EPFL

Thèse n° 7147

Biogeoinformatics for the management of Farm Animal Genetic Resources (FAnGR)

Présentée le 16 juillet 2020

à la Faculté de l'environnement naturel, architectural et construit
Laboratoire de systèmes d'information géographique
Programme doctoral en génie civil et environnement

pour l'obtention du grade de Docteur ès Sciences

par

Solange Catherine GAILLARD

Acceptée sur proposition du jury

Dr J. Skaloud, président du jury
Dr S. Joost, Prof. F. Golay, directeurs de thèse
Dr C. Flury, rapporteuse
Prof. P. Ajmone-Marsan, rapporteur
Prof. R. Schlaepfer, rapporteur

École polytechnique fédérale de Lausanne

2020

Acknowledgements

Je voudrais remercier ici toutes les personnes sans lesquelles cette thèse n'aurait pas vu le jour.

D'un point de vue scientifique, je commencerai bien sûr par mes directeurs de thèse, Prof. François Golay et plus particulièrement Dr. Stéphane Joost qui m'ont supervisé pendant cette aventure. Je voudrais aussi remercier tout le laboratoire du LaSIG pour les échanges aussi bien sociaux que scientifiques; en particulier, j'aimerais mentionner les personnes avec lesquelles j'ai eu le plus d'interaction au sujet de ma thèse: Estelle Rochat, Sylvie Stucki, Elia Vajana, Oliver Selmoni, Kevin Leempoel, Matthew Parkan, Jessie Madrazo, Anaïs Ladoy et Annie Guillaume. Certaines personnes en dehors de l'EPFL m'ont également apporté un précieux soutien. La première personne à laquelle je pense est sans aucun doute Christine Flury, qui, sans avoir été officiellement co-directrice de ma thèse m'a régulièrement suivi et m'a prodigué de nombreux conseils sur les aspects liés aux animaux d'élevage. Ensuite, je pense également à toutes les personnes que j'ai pu rencontrer lors de conférences ou de projets, tels que les projets NEXTGEN, ClimGen et bien d'autres encore.

Mais le support scientifique ne serait rien sans ma famille. Mes parents d'abord, Gilbert et Claire-Lise Duruz, qui m'ont donné la vie et m'ont soutenu pendant mes études. Mes frères et sœurs, Christelle, Cyril, Myriam et Priscille, qui m'ont toujours soutenu dans les moments difficiles. Finalement mon mari et mes enfants, Jérémie, Judith and Simon Gaillard qui réchauffent mon foyer quand je rentre fatiguée le soir à la maison.

Summary

In the context of both severe selection in farm animals and potential effects of climate change, it is crucial to implement a sustainable management of the breeding practice, supported by a judicious use of geographic information technologies. Based on this observation, this thesis advocates the use of *biogeoinformatics* (the combined use of biology, geographic information and informatics) to cope with the challenges encountered by the livestock sector. Indeed, although *biogeoinformatics* can provide key insights for FAnGR (Farm Animal Genetic Resources) management, the variety and complexity of tasks involved hinders a wider usage of this type of analyses. The thesis shows how novel dedicated tools are likely to facilitate the adoption of *biogeoinformatics* by animals scientists and by stakeholders in the livestock sector, while investigating three main challenges related to FAnGR management, namely i) erosion of genetic diversity, ii) effects of climate change on the breeding activity, and iii) pressure on typical cultural breeding practices such as high alpine grazing. On this basis, the thesis is organised around three axes:

1) Preserving locally adapted breeds

In order to prevent the erosion of genetic diversity, locally adapted breeds should be monitored to prevent their extinction. To this end, we developed the open source GENMON WebGIS platform, able to monitor FAnGR and to evaluate the degree of endangerment of livestock breeds. The system integrates various sources of information that are linked with the help of geographic information: pedigree and introgression, geographical concentration of animals, cryo-conservation and sustainability of breeding activities. The score can be visualised on a map and allows a fast and regional identification of breeds in danger.

2) Preserving locally adapted genetic variations

Considering the pace at which genetic diversity is being eroded, it has become urgent to identify and then preserve important genetic variations linked to locally adapted phenotypes. In this context, the recent *Samβada* software was designed to search for signatures of local adaptation through the study of genome–environment association. However, pre- and postprocessing of data for this analysis can be labour-intensive, and, we therefore developed the *R.SamBada R* package providing a pipeline for landscape genomic analyses. Based on *Samβada*, it spans from the retrieval of environmental conditions at sampling locations to gene annotation using the Ensembl genome browser. As a result, it grants access to *biogeoinformatic* analyses to researchers with no skills in geography.

3) Preserving a traditional farming technique suited for local breeds

The preservation of locally adapted breeds is also strongly linked with the conservation of traditional farming techniques, which in Switzerland, include the grazing of high alpine pastures during summer. One major effect of this transhumance on cows is that milk production declines due to food shortage and climatic stress. Here, we developed a new mathematical model to fit a lactation curve for mountain-pastured cows, and tested the influence of environmental, physiological, and morphological factors on the production using five million monthly milk records from Braunvieh cows. When compared to physiological factors, environmental variables show a limited impact on milk production at alpine pastures, precipitation in spring being the most important.

Keywords

Biogeoinformatics, Farm Animal Genetic Resources (FAnGR), livestock, local adaptation, local breeds, erosion of genetic diversity, climate change, transhumance, GIS, landscape genomics, integrated tools.

Résumé

Dans le contexte de sélection drastique des animaux d'élevage et de changements climatiques, il est devenu crucial de mettre en œuvre une gestion durable de la pratique de l'élevage, soutenue par une utilisation judicieuse des technologies de l'information géographique. Sur la base de cette observation, cette thèse préconise l'utilisation de la *biogéoinformatique* (l'utilisation combinée de la biologie, de l'information géographique et de l'informatique) pour faire face aux défis rencontrés dans le secteur de l'élevage. En effet, si la *biogéoinformatique* peut fournir des informations capitales pour la gestion des ressources génétiques des animaux d'élevage, la variété et la complexité des tâches à accomplir empêchent une utilisation plus large de ces analyses. Cette thèse montre donc comment de nouveaux outils spécialisés sont susceptibles de faciliter l'adoption de cette approche par les biologistes et les acteurs du secteur de l'élevage, tout en étudiant trois grands défis des animaux d'élevage : i) l'érosion de la diversité génétique, ii) les effets du changement climatique sur l'élevage, et iii) la pression sur les pratiques d'élevage traditionnelles. La thèse s'agence selon trois axes :

1) Préserver les races adaptées localement

Afin de prévenir l'érosion de la diversité génétique, les races adaptées localement devraient être surveillées pour éviter leur extinction. Nous avons donc développé une plateforme open source Web SIG appelée GENMON, capable de d'évaluer le degré de menace des races d'animaux d'élevage. Le système intègre plusieurs informations, liées entre elles grâce à l'information géographique : pedigree et introgression, concentration géographique, cryoconservation et durabilité des activités d'élevage. Le score peut être visualisé sur une carte et permet une identification rapide et régionale des races en danger.

