLD-Genotyping Array	Low	11 196	Belclare, Charollais, Suffolk, Texel, Vendeen	Useful for biodiversity research, GWAS, parentage verification, breed assignment, aneuploidy detection, mating designs, traceability.
	Affymetrix Axiom	Ovine50K-Genotyping Array	Medium	50 000	74 breeds sampled from Asian, African, South-West Asian, Carribean, North American, South American, European and Australasian sheep	Useful for biodiversity research, GWAS, parentage verification, breed assignment, aneuploidy detection, mating designs, traceability.
	GeneSeek Genomic Profiler (Illumina-Infinium)	GGP Ovine50K		50 000	NA	Useful for biodiversity research, GWAS, genomic selection, identification of QTL, genetic evaluation, linkage disequilibrium studies

Species	Platform	Array name	Marker density	SNPs (N)	SNP ascertainment and/or validation breeds/populations	Special features/potential applications
Sheep	Affymetrix Axiom	Ovicap Mutispecies Array	Medium	54 236	74 breeds sampled from Asian, African, South-West Asian, Carribean, North American, South American, European and Australasian sheep	Useful for biodiversity research, GWAS, mapping studies
	Illumina Infinium	Ovine SNP50-Genotyping-Beadchip		54 241	>70 breeds sampled from Asian, African, South-West Asian, Carribean, North American, South American, European and Australasian sheep	Useful for biodiversity research, GWAS, genomic selection, genetic evaluation, linkage disequilibrium studies

Appendix 8

Bioinformatics pipeline for quality control of genomic data

Quality control (QC) of large datasets of single nucleotide polymorphisms is an important step to ensure reliable results. Fortunately, a large collection of software has already been developed to aid in this task. This appendix describes the step-by-step application of a "QC pipeline" that combines the use of command line and R commands to manipulate and perform quality checks on binary PLINK formatted data. This pipeline outlines a standard approach to quality check genotype data. However, parameters and QC steps and their order might change when applied to different data or depending on downstream analyses. Finally, although some of the visual checks and simplest calculations might be performed using other statistical tools (e.g., Excel spreadsheet) the use of R is highly recommended due to its scalability.

REQUIREMENTS

The dataset used here is a collection of 60k SNP chip genotype data of Italian goat breeds (Cortellari *et al.*, 2021). The dataset is available in binary PLINK format: `Cortellari2021.{bed, bim, fam}` at https://data.mendeley.com/datasets/hnd59x6gmg/1.

The software used are:

- PLINK v1.9 (https://www.cog-genomics.org/plink/)
- PLINK v2.0 (https://www.cog-genomics.org/plink/2.0/)
- R v4.1.0 (https://www.r-project.org/)

All PLINK commands are executed in the system command line terminal (identified here by a preceding "`$`" symbol). R commands require the R console (identified here by a preceding ">" symbol)

The R package data.table is required to efficiently load file content in R.

```
> install.packages("data.table")
> library(data.table)
```

QUALITY CONROL

Missingness

The first step is to determine the rates of missing data both at the individuals and loci level. The `--missing` flag is applied to compute both using one single command. The chromosomal setup (29 autosomes in goats) can be specified using the --cow flag (cattle also have 29 autosomes).

```
$ plink --cow --bfile Cortellari2021 --missing --out Cortellari2021
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
53347 variants loaded from .bim file.
1071 cattle (16 males, 51 females, 1004 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1068 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1591 het. haploid genotypes present (see Cortellari2021.hh ).
Total genotyping rate is 0.977168.
--missing: Sample missing data report written to Cortellari2021.imiss, and
variant-based missing data report written to Cortellari2021.lmiss.
```

FIGURE A6.1

Screenshot following the test for missingness.

The command produces two files with extensions, `.lmiss` and `.imiss`, containing loci and individual missingness information for loci and individual animals, respectively.

