GENOMIC CHARACTERIZATION OF ANIMAL GENETIC RESOURCES TECHNICAL GUIDELINES

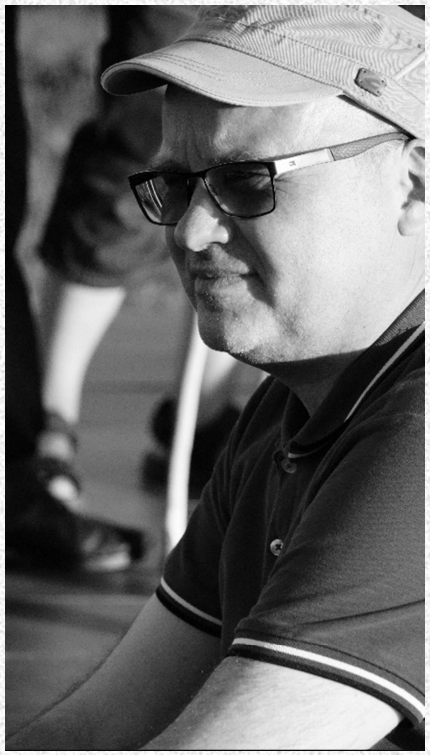
Catarina Ginja
Juha Kantanen
section editors

Asking simple questions to explain complex demographic scenarios

SECTION 4 – Applications of genomics

SECTION EDITORS

CATARINA GINJA
UNIVERSITY OF PORTO, BIOPOLIS



JUHA KANTANEN
NATURAL RESEOURCES INSTITUTE - LUKE

SECTION 4 – Applications of genomics

Major goals

- Characterize patterns, e.g. genetic diversity, differentiation, classification, genetic clines, etc.
- Infer demographic processes, e.g. expansions, contractions, admixture, gene flow, etc.

SECTION 4 – Applications of genomics

The resulting patterns are better interpreted if the underlying models and processes are understood using integrative analyses.

Authors

https://commons.wikimedia.org/wiki/File:Standard_map_of_Europe_(blank).png

Authors – Text boxes

SECTION 4 – Applications of genomics

ASSESSMENT OF GENOMIC VARIATION WITHIN-POPULATIONS

Measures of genetic diversity
Inbreeding and runs of homozygosity
Effective population size

SECTION 4 – Applications of genomics

Measures of genetic diversity

observed (Ho) and expected (He) heterozygosities – Ho is based on the actual distribution of genotypes, whereas He is based on allele frequencies

nucleotide diversity (π) – gives the average number of nucleotide differences per site between two DNA sequences chosen randomly from the studied population (~Ho)

effective population size (Ne) - represents the number of reproducing individuals given a number of conditions

SECTION 4 – Applications of genomics

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

Inbreeding increases the homozygosity in individuals and thus decreases Ho, but it does not change He based on the allele frequencies, resulting in a heterozygote deficit and a departure from the Hardy-Weinberg equilibrium (HWE) of homozygous and heterozygous genotypes