2) Préserver les variations génétiques adaptées localement

Compte tenu du rythme auquel la diversité génétique s'érode, il est crucial d'identifier pour ensuite préserver les variations génétiques liées à un phénotype adapté localement. Dans ce contexte, le logiciel *Samβada* avait été conçu pour rechercher des signatures d'adaptation locale via l'étude de l'association génome–environnement. Cependant, le pré- et le post-traitement peuvent être laborieux, et nous avons donc développé un package *R* nommé *R.SamBada*, qui fournit une chaîne de traitement pour la génomique environnementale. Basé sur *Samβada*, il va de la recherche des conditions environnementales jusqu'à l'annotation des gènes. Ainsi, *R.SamBada* facilite l'accès à des analyses *biogéoinformatiques* pour des chercheurs sans connaissance en géographie.

3) Préserver une technique d'élevage traditionnelle, adaptée aux races locales

La préservation des races locales est aussi liée à la conservation de techniques d'élevage traditionnelles, qui en Suisse, comprennent la montée à l'alpage pendant l'été. L'un des principaux effets de l'alpage sur les vaches est la baisse de la production laitière. Nous avons développé un nouveau modèle mathématique pour ajuster la courbe de lactation des vaches alpées et avons testé l'influence des facteurs environnementaux, physiologiques et morphologiques en utilisant cinq millions de relevés laitiers mensuels de vaches Braunvieh. Comparées aux facteurs physiologiques, les variables environnementales montrent un impact limité sur la production de lait à l'alpage, les précipitations au printemps étant néanmoins la plus importante.

Mots-clés

Biogéoinformatique, Ressources Génétiques des Animaux d'Elevage, adaptation locale, races locales, érosion de la diversité génétique, changements climatiques, transhumance, SIG, génomique environnementale, outils intégrés

Table of contents

List of Figures

List of Tables

List of Equations

Chapter 1
Introduction

Livestock production is the largest land-use system on Earth occupying 30% of the world's ice-free surface. This sector contributes 40% of global agricultural gross domestic product, provides income for more than 1.3 billion people and nourishes at least 800 million food-insecure people, and uses one-third of the earth's freshwater. Each year the livestock sector globally produces 586 million tons of milk and 285 million tons of meat. There is probably no other single human activity that has a bigger impact on the planet than the raising of livestock. (Herrero et al. 2013)

Beyond these numbers, FAO (2015) highlights the side-functions of the livestock sector, such as the key role it plays in the sociocultural landscape (religion, shows, sports) and the ecological regulation it provides, particularly in mountainous and arid areas (for instance fight against invasive species and wildfires).

These side-functions are particularly crucial in Switzerland, where one third of the total surface of the country is used and therefore maintained by agriculture, of which one third is located in mountains. In 2017 alone, Swiss livestock produced 474000 tons of meat and 3.8 Mio litters of milk. (FSO 2019)

Nowadays, agriculture, and especially the livestock sector, is undergoing stress due to environmental and structural changes that include modern farming methods aimed at improving productivity. This thesis advocates the use of *biogeoinformatics* (a new field emerging from the combined use of biology, geography and informatics) to cope with the challenges of Farm Animal Genetic Resources (FAnGR), notably erosion of genetic diversity and climate change adaptation. While relying on data from different countries around the world, it will give particular emphasis to Swiss case studies. Indeed, although *biogeoinformatics* can provide interesting insights for livestock management, the variety and complexity of tasks involved hinders a wider usage of this type of analyses. Dedicated tools as well as case studies demonstrating the benefit of the methods would facilitate its adoption.

This introduction describes three of the main challenges faced by the livestock sector, reviews the methods available in *biogeoinformatics* that could be helpful in this context and defines the problem statement and the subsequent structure of the thesis.

Definitions

Farm Animal Genetic Resources (FAnGR): "all animal species, breeds/strains and populations used for food and agricultural production and their wild and semi-domesticated relatives" (Kohler-Rollefson 2004)

Biology: "the study of living organisms, divided into specialised fields that cover their morphology, physiology, anatomy, behavior, origin, and distribution" (Lexico)

Geography: the study of the physical features of the earth and its atmosphere (Lexico)

Informatics: "the science of processing data for storage and retrieval" (Lexico)

Bioinformatics: “the science of collecting and analysing complex biological data […]” (Oxford Dictionary)

Genomics: “the branch of molecular biology concerned with the structure, function, evolution, and mapping of genomes” (Oxford Dictionary)

1 Challenges in the livestock sector

As a consequence of population growth, biodiversity in general is experiencing pressure which tends to decrease its genetic diversity (Sarkar et al. 2006). To address this problem, the United Nations first organised a conference in Rio in 1992, during which 150 countries signed the Convention on Biological Diversity (UN 1992). More specifically focused on farm animals, the Food and Agriculture Organization (FAO) coordinated a conference held in Interlaken in 2007 which lead to the Declaration on Animal Genetic Resources. In this declaration, the Global Plan of Action (FAO 2007) was adopted, which defines strategic priorities aimed at ceasing any further loss of genetic diversity in farm animals. In this plan, several challenges of the livestock sector are highlighted and three of them are described here, as they will be further treated in this thesis.

1.1 Erosion of genetic diversity

The erosion of genetic diversity of farm animals is the main focus of the Global Plan of Action. It arose mainly as a consequence of the gradual substitution of locally adapted breeds with a limited number of highly productive transboundary breeds such as e.g. the Holstein Friesian cattle selected for milk production (Bruford, Bradley, and Luikart 2003). These selective breeding and controlled reproduction of a limited number of high performance individuals have gradually led to a general loss of genetic diversity within breeds (Bruford, Bradley, and Luikart 2003). In poultry for example, genetic diversity of highly industrialised breeds is lower than other breeds of the species (Crawford 1990). These commercial breeds have been shown to require high inputs (e.g. high quality food, medicine) and to be prone to diseases and stress factors (Thrupp 1997).

Furthermore, erosion of standing genetic variation can threaten the species' evolutionary potential. Indeed, some genetic variants can result neutral or even deleterious under the current conditions, while conferring an adaptive advantage with changing selection regimes (see for instance the case of global warming and its consequences). The depletion of variation could therefore affect the species ability to survive to environmental change due to the loss of important adaptive genetic variants (Taberlet et al. 2008). In FAnGR management, the needs to adapt can arise from changing environmental conditions or breeding practice (Notter 1999) and include: extreme temperature, food shortage, disease and water scarcity (FAO 2015).