The missingness distribution is then assessed further by using R:

```
> imiss <- fread("Cortellari2021.imiss", select = 6)
> lmiss <- fread("Cortellari2021.lmiss", select = 5)
> par(mfrow = c(1, 2))
> hist(imiss$F_MISS, xlab = "missingness freq", main = "individual
missingness", breaks = 50)
> hist(lmiss$F_MISS, xlab = "missingness freq", main = "loci missingness",
breaks = 50)
> par(mfrow = c(1, 1))
```

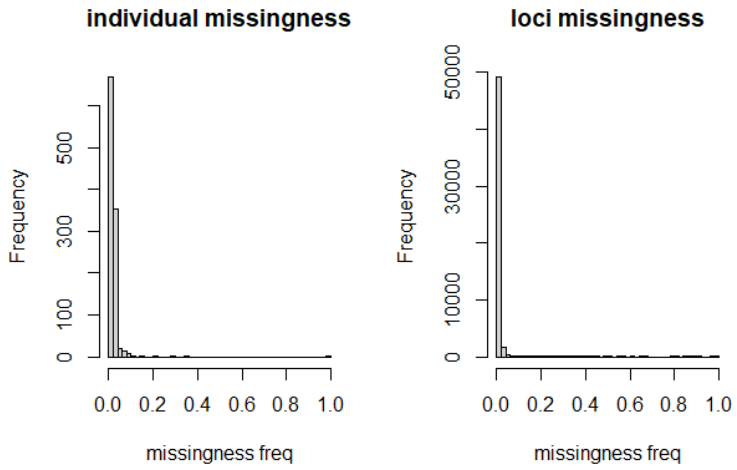

FIGURE A6.2

Distributions of missingness according to individuals and loci

To maximise the number of individuals, first prune for loci missingness, followed by individual missingness then. Pruning for missingness can be performed in PLINK using the `--geno` and `--mind` for loci and individuals, respectively.

```
$ plink --cow --bfile Cortellari2021 --geno 0.05 --make-bed --out
Cortellari2021_g05
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
53347 variants loaded from .bim file.
1071 cattle (16 males, 51 females, 1004 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021_g05.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1068 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1591 het. haploid genotypes present (see Cortellari2021_g05.hh ).
Total genotyping rate is 0.977168.
2228 variants removed due to missing genotype data (--geno).
51119 variants and 1071 cattle pass filters and QC.
Note: No phenotypes present.
--make-bed to Cortellari2021_g05.bed + Cortellari2021_g05.bim +
Cortellari2021_g05.fam ... done.
```

FIGURE A6.3

Screenshot following pruning for loci missingness

```
$ plink --cow --bfile Cortellari2021_g05 --mind 0.05 --make-bed --out
Cortellari2021_g05m05
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
51119 variants loaded from .bim file.
1071 cattle (16 males, 51 females, 1004 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021_g05m05.nosex .
23 cattle removed due to missing genotype data (--mind).
IDs written to Cortellari2021_g05m05.irem .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1045 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1450 het. haploid genotypes present (see Cortellari2021_g05m05.hh ).
Total genotyping rate in remaining samples is 0.99708.
51119 variants and 1048 cattle pass filters and QC.
Note: No phenotypes present.
--make-bed to Cortellari2021_g05m05.bed + Cortellari2021_g05m05.bim +
Cortellari2021_g05m05.fam ... done.
```

FIGURE A6.4

Screenshot following pruning for individual missingness

The missingness statistics are then checked:

```
$ plink --cow --bfile Cortellari2021_g05m05 --missing --out
Cortellari2021_g05m05
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
51119 variants loaded from .bim file.
1048 cattle (16 males, 50 females, 982 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021_g05m05.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1045 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1450 het. haploid genotypes present (see Cortellari2021_g05m05.hh ).
Total genotyping rate is 0.99708.
--missing: Sample missing data report written to Cortellari2021_g05m05.imiss,
and variant-based missing data report written to Cortellari2021_g05m05.lmiss.
```

FIGURE A6.5

Screenshot when checking missingness statistics

Missingness ratios are now negligible; a call rate of ~0.998 was obtained at the cost of only 2,228 loci (from 53,347 to 51,119) and 23 individuals (1,071 to 1,048), comparing Figures A6.4 and A6.5.