SECTION 4 – Applications of genomics

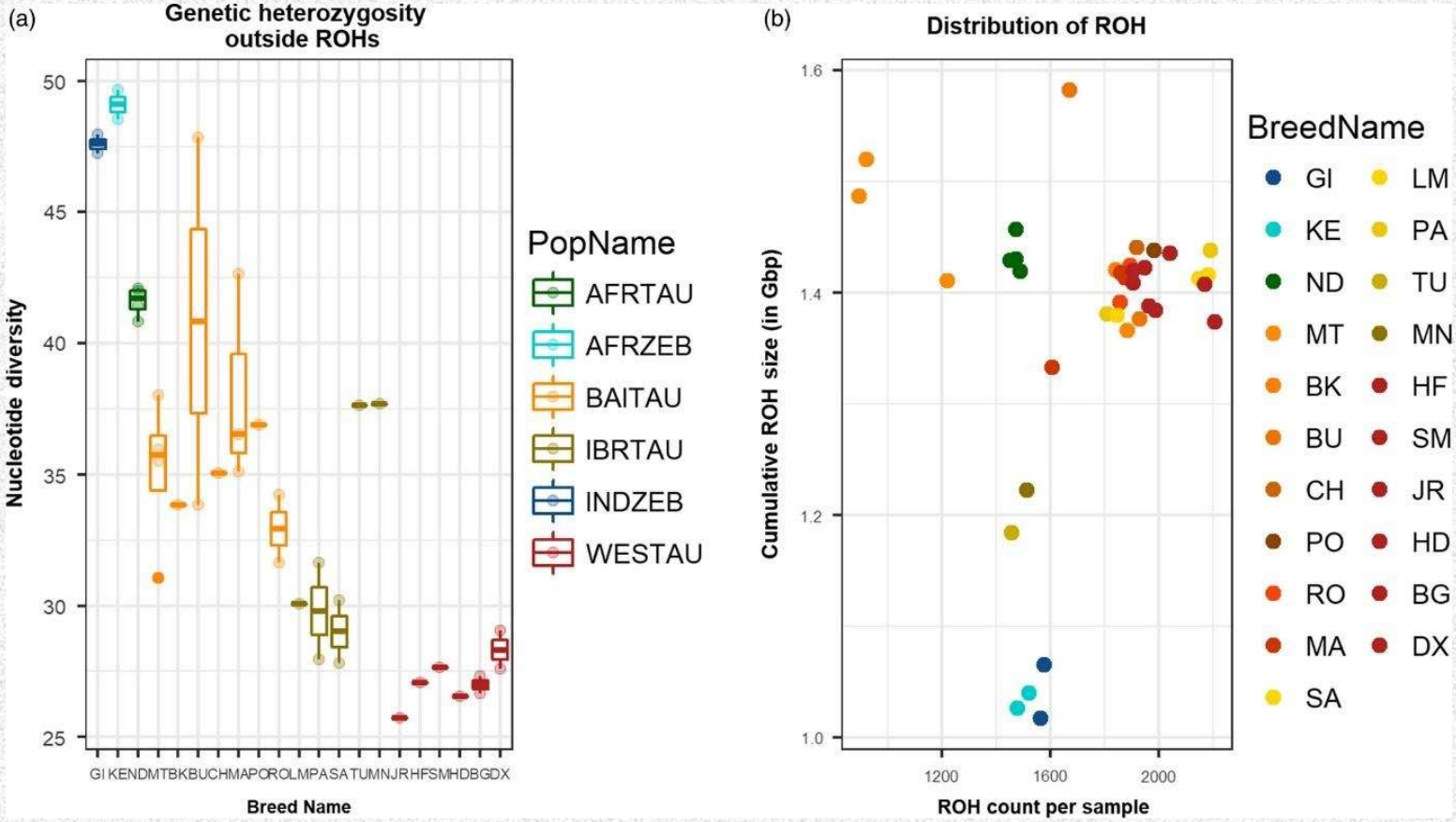
Inbreeding and runs of homozygosity

Wright's F-statistics based on comparing Ho and He can be used to infer inbreeding (FIS) from molecular data

The genomic coverage by runs of homozygosities (ROHs) has become a commonly used measure for inbreeding

SECTION 4 – Applications of genomics

Genomic variation within-populations



Boxplots showing average heterozygosities calculated in 10-kbp windows outside ROH for groups of cattle breeds (a) and distribution of ROH in each breed (b) using whole-genome sequencing data.

SECTION 4 – Applications of genomics

Inbreeding and runs of homozygosity

Inferring the frequency and length of ROH helps to understand population histories including the occurrence of bottlenecks, estimate inbreeding and identify signatures of selection

long ROH segments suggest recent inbreeding (consanguinity) and low genetic diversity,
e.g. Holstein-Friesian cattle

many short ROHs with only a few long ROHs indicate a reduced population size in the past and little recent inbreeding

SECTION 4 – Applications of genomics

Effective population size

The effective size of a population (Ne) corresponds to the number of individuals in an idealized Wright-Fisher population that would become inbred or lose diversity at the same rate as this population

Ne 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

Various coalescent-based methods have been developed, that allow to infer the population demography over time including the variation of Ne or connectivity