1.2 Pressure on traditional farming techniques

The Global Plan of Action describes as strategic priority n° 6 the need to “support indigenous and local production systems”, as these are ancient techniques that over the ages became suited for local breeds. The current structural changes faced by agriculture results in an increased pressure on these production techniques, considered as less efficient than modern farming systems, and

consequently enhances stress on local breeds. Particular attention should be given to maintain pastoralism and small farms. (FAO 2007)

One example of such indigenous production system is the transhumance technique, which consists of livestock spending summer months in high mountain pastures, and which plays a considerable role in several sectors. First, it is involved in the preservation of biodiversity as some alpine plants heavily depend on grazing (Herzog et al. 2005), and in creating a very specific ecosystem (Bunce, Pérez-Soba, and Smith 2009; Olea and Mateo-Tomás 2009). Secondly, the transhumance system is also a key aspect for the maintenance of the socio-cultural landscape, including tourism-related activities. Indeed, high mountain pastures represent a unique traditional cultivation technique (Gellrich and Zimmermann 2007). Agricultural land abandonment in mountain areas also results in higher risks of avalanches (Newesely et al. 2000) and wild fires (Gellrich and Zimmermann 2007).

Even though this farming practice plays a considerable role in Switzerland, it is under increasing pressure, being criticised as inefficient (Jurt, Häberli, and Rossier 2015). Nowadays, alpine agricultural areas, including high summer grazing represents 12% of the total area of Switzerland (FSO 2019) and 100'000 dairy cows are brought up in high mountain pastures (OFS 2013). Yet, in recent years, the total surface dedicated to this practice has decreased (Lauber 2013). Mack *et al.* (2008) further investigate how to revert this trend and draw the conclusion that this system heavily depends on subsidies of which should be raised if it is to be preserved it. Several initiatives have been deployed to better understand and maintain this breeding technique, including the trans-disciplinary and international research project AlpFUTUR (Herzog et al. 2009), but some authors deplore the lack of research undertaken to date to better understand the transhumance system (Jurt, Häberli, and Rossier 2015)

1.3 Climate change

The Global Plan of Action also sets several priorities for the establishment of a sustainable use of genetic resources, with the aim of adapting to the consequences of climate change and other structural changes. As a result, environmental and socio-economic trends should be evaluated for the purpose of taking adequate political actions. (FAO 2007)

Today, climate forecasts tend to predict hotter and drier environmental conditions for the future (IPCC 2014). In Switzerland, these changes are expected to amount on average to around 10-20% decrease in summer precipitation and an increase of 2-3°C in temperature by 2050 (C2SM et al. 2011; OcCC 2007). These shifts are likely to impact breeding activities, as cattle are sensitive to heat stress (Hayes et al. 2009) and forage quality will probably decrease (Craine et al. 2010). Additionally, a hotter climate will probably be accompanied with the arrival of new diseases (Tabachnick 2010) and potential water scarcity (Milano et al. 2015).

These changes are particularly challenging in the context of the above-mentioned erosion of genetic diversity experienced by the livestock sector. Yet, some studies provide interesting leads to address these issues, such as the highlight of genetic variation associated with better dairy milk production characteristics when temperature humidity index (THI) is high and when supplemental feeding is low (Hayes et al. 2009). However, it is worth mentioning that these cows with a more efficient milk production in hot climates are not as productive as others when heat stress is low and when concentrates are widely-used.

The limited number of existing studies on the impact of climate change on high alpine grazing shows that biomass productivity and net assimilation rate, while being very site-specific, are most affected in the alpine site, which receives the least amount of annual precipitation (Gilgen and Buchmann 2009; Signarbieux and Feller 2008).

2 Biogeoinformatics

Biogeoinformatics results from the concatenation of three words: biology, geography and informatics. While bioinformatics is a commonly used word, *biogeoinformatics* is a more recent and less frequent expression, used for example in Fautin & Buddemeier (2001) where they "interface geospatial, taxonomic, and environmental data" to study the marine wildlife. We could therefore extend Oxford's definition (see text box) for *biogeoinformatics* as the science of collecting and analysing complex biological and spatial data, such as environmental conditions. Similar to bioinformatics (Xiong 2006), the scope of *biogeoinformatics* can be defined as being the development of new tools and methods to better understand what biological processes are involved in a specific location or situation.

Supported by expert-based decision-making approaches, *biogeoinformatics* is able to take into account animal demographics, adaptation aptitudes and to simultaneously assess the sustainability of breeding activities in areas of interest (Joost 2014; FAO 2015). Besides farm-animal related studies, *biogeoinformatics* has also been applied to manage other living organisms such as crops (Waltman et al. 2004; Patil, Bhat, and Joshi 2007).

Biogeoinformatics calls on specific contributions of the underlying sciences (biology, geographic information, and informatics) as shortly presented in the next paragraphs.

2.1 Biology

Several fields of biological study are employed in this thesis, particularly molecular biology which enables a variety of indices and a wide range of analyses empowering a better understanding of the genome. The use of DNA sequences to scan for molecular markers like Single Nucleotide Polymorphisms (SNPs) is now available at affordable cost through the use of SNP chips (Calus, de Haas, and Veerkamp 2013), revealing sites in the strand of DNA that vary from one individual to another. SNPs data can reveal the mechanism of local adaptation (Edea et al. 2014; de Simoni Gouveia et al. 2017) through a wide variety of approaches. F_{st}-methods for example compare the genetic variance of two groups of individuals, whereas the identification of selective sweeps brings to light regions of the genome experiencing a strong linkage disequilibrium (i.e. correlated SNPs located in a nearby place). Landscape genomics studies in turn takes advantage of the association between the genome and its environment (see next section).

Molecular data is also used to estimate the genetic diversity of a breed. In particular, the inbreeding coefficient, which defines the level of consanguinity in the ancestry of an animal, is highly valued in livestock breeding practice. Generally speaking, the higher the average inbreeding of a breed is, the lower its genetic diversity is (Charlesworth 2003). It can be computed either from genetic data or pedigree information. In FAnGR management, complete pedigree data are often available, while genetic data are often missing (Cunningham et al. 2001). This makes pedigree analyses essential for the monitoring of genetic diversity. The effective population size is an additional suitable parameter, that calculates the theoretical size of the population that would have occurred under random mating (Hedrick 2011). This measure captures the inbreeding rate, the fitness of the

individuals and the genetic erosion of the breed (Gutiérrez et al. 2008). Several genetic- and pedigree-based methods have been developed to estimate it (Cervantes et al. 2011; Groeneveld et al. 2009)

Besides genetic analyses, dairy science and more specifically the study of lactation curves are of special interest here. Lactation curves represent the milk yield (quantity in kg) of a cow during a milking season and have been extensively studied and mathematically modelled (Macciotta, Vicario, and Cappio-Borlino 2005). Yet interestingly, the impact of transhumance to high alpine grazing, while being decisive, has not yet been described in detail.