Results are then visualized in R:

```
> imiss.pruned <- fread("Cortellari2021_g05m05.imiss", select = 6)
> lmiss.pruned <- fread("Cortellari2021_g05m05.lmiss", select = 5)
> par(mfrow=c(1, 2))
> hist(imiss.pruned$F_MISS, xlab = "missingness freq", main = "individual
missingness (after pruning)")
> hist(lmiss.pruned$F_MISS, xlab = "missingness freq", main = "loci
missingness (after pruning)")
```

```
> par(mfrow = c(1, 1))
```

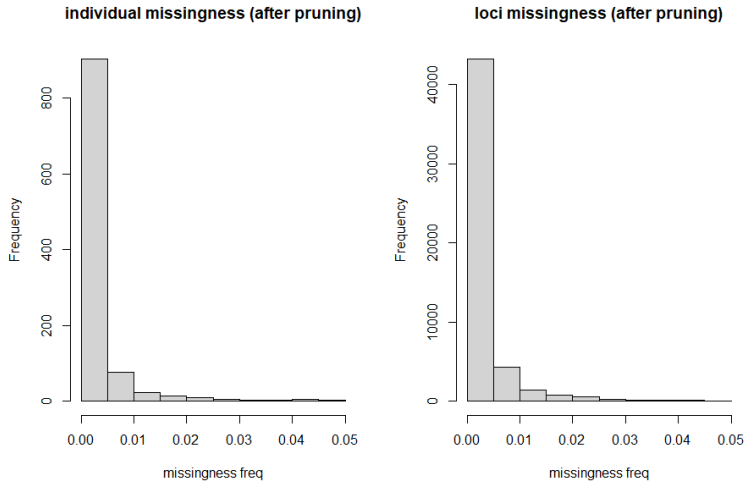

FIGURE A6.6

Graphical representations of missingness statistics for individuals and loci

Minor allele frequency

The allele frequency spectrum is then compared using the `--freq` flag in PLINK.

```
$ plink --cow --bfile Cortellari2021_g05m05 --freq --out
Cortellari2021_g05m05
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
51119 variants loaded from .bim file.
1048 cattle (16 males, 50 females, 982 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021_g05m05.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1045 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1450 het. haploid genotypes present (see Cortellari2021_g05m05.hh ).
Total genotyping rate is 0.99708.
--freq: Allele frequencies (founders only) written to Cortellari2021_g05m05.frq
```

FIGURE A6.7

Screenshot when assessing allelic frequencies

Results are then plotted with R (Figure A6.8).

```
> maf <- fread("Cortellari2021_g05m05.frq")
> hist(maf$MAF, xlab = "MAF", main = "Allele Frequency Spectrum")
```

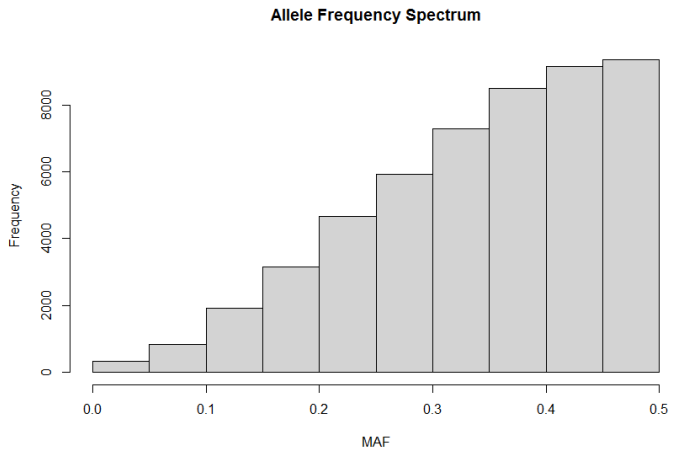

FIGURE A6.8

Distribution of minor allele frequencies (MAF)

Low frequency variants appear heavily under-represented due to ascertainment bias. We set the MAF threshold at 0.02 following the rule of thumb *10/N* where *N* = number of individuals.

```
> round(10/nrow(fread("Cortellari2021_g05m05.fam")), 2)
```

`0.01` (Output)