SECTION 4 – Applications of genomics

ASSESSMENT OF POPULATION STRUCTURE AND BETWEEN-BREED GENOMIC VARIATION

Principal component analysis
Model-based clustering
Genetic distances
Genetic distances between individuals
Phylogenetic trees and networks
NeighbourNet graphs
Detection of admixture using f3, f4 and D statistics
Phylogeny across the genome

SECTION 4 – Applications of genomics

Principal component analysis

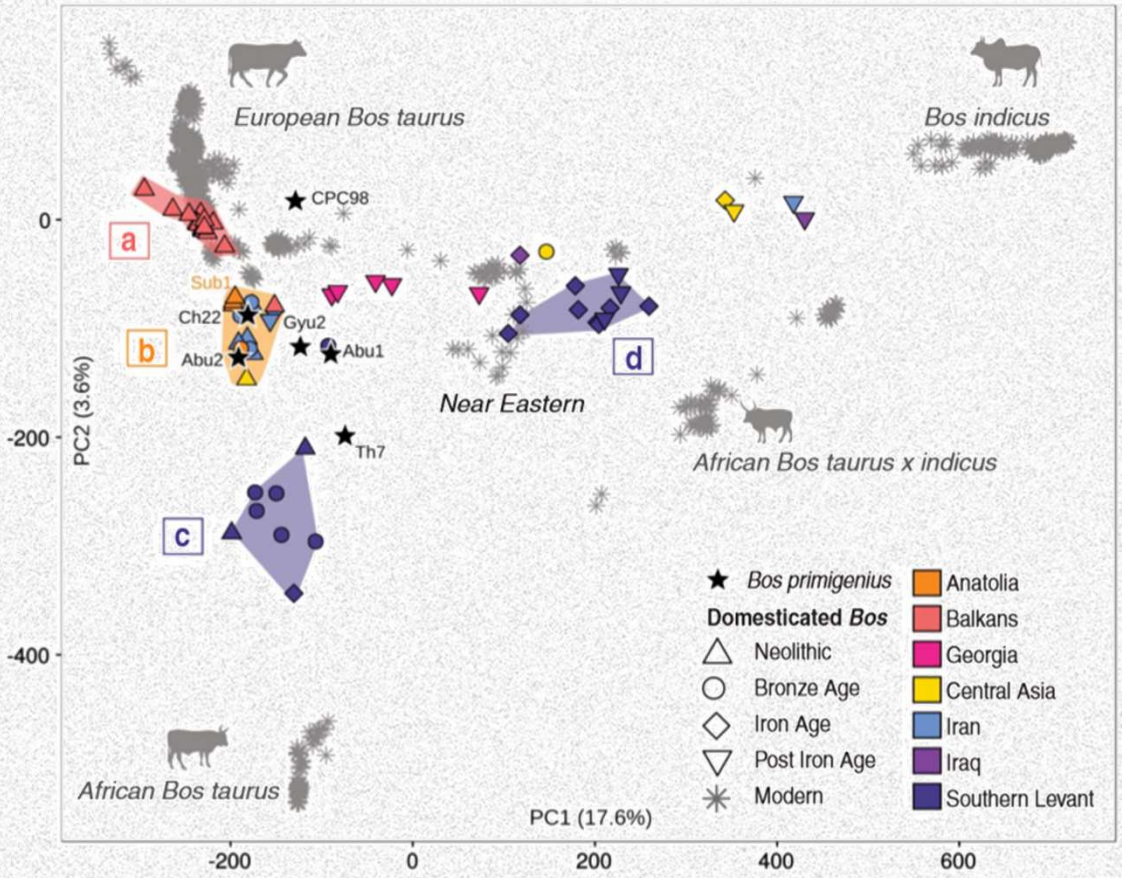
PCA is a statistical method to capture the variability associated with many variables (such as marker genotypes) into a much smaller number of variables that still contain most of the original variation

the principal components (PCs) - are ranked in descending order according to the amount of variation they explain

Distinguishing the groups is possible if (and only if) the product of the number of samples and markers is larger than $(1 / FST)^2$

SECTION 4 – Applications of genomics

Population structure and between-breed variation



PCA showing ancient DNA data projected over that of extant cattle using whole-genome sequencing data (Verdugo et al. 2019).