2.2 Geographic information

Geographic information, and Geographic Information Systems (GIS), also have an important role to play in FAnGR management (Joost 2006; Bertaglia et al. 2007, Joost et al 2010; Paul J. Boettcher et al. 2014; Joost et al. 2015), though its use has not spread to the entire farm animal community yet. The contribution of GIS is fundamental as it creates a bond in the geographical space between the characteristics of an animal and its environmental and socio-economic conditions. In the context of climate change, the information entailed in the location is crucial, as it enables to identify the environmental conditions in which the individual lives and will live using environmental databases and climatic scenarios.

In 2008, FAO and WAAP published a report on production environment descriptors for animal genetic resources. One of the main conclusions was that it was necessary to quickly systematise the recording of breeds' geographical coordinates worldwide in order to enable links to any kind of information available in other geo-referenced databases. Indeed, the management of FAnGR requires complementary data on population and evolutionary genetics, on animal husbandry practices, but also data characterising the socio-economic and environmental conditions of the regions where animals are bred. Only the integration of these different information levels by means of geographical coordinates and GIS is likely to empower the development of FAnGR monitoring systems able to identify endangered breeds.

Joost *et al.* (2010) propose a review of the use of GIS in livestock science, and identify five subfields in which GIS was successfully applied : 1) impact of livestock on the environment 2) management of landscapes and pasture surface 3) disease control/health epidemiology 4) rural economy and development 5) FAnGR conservation.

The last point, being the focus of this thesis, deserves particular attention. Bertaglia *et al.* (2007) identify marginality regions based on geographic information, which are intricately linked to the conservation of locally adapted sheep and goat breeds. Many other studies perform landscape genetics (or landscape genomics) analyses defined by Manel *et al.* (2003) as a way to "provide information about the interaction between landscape features and microevolutionary processes, such as gene flow, genetic drift and selection". This kind of approach has been applied repeatedly to identify SNPs under selection in domesticated animals, such as sheep (Joost et al. 2007; Lv et al. 2014; Vahidi et al. 2016), goats (Colli et al. 2014) and cattle (Stucki et al. 2017). The main principle common to all these studies is to assess the correlation between environmental variables and genetic variations (a description the different methods is available in the next section).

While most landscape genomics analyses make use of general environmental data (i.e. mean annual parameters such as in the WorldClim database; Hijmans et al. 2004), temporal information can also be stored to complement geographic coordinates, so that weather conditions on the day

of measurement can be retrieved. This enables a refined analysis to study the effect of climate on living organisms, which can be applied for example to crop (Magarey et al. 2007), wild species (Masterman et al. 1996) and livestock (Bryant et al. 2007; Ugurlu et al. 2014).

2.3 Informatics

The analysis of complex biological and environmental data as required by *biogeoinformatics* is usually done via statistical, mathematical and algorithmic models (Roy, Pantanowitz, and Parwani 2014).

Statistics are key aspects of landscape genomic studies. A large amount of approaches are available to test the association between a genetic variation and the environment, including LFMM (Frichot et al. 2013), XPCLR (Chen, Patterson, and Reich 2010) and *Samβada* (Stucki et al. 2017). LFMM introduces an unobserved variable representing population structure as latent factor. XPCLR uses Brownian motion to model genetic drift under neutrality and compare it to a deterministic model determining the effect of selection. *Samβada* in turn offers a way to approximate the effect of population structure by constructing a multivariate logistic model including one or several population variables. The identification of regions of the genome adapted to local environments is key to manage livestock in a sustainable way (Hayes et al. 2009).

Mathematics in turn can be very informative in dairy science, with the modelling of lactation curves enabling the investigation of the impact of several factors including climatic conditions, forage quality (and possible concentrates supplement), calving year, calving season, age, service period (days in milk) (Hayes et al. 2009; Tekerli et al. 2000). The transhumance system has an important yet understudied effect on lactation.

In this context, computer science plays an increasingly important role in animal management, mainly due to the large amount of data that have to be processed, particularly with molecular, meteorological and phenotypical data. For example, SNP arrays nowadays reach very high density and their analysis requires high computational power (Calus, de Haas, and Veerkamp 2013). As a consequence, it has become increasingly important to possess suitable methods to process them (Aulchenko, De Koning, and Haley 2007; Stucki et al. 2017), including efficient coding, compiled rather than interpreted programming language and parallel computing. All cited software in landscape genomics, and especially *Samβada*, give special emphasis to High Performance Computing (HPC). HPC is also critical in pedigree analyses, which generate heavy workloads, whereby Poprep (Groeneveld et al. 2009) and ENDOG (Gutiérrez and Goyache 2005) are good examples of software that efficiently perform pedigree analyses. Similarly, dealing with high resolution weather data is a challenging task and requires expertise both in GIS and computer science. In Switzerland for example, 1km-grid for different weather parameters are available on a daily basis (Meteoswiss, n.d.), resulting in a multitude of heavy layers, when various parameters are analysed over several years.

Last but not least, computer science also holds a consequential role in data integration. Databases are efficient means of compiling different sources of information and store them at a common level. Some publically available databases used world-wide in farm animal management are worth mentioning: the Ensembl database provides an annotation of the genomes of many species (human, mouse, livestock, …) with location of the known genes. The AnimalQTLdb is designed to store information on regions of the genomes associated with specific phenotypes of livestock species, while the DAD-IS (managed by FAO, dad.fao.org) and the more advanced FABISnet network (Groeneveld et al. 2007) are meant to store data on biodiversity in agriculture. At a national level,

different organisations in several countries are responsible for the storage of all routinely collected data on livestock (births, inseminations, phenotypes, milk production, movements, health). A few Swiss examples are dbmilch, Tierverkehrsdatenbanken (TVD), Auswertung Lebensmittelsicherheit Veterinärwesen Public Health (ALVPH). Most breeding organisations also manage their own database storing information on pedigree and individual performances of animals.

3 Problem statement and plan of the thesis

On the basis of the observations of this chapter, this thesis will demonstrate **how *biogeoinformatics* can be harnessed for the management of livestock** with special emphasis on three current **challenges** that are

- The erosion of genetic diversity, which requires a constant monitoring to prioritise endangered breeds as well as the identification of locally adapted genetic variants that should be preserved.
- The pressure on traditional farming techniques and more specifically the transhumance system, which might be eased by the understanding of such system and of the effect of different factors on the lactation. This would ease the preservation of this farming technique and its adaptation to future conditions
- Climate change and its consequent impact on livestock, necessitating a better knowledge of problematic environmental variables and useful genetic variations in this context.

Available **methods** in *biogeoinformatics* to address these issues are

From biology

- The computation of the inbreeding and effective size of the population using pedigree or molecular data to monitor the genetic diversity of breed.
- Landscape genomics enabling the identification of genomic regions that are locally adapted to harsh environmental conditions, so as to select animals resilient to future climatic conditions.
- Dairy science to study lactation curves.