MAF pruning can be performed with the `--maf` flag in PLINK.

```
$ plink --cow --bfile Cortellari2021_g05m05 --maf 0.01 --make-bed --out
Cortellari2021_g05m05f01
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05f01.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
51119 variants loaded from .bim file.
1048 cattle (16 males, 50 females, 982 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021_g05m05f01.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1045 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1450 het. haploid genotypes present (see Cortellari2021_g05m05f01.hh
).
Total genotyping rate is 0.99708.
151 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
50968 variants and 1048 cattle pass filters and QC.
Note: No phenotypes present.
--make-bed to Cortellari2021_g05m05f01.bed + Cortellari2021_g05m05f01.bim +
Cortellari2021_g05m05f01.fam ... done.
```

FIGURE A6.9

Screenshot when pruning based on allelic frequencies

Linkage disequilibrium reduction

A dataset with reduced linkage disequilibrium (LD) can be used to feed all those analyses requiring approximate linkage equilibrium. The LD is scanned in sliding windows of 50 SNPs, sliding forward 5 SNPs at each step. The LD for each window is reduced to $r^2 = 0.2$.

```
$ plink –cow –bfile Cortellari2021_g05m05f01 –indep-pairwise 50 5 0.2 –out
Cortellari2021_g05m05f01
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05f01.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
50968 variants loaded from .bim file.
1048 cattle (16 males, 50 females, 982 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021_g05m05f01.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1045 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1450 het. haploid genotypes present (see Cortellari2021_g05m05f01.hh
).
Total genotyping rate is 0.997082.
50968 variants and 1048 cattle pass filters and QC.
Note: No phenotypes present.
--indep-pairwise: Ignoring 597 chromosome 0 variants.
Pruned 367 variants from chromosome 1, leaving 2816.
Pruned 266 variants from chromosome 2, leaving 2473.
Pruned 250 variants from chromosome 3, leaving 2051.
Pruned 281 variants from chromosome 4, leaving 2095.
Pruned 217 variants from chromosome 5, leaving 1983.
Pruned 280 variants from chromosome 6, leaving 2078.
Pruned 253 variants from chromosome 7, leaving 1885.
Pruned 243 variants from chromosome 8, leaving 2027.
Pruned 127 variants from chromosome 9, leaving 1711.
Pruned 212 variants from chromosome 10, leaving 1825.
Pruned 248 variants from chromosome 11, leaving 1849.
Pruned 235 variants from chromosome 12, leaving 1475.
Pruned 155 variants from chromosome 13, leaving 1456.
Pruned 230 variants from chromosome 14, leaving 1652.
Pruned 156 variants from chromosome 15, leaving 1447.
Pruned 158 variants from chromosome 16, leaving 1403.
Pruned 114 variants from chromosome 17, leaving 1286.
Pruned 156 variants from chromosome 18, leaving 1119.
Pruned 93 variants from chromosome 19, leaving 1096.
Pruned 118 variants from chromosome 20, leaving 1337.
Pruned 145 variants from chromosome 21, leaving 1259.
Pruned 126 variants from chromosome 22, leaving 1017.
Pruned 94 variants from chromosome 23, leaving 898.
Pruned 135 variants from chromosome 24, leaving 1158.
Pruned 71 variants from chromosome 25, leaving 784.
Pruned 88 variants from chromosome 26, leaving 931.
Pruned 89 variants from chromosome 27, leaving 812.
Pruned 68 variants from chromosome 28, leaving 841.
Pruned 76 variants from chromosome 29, leaving 881.
Pruned 376 variants from chromosome 30, leaving 1299.
Pruning complete.  5427 of 50371 variants removed.
Marker lists written to Cortellari2021_g05m05f01.prune.in and
Cortellari2021_g05m05f01.prune.out .
```

FIGURE A6.10

Screenshot when pruning based on linkage disequilibrium among loci

The SNPs to retain are listed in the output file with extension `.prune.in`. This latter file can be used in combination with the PLINK flag `-extract` to subsets the input file accordingly.