SECTION 4 – Applications of genomics

Principal component analysis – Limitations

PCA plots visualize only a small proportion of the total variation, they do not unambiguously indicate a close relationship of samples and do not clearly indicate duplicates

Unbalanced sample sizes may strongly affect the position of the groups on the plot, e.g. those with large sample sizes are typically pushed towards the centre of the plot

If the dataset contains a breed that, by genetic isolation, differs substantially from all others, the major PCs mainly reflect this difference, which corresponds to the bias due to inbreeding

SECTION 4 – Applications of genomics

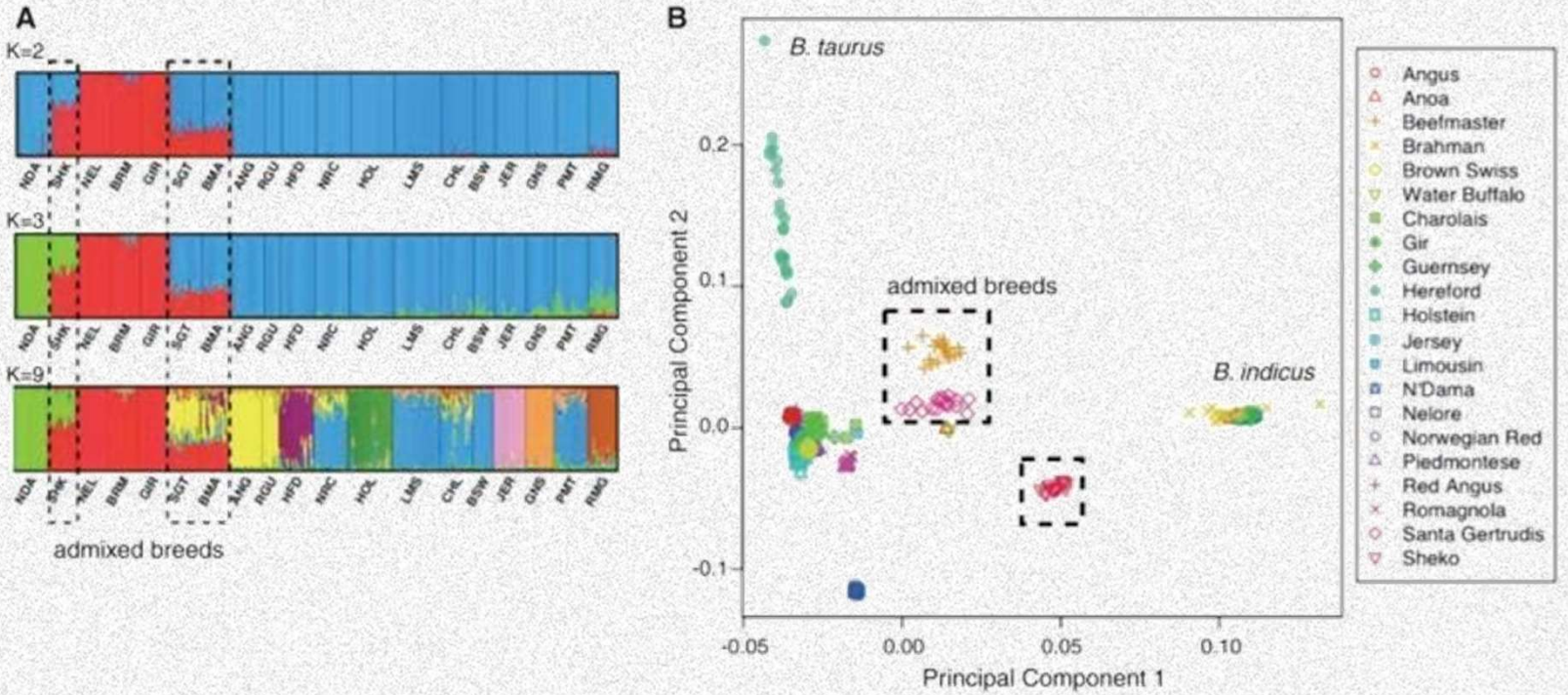
Model-based clustering

These clustering methods assume there are K genetic groups (or clusters), and all polymorphic sites are at HWE and linkage equilibrium within these groups.