From geography

- The assessment of the geographical concentration of the breed using the location of the animal, as being an important factor to determine the level of endangerment of a breed.
- The retrieval of environmental and socio-economic conditions inferred from the spatial (and temporal) location of the animal.

From informatics

- Statistics to test the significance of the association between the genome and environmental conditions and milk production and influencing factors.
- Mathematical modelling of lactation curves.
- High performance computation to treat large molecular, phenotypical and meteorological data.
- Databases to efficiently store and share different types of information.

From what we have observed, despite the undeniable advantages of *biogeoinformatics*, this field is rarely used in the context of farm animal. Consequently, this thesis is an attempt to fill in two

technologic gaps for a subsequent use of *biogeoinformatics* in livestock management and proposes

Two tools to

- Automate the monitoring of FAnGR using a WebGIS platform (GENMON).
- Facilitate the joint use of GIS and biology in the context of landscape genomics with a dedicated *R* package (*R.SamBada*).

Detailed case studies demonstrating the benefit of *biogeoinformatics* using

- Herdbook information and geolocation of animals from three Swiss breeds to test the GENMON platform.
- SNPs and geolocation from hundreds of individuals of Spanish cattle and Moroccan sheep to illustrate the use of *R.SamBada*.
- Milk records and geolocation of mountain-pastured Braunvieh cows to better understand the impact of the alping system on milk production.

To cope with the above-mentioned challenges, **actions** can be taken at **different levels**

- Preservation of local breeds: in particular the monitoring of their genetic diversity to prevent the loss of locally adapted breeds.
- Preservation of locally adapted genetic variations of local breeds that are essential in a context of climate change. Identified variations can then be included in breeding programs.
- Preservation of traditional farming practices suited for local breeds, such as the transhumance from lowland to mountain, with a better understanding of their dynamics and how they are impacted by climate change.

As a consequence, the rest of the document is divided into three parts, each section being a published article, in which I am the first author. A short description of the subsequent chapters is given here.

Chapter 2: Preserving locally adapted breeds

In order to prevent the erosion of genetic diversity, locally adapted breeds should be monitored to prevent their extinction. This is one of the main goals of the Global plan of action for Farm Animal Genetic Resources initiated in 2007 by the Food and Agriculture Organisation of the United Nations (FAO). To this end, we developed the GENMON WebGIS platform, able to monitor FAnGR and to evaluate the degree of endangerment of livestock breeds. The system integrates various sources of information that are linked with the help of geographic information: pedigree and introgression, geographical concentration of animals, cryo-conservation plan and the sustainability of breeding activities based on socio-economic data as well as present and future land use conditions. A multi-criteria decision tool supports the aggregation of the multi-thematic indices mentioned above using the MACBETH method, which is based on a weighted average using satisfaction thresholds. The score can be visualised on a geographic map and allows a fast, intuitive and regional identification of breeds in danger. Appropriate conservation actions and breeding programs can thus be undertaken in order to promote the recovery of the genetic diversity in livestock breeds in need. GENMON is an open source software, designed as a monitoring tool to reach subjective decisions made by a government agency. The use of the platform is illustrated by means of an example based on three local livestock breeds from different species in Switzerland.

Chapter 3: Preserving locally adapted genetic variations

Considering the pace at which genetic diversity is being eroded, it has become urgent to identify and then preserve important genetic variations linked to locally adapted phenotypes. In this context, the recent Samβada software was designed to search for signatures of local adaptation through the study of genome–environment association. However, pre- and postprocessing of data for this analysis can be labour-intensive, preventing a wider uptake of the method. Consequently, we developed the *R.SamBada R* package providing a pipeline for landscape genomic analysis based on *Samβada*, spanning from the retrieval of environmental conditions at sampling locations to gene annotation using the Ensembl genome browser. As a result, *R.SamBada* standardises the landscape genomics pipeline and eases the search for candidate genes of local adaptation, granting access to *biogeoinformatic* analyses to researchers with no skills in geography. The efficiency and power of the pipeline is illustrated using two examples: sheep populations from Morocco with no evident population structure and Lidia cattle from Spain displaying population substructuring. In both cases, *R.SamBada* enabled rapid identification and interpretation of candidate genes, which are further discussed in the light of local adaptation. The package is available in the *R* CRAN package repository and on GitHub (github.com/SolangeD/R.SamBada).

Chapter 4: Preserving a traditional farming technique suited for local breeds

The preservation of locally adapted breeds is also strongly linked with the conservation of their native environments and of the production system in which they are involved. In Switzerland, these environments include high alpine pastures, grazed by livestock animals brought to the mountains during summer. This transhumance system plays a considerable role in preserving both local biodiversity and traditions, as well as protecting against natural hazard. In cows, particularly, milk production is observed to decline as a response to food shortage and climatic stress, leading to atypical lactation curves that are barely described by current lactation models. Here, we relied on five million monthly milk records from over 200,000 Braunvieh and Original Braunvieh cows to devise a new model accounting for transhumance, and test the influence of environmental, physiological, and morphological factors on cattle productivity. Climatic variables were retrieved from available high-resolution meteorological data at the sampling location, shortly before records were taken. Counter to expectations, environmental conditions in the mountain showed a globally limited impact on milk production during transhumance, with cows in favourable conditions producing only 10% more compared to cows living in detrimental conditions, and with precipitation in spring and altitude revealing to be the most production-affecting variables. Conversely, physiological factors as lactation number and pregnancy stage presented an important impact over the whole lactation cycle with 20% difference in milk production, and alter the way animals respond to transhumance. Finally, the considered morphological factors (cow height and foot angle) presented a smaller impact during the whole lactation cycle (10% difference in milk production). The present findings help to anticipate the effect of climate change and to identify problematic environmental conditions by comparing their impact with factors that are known to influence lactation.

Chapter 5: General discussion

To conclude the thesis, this chapter will review the main outcomes and perspectives of the three articles presented and give concluding remarks.

Chapter 2
Preserving locally adapted breeds

The erosion of genetic diversity has been identified in the introduction as one of the main challenges of the livestock sector. In order to mitigate this loss, the FAO initiated in 2007 the Global Plan of Action (FAO 2007a), which identifies as a first step the monitoring of genetic diversity with the creation of prioritisation strategy and Early Warning Systems (EWS). In the field of livestock breed conservation, a few one-off studies have been conducted on a selected set of breeds with a special focus on different aspects: Barker et al. (Barker 1999) for example assess the between-breed diversity to sustain breeds that add the most genetic variance while Fabbri et al. (2019) identify problematic mating practice. More global EWS are often very basic, such as the DAD-IS database from FAO which only includes the number of breeding animals, or focus on a single aspect of conservation, namely genetic diversity (Duchev, Distl, and Groeneveld 2006). A true multi-criteria analysis accounting for socio-cultural aspects is rarely present (Verrier et al. 2015) and the inclusion of geography is found in only one example (L. Alderson 2009).