Extreme heterozygosity

Genotypic counts can be obtained by using the PLINK `--het` flag. This analysis is influenced by LD, hence, the `--extract` flag is applied in conjunction with the `.prune.in` file to limit the SNP list to loci in approximate linkage equilibrium.

```
$ plink --cow --bfile Cortellari2021_g05m05f01 --extract
Cortellari2021_g05m05f01.prune.in --het --out Cortellari2021_g05m05f01ld02
```

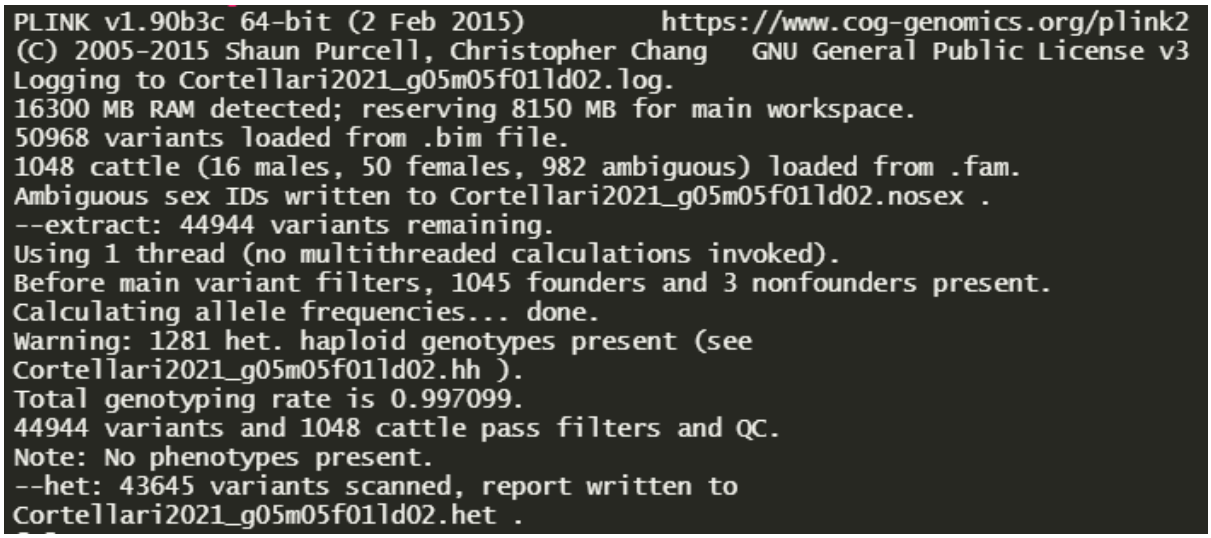

FIGURE A6.11

Screenshot when obtaining genotypic counts

The output with extension `.het` contains the observed Homozygotes counts for each individual. These data are then used to compute in R the observed heterozygosity.

```
> het <- fread("Cortellari2021_g05m05f01ld02.het")
> het$O_HET = (het$`N(NM)` - het$`O(HOM)`)/het$`N(NM)`
```

The heterozygosity distribution per population can then be plotted (Figure A6.12).

```
> boxplot(O_HET~FID, data = het, las = 2)
```