An individual is modelled as a mixture of these K groups with specific proportions. The K clusters and their associated allele frequencies are inferred, and the biological meaning of each cluster is deduced a posteriori.

SECTION 4 – Applications of genomics

Population structure and between-breed variation



The Bovine HapMap Consortium (Gibbs et al., 2009), genotype data at 37,470 SNPs for 497 cattle from 19 geographically and biologically diverse breeds. This dataset was analysed using a Structure like (panel A) and a PCA approach (panel B).

SECTION 4 – Applications of genomics

Model-based clustering

Ideally, this reveals the major ancestral populations and the composition of admixed individuals or populations. It does not model evolutionary processes such as drift, mutation, migration or divergence.

The 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.

SECTION 4 – Applications of genomics

Genetic distances

Common approaches to visualize genetic structure use a matrix of genetic distances between all pairs of individuals. For whole-genome SNPs of a diploid species, distances are based either on the identity by state (IBS) of markers or on genomic relationships of individuals.

These distances can be visualized via MDS analysis or via neighbour-joining trees. 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. The accuracy depends on an adequate sample size per breed.

SECTION 4 – Applications of genomics

RECONSTRUCTION OF POPULATION HISTORY AND DEMOGRAPHIC MODELLING

MITOCHONDRIAL DNA AND THE SEX-CHROMOSOMES

SECTION 4 – Applications of genomics

Reconstruction of population histories and demographic modelling

Methods that aim at detecting, dating and quantifying population size changes and other events such as admixture or population splits. Genomic data enable testing alternative interpretations and models by comparing the genetic consequences of alternative evolutionary scenarios with observed data.

A coalescent approach works backward in time. 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. 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.

SECTION 4 – Applications of genomics

Reconstruction of population histories and demographic modelling

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).

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.

SECTION 4 – Applications of genomics

Reconstruction of population histories and demographic modelling

PSMC/MSMC - estimate the history of population size changes using information on mutation and recombination rates to analyse and interpret the distribution of heterozygous sites along the genome(s).

Other methods use the site or allele frequency spectrum (AFS) to summarize genomic data across many independent loci.

ABC - is particularly well-adapted to study complex demographic datasets and evolutionary models for which the likelihood may be difficult or impossible to compute. It can be used to compare alternative demographic models by computing the relative proportions of simulated data closer to the observed data.