The monitoring of genetic diversity constitutes an excellent application where *biogeoinformatic* procedure can be used for FAnGR management. Indeed, beside the evaluation of genetic diversity, the inclusion of the geographic concentration of the breed as computed by Alderson and the assessment of the local sustainability of agriculture can add a new perspective to determine whether a breed is at risk or not. Furthermore, High Performance Computing is required to perform the analysis of large and complex pedigree files.

The goal of this chapter is to propose a tool to monitor the endangerment level of breeds. In Switzerland, the whole pedigree as well as several additional information are already routinely collected and stored for most local livestock breeds, but resources to enhance these data are lacking. The tool should therefore constitute a quasi-automatic pipeline that easily highlights problematic breeds or regions. The platform is tested with three Swiss local breeds.

Since the publication of this article and its associated platform, two studies are worth mentioning in the context of livestock conservation. Wainwright (2019) focuses on the idea of economical optimisation and performs a sensitivity analysis to highlight the impact of choices, such as weights of criteria. This kind of analysis is relevant for the presented platform since its concept relies on the use of weights to prioritise criteria. Furthermore, conservation efforts through the use of subsidies have been proven effective to increase the number of bred animals but still presents challenges (Gicquel et al. 2019). Knowing the impact of subsidies to endangered breeds is crucial, as the use of a monitoring tool will lead to prioritisation of breeds, whereby the most endangered breeds can then be supported by increased subsidies.

As the first author of this article, I completed most of the tasks, both in the implementation of the platform and in the writing of the manuscript. The other authors provided advices on the methods to be used and assistance in writing the paper, whereby a few paragraphs were written by experts in specific fields. The complete list of contributions is available at the end of the chapter.

Duruz, S., Flury, C., Matasci, G., Joerin, F., Widmer, I., & Joost, S. (2017). A WebGIS platform for the monitoring of Farm Animal Genetic Resources (GENMON). *PloS one,* 12(4).
doi: 10.1371/journal.pone.0176362

1 Introduction

1.1 Erosion of livestock genetic resources and global strategy for the management of Farm Animal Genetic Resources (FAnGR)

Agricultural biodiversity is the basis of the functioning and the productivity of agricultural systems and is thus essential, for example to satisfy human nutritional needs. Nowadays, agriculture is facing increasing stress due to structural changes and modern farming methods aimed at improving productivity. In the livestock sector, locally adapted breeds have been gradually substituted with a limited number of highly specialised transboundary breeds (such as. the Holstein Friesian cattle selected for milk production) (Bruford, Bradley, and Luikart 2003) requiring high inputs (e.g. high quality food, medicine) and which are prone to diseases and stress factors which naturally occur (Thrupp 1997). Selective breeding and controlled reproduction of a limited number of high performance individuals have gradually led to a general loss of genetic diversity within breeds (Bruford et al. 2003). This might reduce productivity through a drop in individual fitness in non-optimal environments, and over the longer term the capacity of the breeds to evolve and adapt to (changing) local environmental conditions (such as climate, pests or diseases; Notter 1999).

In order to counteract the current trend of erosion and underutilisation of animal genetic resources, the Food and Agriculture Organisation of the United Nations (FAO) initiated a global strategy for the management of Farm Animal Genetic Resources (FAnGR) in 2007 (FAO 2007b). This strategy has been recently reinforced by a recent second report on the state of the world's animal genetic resources (FAO 2015). The ultimate goal of this plan is to lead to policies aimed at promoting and conserving livestock biodiversity and using animal genetic resources in a sustainable way (e.g. priority measures for a sustainable use, development and conservation of animal genetic resources). The FAnGR strategy was discussed and fixed during the Interlaken Conference in 2007 (Interlaken Declaration) and since then, the governments of UN countries are encouraged to implement it. The main objectives of this plan are to identify genetic resources, characterise and protect them in order to stop further genetic erosion and to promote genetic diversity in farm animal resources. An important step to reach this goal is to develop better indicators that can be applied to monitoring genetic trends in domestic populations (Bruford et al. 2015), and to use monitoring systems to identify endangered breeds, to prioritise them, and to initiate as well as support conservation programs (Boettcher et al. 2010).

1.2 A multi-criteria approach

One of the major challenges is the definition of meaningful criteria to identify endangered breeds. FAO created a scale of endangerment based on the number of breeding females and males (FAO 2007c); this approach has the advantage of being easily implemented but is a simplistic view of the problem. Several other systems to categorise endangered livestock breeds have been developed

on national (Lawrence Alderson 2009; BMELV n.d.; Ruane 2000; Verrier et al. 2015) and international levels (Alderson 2003; Avon 1992; Gandini et al. 2004; Loftus and Scherf 1993; Reist-Marti et al. 2003; Simon and Buchenauer 1993). With a few exceptions (Lawrence Alderson 2009; Gandini et al. 2004; Verrier et al. 2015), existing systems rarely provide a standardised definition and measurement of the most significant factors (Lawrence Alderson 2009).

In accordance with the FAO Global Plan, the top strategic priority is given to the characterisation of animal genetic resources (AnGR), to the monitoring of trends and risks to these resources, and to the establishment of breed endangerment Early Warning Systems (EWS) (FAO 2007b). In order to obtain an overview of the diversity, status and trends of animal genetic resources, the measurement of genetic diversity is a basic component of monitoring systems. However, beside genetic diversity, other criteria should also be considered. For instance, Alderson (Alderson 2010) specifies that introgression - the process of uncontrolled entrance of genes from another gene pool through mating with another breed (Giuffra et al. 2000) - is another important and relevant criterion since it dilutes specific traits that might be worth conserving. Moreover, the UK monitoring system accounts for geographical concentration of the breeds, as much as a breed that is clustered in a small region is more vulnerable to epidemics (Alderson 2010). In addition, the presence of cryo-conserved gametes is an important element to consider, as it can refresh genetic resources of very small breeds (Meuwissen 2009) and even help to bring a breed back to life after a critical point has been reached (Curry 2000). Finally, according to the FAO protocol, the supervision of the genetic diversity should be completed by the establishment of national sustainable use policies taking into account environmental and socio-economic aspects, including demographic changes, climate change, and conducting economic and cultural valuation (Strategic Priority 3, FAO 2007b).

1.3 Data integration: Geographic information system (GIS) and multi-criteria decision analysis (MCDA)

Animal genetic resources have to be monitored and conserved at local, regional and global levels (FAO 2007b) and thus geography is an important component in this effort. Intriguingly, despite the issue of AnGR conservation being composed of entities which are totally embedded in lands and distributed over territories, geography is only considered in the UK system (Lawrence Alderson 2009). However, this monitoring system does not assess the level of sustainability of regional or local breeding conditions, comprised of socio-economic, socio-demographic, and environmental characteristics.