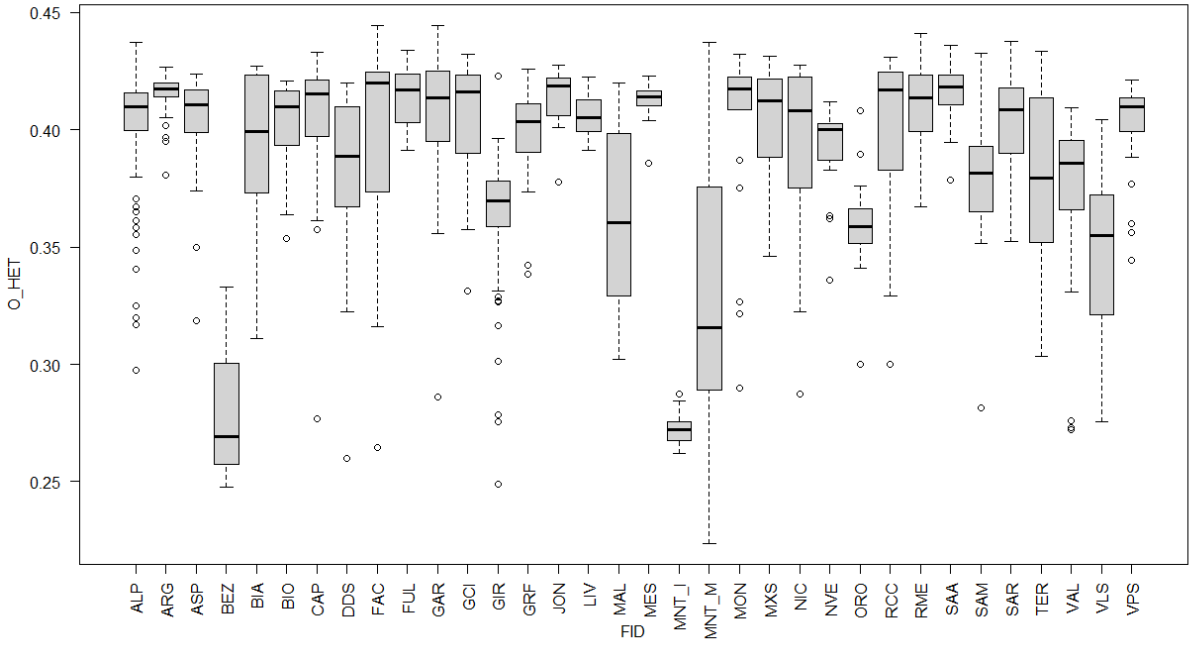

FIGURE A6.12

Distribution of heterozygosity according to population (breed)

R can then be used to identify those individuals having within-population extreme (>3 SD) heterozygosity.

```
> ExtHet <- NULL
> for(p in unique(het$FID)){
> subS = het[het$FID == p, ]
> ExtHet = rbind(ExtHet, subS[as.vector(abs(scale(subS$O_HET)) > 3),
c("FID", "IID")])
> }
> table(ExtHet$FID)
```

Output:

```
ALP ARG ASP CAP DDS FAC GAR GCI GIR MES NIC SAA
  4   1   1   1   1   1   1   1   1   1   1   1
```

The output above shows that the process identified 15 Individuals (i.e. 4 from the ALP breed and 1 from each of the others) showing within population extreme Heterozygosity. The list is then saved in a tab delimited file.

```
> fwrite(ExtHet, "IIDtoremove.txt", sep = "\t")
```

The data can then be filtered to remove individuals with extreme observed heterozygosity (*Ho*) while keeping SNP in approximate linkage equilibrium by combining the PLINK flags `--remove` (to exclude individuals) and `--extract` (to retains SNPs).

```
$ plink --cow --bfile Cortellari2021_g05m05f01 --extract
Cortellari2021_g05m05f01.prune.in --remove IIDtoremove.txt --make-bed --out
Cortellari2021_g05m05f01ld02h3
```

```
PLINK v1.90b3c 64-bit (2 Feb 2015)         https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05f01ld02h3.log.
16300 MB RAM detected; reserving 8150 MB for main workspace.
50968 variants loaded from .bim file.
1048 cattle (16 males, 50 females, 982 ambiguous) loaded from .fam.
Ambiguous sex IDs written to Cortellari2021_g05m05f01ld02h3.nosex .
--extract: 44944 variants remaining.
--remove: 1033 cattle remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1030 founders and 3 nonfounders present.
Calculating allele frequencies... done.
Warning: 1215 het. haploid genotypes present (see
Cortellari2021_g05m05f01ld02h3.hh ).
Total genotyping rate in remaining samples is 0.997085.
44944 variants and 1033 cattle pass filters and QC.
Note: No phenotypes present.
--make-bed to Cortellari2021_g05m05f01ld02h3.bed +
Cortellari2021_g05m05f01ld02h3.bim + Cortellari2021_g05m05f01ld02h3.fam ...
done.
```