SECTION 4 – Applications of genomics

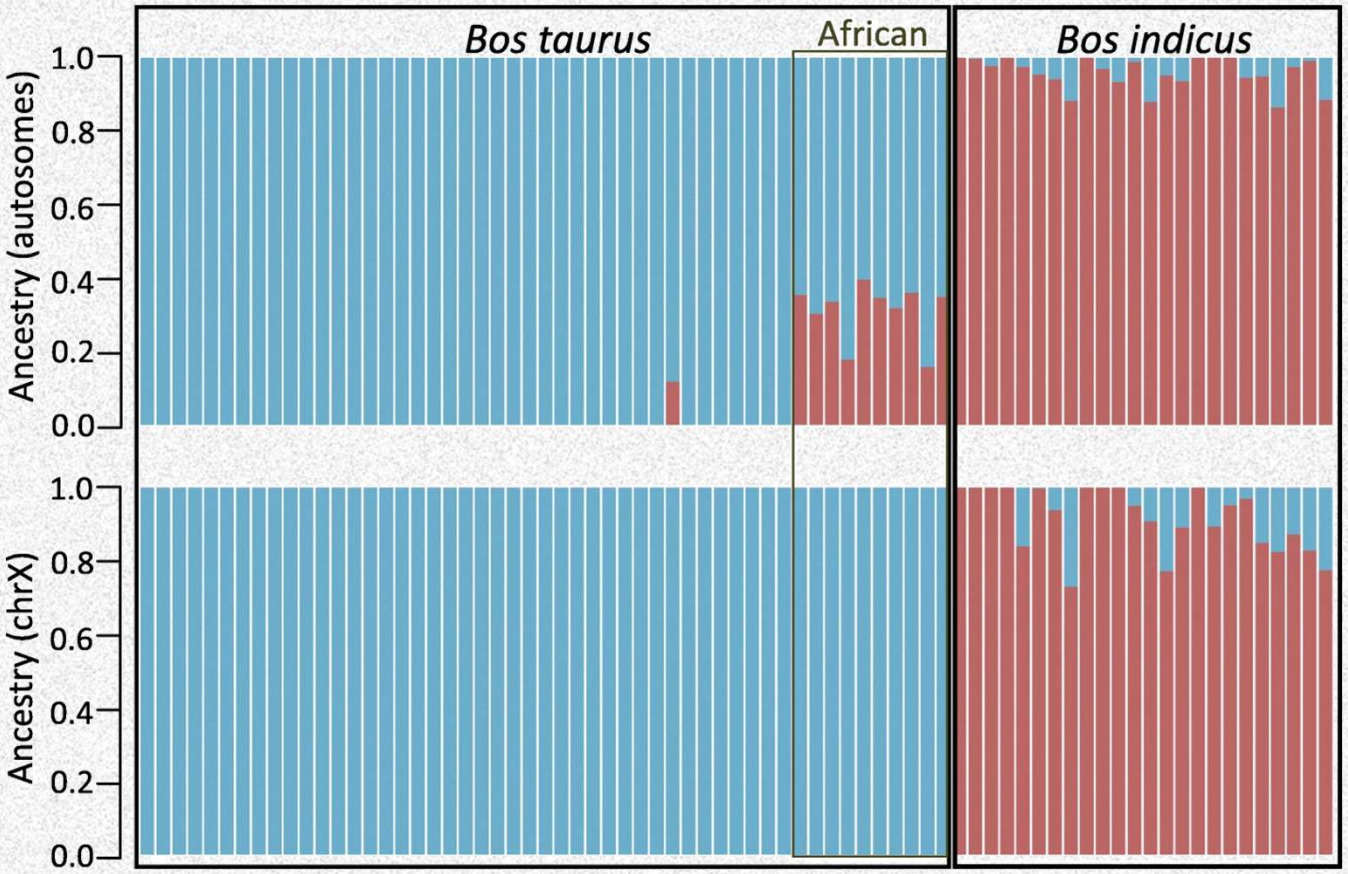
Mitochondrial DNA and the sex-chromosomes

Because of their lack of recombination, the evolution of mtDNA and Y-chromosomal DNA is less complex than the evolution of autosomes, and can be reconstructed largely by obtaining a phylogenetic tree and coalescence analysis.

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, but is considerably less informative than autosomal DNA for reconstructing breed histories.

SECTION 4 – Applications of genomics

Information on autosomal vs the sex-chromosomes



Population structure at K = 2 determined using the female individuals only. The indicine contribution to African taurine (N'Dama) is not observed in sex chromosome X (bottom) compared to the autosomes (top). (Da Fonseca et al. 2019).

Recommendations

Studies of the present patterns of genetic diversity of livestock are complemented by the analysis of ancient DNA to provide a historic context.

Understand the statistical methods and critically evaluate the results. Interpretation of statistical “significance” must be considered in the context of a meaningful biological message.

Recommendations

Interpret results in terms of population genetic, phylogeographic or evolutionary events, such as common ancestry, divergence, population bottlenecks, genetic drift, admixture, migrations, formation of clines, selection, adaptation, and/or in terms of specific changes in the genome and within genes in relation to biochemical and physiological mechanisms.

CATARINAGINJA@CIBIO.UP.PT

BIOPOLIS

CIBIO
Centro de Investigação em Biodiversidade e Recursos Genéticos

InBIO
Rede de Investigação em Biodiversidade e Biologia Evolutiva

U.PORTO

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

contract grant
2020.02754.CEECIND