Here we propose the application of a GIS-based multi-criteria analysis for the integration of multi-disciplinary data in order to monitor animal genetic resources at different scales. Geographic Information Systems (GIS) offer an appropriate basis in a monitoring perspective as it integrates different categories of information (demography, phenotypes, husbandry practices, socio-economy, natural environment, etc.) on different geographical scales (local, regional or global) (Joost et al. 2015). Furthermore, GIS analysis exhibits other advantages such as a direct comparison between available thematic layers according to geographic coordinates, and the production of valuable outputs like maps, graphs, and tables) (Joost et al. 2010).

Whether combined with the use of GIS or not, multi-criteria decision analysis (MCDA) has often been applied to support decisions involving environmental issues and biodiversity conservation (Bertaglia, Joost, and Roosen 2007; Huang, Keisler, and Linkov 2011; Sarkar et al. 2006; Verrier et al. 2015). However, the application of GIS-based multi-criteria analysis are rare in the domain of livestock species conservation.

In parallel, when dealing with multi-criteria analyses, we are often confronted to the problem of non-commensurability, which occurs in situations where criteria are assessed onto different and incomparable scales of measure (Martinez-Alier, Munda, and O'Neill 1999). This difficulty is often circumvented by means of mathematical tools. Within this type of approach, we find many methods, including the Multi-Attribute Utility Theory (MAUT; Keeney and Wood 1977) or the Analytic Hierarchy Process (AHP; Saaty 1990) methods, as well as the use of weighted average methods. However, criteria are not necessarily comparable, and this is the reason why methods referred to as "outranking methods" were introduced, giving the possibility that two scenarios might be incomparable (Roy 1991; Roy and Vincke 1981). This situation is recurrently encountered when dealing with FAnGR evaluation, in which criteria cover a wide variety of fields measured with different units, from pedigree information to socio-economic data. Nevertheless, outranking methods also present drawbacks, such as the difficulty in explaining and implementing them as well as a reduced performance when the number of variants is very large. These are the reasons why here we consider the MACBETH method (Costa, Bana, and Vansnick 1994), which is a weighted average method with satisfaction thresholds defined by experts in the disciplines considered (see section 2.3).

1.4 GENMON: a WebGIS platform to monitor breed endangerment

To cope with the challenge of the identification of endangered breeds, we propose an easy-to-use WebGIS platform (GENMON), designed to facilitate decision-making which should favor sustainable use and conservation of livestock breeds via the integration of five important categories of information: pedigree analysis, introgression, geographical concentration, cryo-conserved material and agriculture sustainability (this being calculated on the basis of socio-economic and environmental data). Investigated breeds are then ranked in order to identify the most endangered ones using weighting of the various criteria. The GENMON application has been designed in the Swiss context and uses data available in this country; however, the system can easily be adapted to the data available in other countries.

The usefulness of the GENMON approach is demonstrated here through its application to three local livestock breeds, the Swiss Original Braunvieh cattle (OBV), the Valais Blacknose sheep (VBN) and the Franches-Montagnes horse (FM). The OBV is from central Switzerland, not ranked among endangered breeds but under supervision because of its international interest due to valuable genetic heritage (FOAG 2002). The VBN is a sheep breed mainly reared in Valais, which is recognised for its genetic uniqueness (Burren et al. 2014; Glowatzki-Mullis et al. 2009). The FM horse breed is the only Swiss native horse breed (Glowatzki-Mullis et al. 2006) and is mostly bred in the Jura mountains. For reproducibility reasons, a fourth breed has been simulated (SIM), whose pedigree has been made publically available (see section 3.1.2)

2 Materials and Methods

The process implemented in GENMON is based on multiple stages and its overview is represented in Fig 1, whereas a more detailed description is given in Fig 2. GENMON relies on the aggregation of indices (pedigree information, introgression, geographic distribution, cryo conservation plan and socio-economic and environmental information) into one final score (Fig 1). The process as a whole (Fig 2) begins with the upload of different sources onto the WebGIS portal. These data are geographic borders (ZIP-codes and municipality areas), herdbook information (pedigree data), general information on the breed (such as cryo-conservation or cultural value), municipality-based

socio-economic and environmental data, and land use change scenarios. A set of criteria is obtained by slightly modifying the inputs (such as the trend of the number of farms being computed out of current and past number of farms; see Fig 2 for more examples), and are then aggregated at the breed and at the municipality levels before the ranking based on a global index of sustainability characterising breeds can be processed.