FIGURE A6.13

Screenshot when filtering animals with extreme heterozygosity

Relatedness

Plink v2.0 implements the KING robust kinship estimation (Manichaikkul *et al.*, 2010). We can generate a table of kinship to check pairwise kinship across the dataset using the `--make-king-table` flag:

```
$ plink2 --cow --bfile Cortellari2021_g05m05f01ld02h3 --make-king-table --
out Cortellari2021_g05m05f01ld02h3
```

```
PLINK v2.00a2 64-bit (9 Oct 2019)             www.cog-genomics.org/plink/2.0/
(C) 2005-2019 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05f01ld02h3.log.
Options in effect:
  --bfile Cortellari2021_g05m05f01ld02h3
  --cow
  --make-king-table
  --out Cortellari2021_g05m05f01ld02h3

Start time: Mon Jul 19 10:08:25 2021
16300 MiB RAM detected; reserving 8150 MiB for main workspace.
Using up to 8 compute threads.
1033 samples (49 females, 15 males, 969 ambiguous; 1030 founders) loaded from
Cortellari2021_g05m05f01ld02h3.fam.
44944 variants loaded from Cortellari2021_g05m05f01ld02h3.bim.
Note: No phenotype data present.
Excluding 1299 variants on non-autosomes from KING-robust calculation.
--make-king-table: Scanning for singletons and monomorphic variants... done.
No singletons/monomorphics found.
--make-king-table pass 1/1: Writing...
--make-king-table: 43645 variants processed.
Results written to Cortellari2021_g05m05f01ld02h3.kin0 .
End time: Mon Jul 19 10:08:26 2021
```

FIGURE A6.14

Screenshot when generating a kinship table

The output is then loaded into R:

```
> kin <- fread("Cortellari2021_g05m05f01ld02h3.kin0")
```

A cutoff threshold = 0.354 is used to identify duplicate samples and monozygotic twins, with 0.354 being the geometric mean of kinship = 0.5 (for duplicates and monozygotic twins) and kinship= 0.25 (for parent-child, full siblings). The inferred kinship values can then be plotted (Figure A6.15) and thresholds set to identify relatedness proportions in the data:

```
> hist(kin$KINSHIP, xlim = c(0, 0.5))
> rel <- c(sqrt(0.5*0.25), sqrt(0.25*0.125), sqrt(0.125*0.0625))
> abline(v = rel, col = 2:4)
> text(rel+.02, 100000, round(rel, 3))
> legend("topright", legend = c("Dup/MZ", "PO/FS", "2nd"), col = 2:4,
lty=1)
```

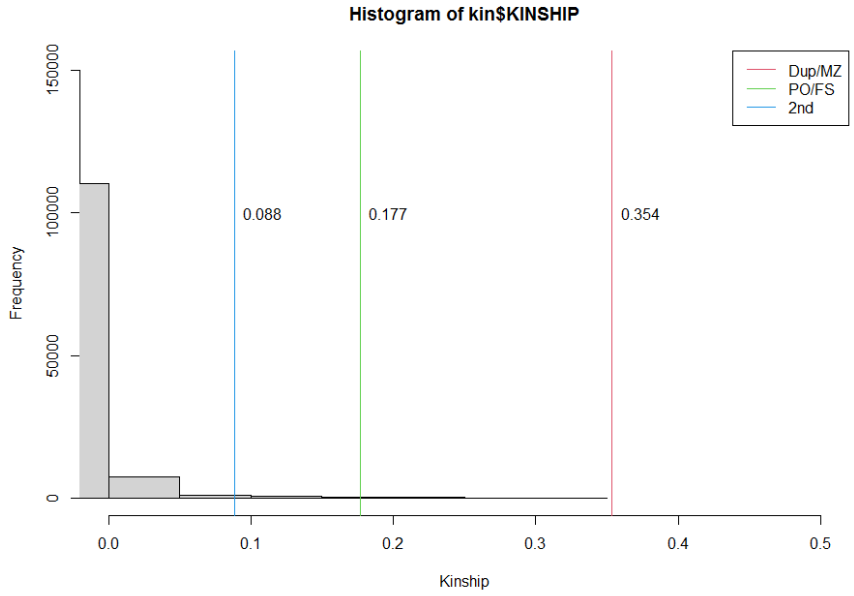

FIGURE A6.15

Distribution of kinship values

The `--king-cutoff` flag is then used in PLINK v2.0 to exclude one member of each pair of samples with relatedness ≥ PO (0.177 in the following example).

```
$ plink2 --cow --bfile Cortellari2021_g05m05f01ld02h3 --king-cutoff 0.177 -
-make-bed --out Cortellari2021_g05m05f01ld02h3rel
```

```
PLINK v2.00a2 64-bit (9 Oct 2019)              www.cog-genomics.org/plink/2.0/
(C) 2005-2019 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Cortellari2021_g05m05f01ld02h3rel.log.
Options in effect:
  --bfile Cortellari2021_g05m05f01ld02h3
  --cow
  --king-cutoff 0.177
  --make-bed
  --out Cortellari2021_g05m05f01ld02h3rel

Start time: Mon Jul 19 10:10:15 2021
16300 MiB RAM detected; reserving 8150 MiB for main workspace.
Using up to 8 compute threads.
1033 samples (49 females, 15 males, 969 ambiguous; 1030 founders) loaded from
Cortellari2021_g05m05f01ld02h3.fam.
44944 variants loaded from Cortellari2021_g05m05f01ld02h3.bim.
Note: No phenotype data present.
Excluding 1299 variants on non-autosomes from KING-robust calculation.
--king-cutoff: Scanning for singletons and monomorphic variants... done.
No singletons/monomorphics found.
--king-cutoff pass 1/1: Condensing...
--king-cutoff: 43645 variants processed.
--king-cutoff: Excluded sample IDs written to
Cortellari2021_g05m05f01ld02h3rel.king.cutoff.out.id , and 897 remaining sample
IDs written to Cortellari2021_g05m05f01ld02h3rel.king.cutoff.in.id .
Writing Cortellari2021_g05m05f01ld02h3rel.fam ... done.
Writing Cortellari2021_g05m05f01ld02h3rel.bim ... done.
Writing Cortellari2021_g05m05f01ld02h3rel.bed ... done.
End time: Mon Jul 19 10:10:16 2021
```

FIGURE A6.16

Screenshot when removing closely related (and possible duplicate) animals

For these data, 136 individuals were removed due to excessive relatedness leaving 897 individuals for further analysis (Figure A6.16).

Importantly, although it is often preferred to remove excess heterozygosity or relatedness from a working dataset, such features can be diagnostic of underlying population dynamics of interest. Hence, it is always advisable to perform these pruning steps keeping in mind the demographic history of the populations.

REFERENCES

Cortellari, M., Barbato, M., Talenti, A., Bionda, A., Carta, A., Ciampolini, R., Ciani, E. *et al.* 2021. The climatic and genetic heritage of Italian goat breeds with genomic SNP data. *Scientific Reports*, 11, 10986. https://doi.org/10.1038/s41598-021-89900-2

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. & Chen, W-M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26: 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Appendix 9

Typical steps to be taken in a genomic characterization study

ALL DATASETS

- Genotyping
- Formatting of dataset (PLINK)[13]
- Quality control (PLINK, KING, and R - See Appendix 8)
- Optional: merging with literature data (PLINK)
- Calculation and display of interindividual distances for detection of duplicates and outliers (mislabelling, crossbreeding) and for checking breed-level differentiation (PLINK)
- Calculation of summary statistics per breed: nucleotide diversity (genomes), expected and observed heterozygosity, heterozygote deficiency (PLINK)

MOST DATASETS

- Inferring of relationships of breeds and genetic clines
  - PCA (Eigensoft, PLINK and specialized R functions)
  - model-based clustering (Admixture, Frappe, Structure)
  - NeighborNet graphs (SplitsTree)

DEPENDING ON THE DATASET AND OBJECTIVES OF THE STUDY

- Phasing and detection of haplotype-based clusters by fine-structure (AphaPhase, EAGLE, fastPhase)
- ROH content (KING, PLINK)
- Inferring of gene flows by the f3, f4 and of the D statistics (ADMIXTUREGRAPH, MIXMAPPER, TreeMix,)
- Examine the genetic basis of traits
  - GWAS - when individual phenotypes are available (BayesR, CGTA, EMMAS, PLINK,)
  - Selection signatures for population-wide characteristics
    - Within breeds (cgaTOH, rehh, Selscan, SweeD, Sweep)
    - Across breeds (HierfFstat, PLINK, rehh)
- Coalescence analysis (MSMC and PSMC R packages)
- Reconstructing the history of a population (ABCtoolbox, Fastsimcoal2, SliM)

[13] Common software options are indicated in parenthesis