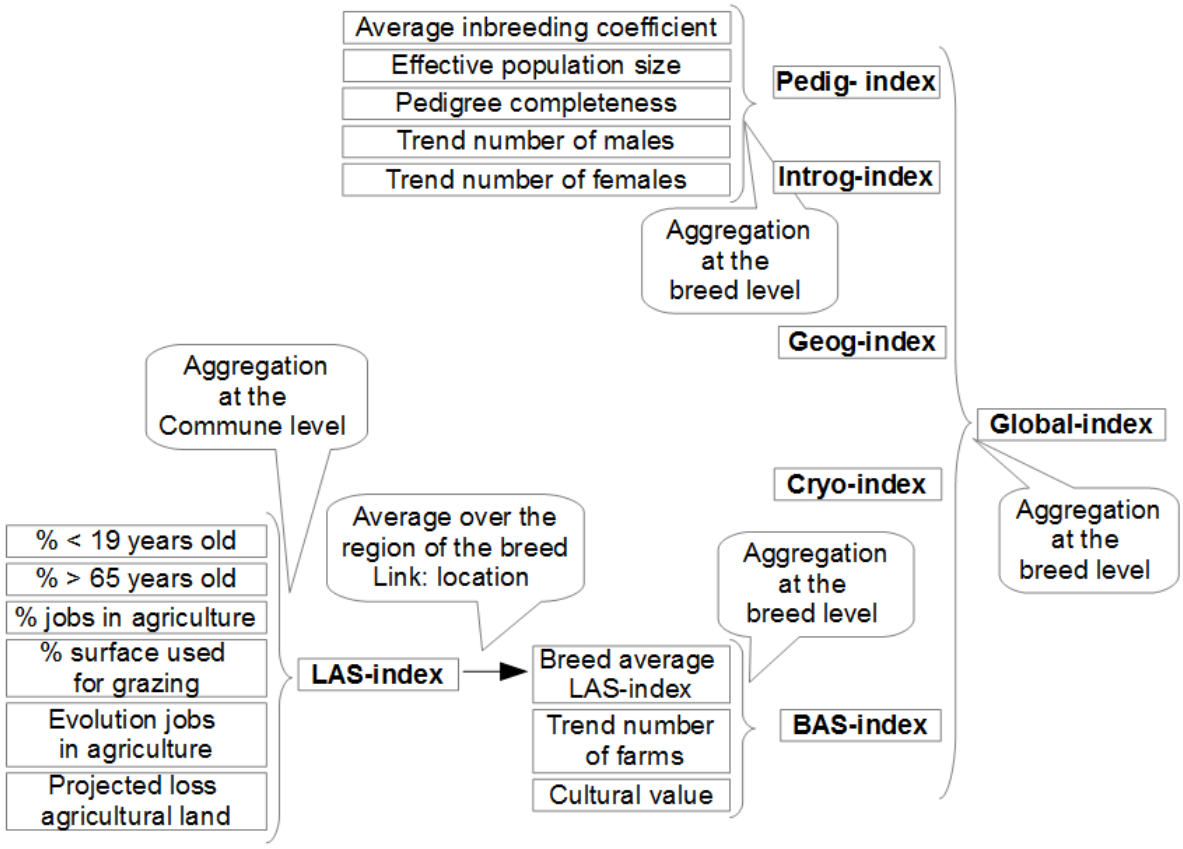


Figure II. 1: Simplified GENMON process. GENMON takes into account five main categories (or indices) aggregated into one final score: pedigree information (pedig-index); introgression (introg); geographic distribution (geog); cryo conservation plan (cryo); socio-economic and environmental information (BAS, standing for breed agriculture sustainability). Some of these indices come from an aggregation of criteria themselves.

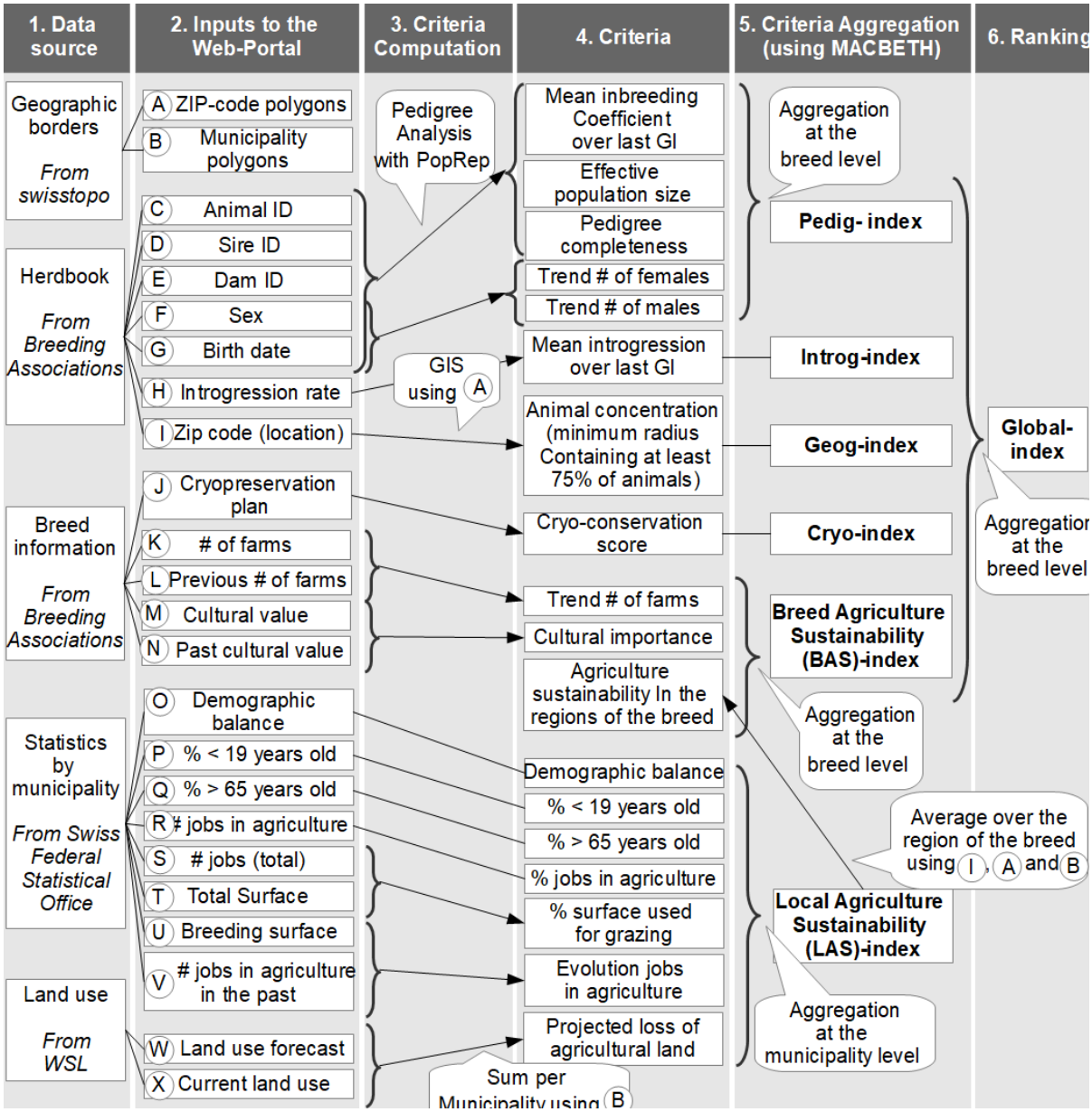


Figure II. 2: Overall GENMON Process. The process starts with data input followed by criteria processing, integration and aggregation; GI: generation interval, GIS: Geographic Information System, Pedig-Index: index accounting for pedigree and genetic diversity, Introg-Index: introgression index, LAS/BAS Index: Local/Breed Agriculture Sustainability indices, accounting for socio-economic and environmental sustainability of breeding conditions; swisstopo is the Swiss Federal Office of Topography (http://www.swisstopo.admin.ch/, WSL is the Swiss Federal Institute for Forest, Snow and Landscape Research.

2.1 Data

In this section we list and describe the different input data (see Fig 2, input column), their source and how they are pre-processed to extract the information for the different criteria.

2.1.1 Geodata: Swiss municipalities and ZIP-codes

The GENMON application has been developed to integrate different categories of information in order to monitor animal genetic resources. All categories have a geographical component and the GIS-based multi-criteria analysis platform developed allows the different data sources to be linked and compared according to geographic coordinates (concept of spatial coincidence). Thus geographical units are key to this platform. Here we use a shapefile containing the geographic coordinates of Swiss ZIP-code areas corresponding to the situation in June 2013 (Swisstopo 2014a) as defined by Swiss Post (http://www.post.ch/). With this file, the animals to be geo-referenced according to the unique identifier (ZIP-code) of the mail delivery area where they are bred. The shapefile contains 4,191 polygons covering the whole country. In parallel, a shapefile with 2,564 municipalities (2013) is used (Swisstopo 2014b) to geo-reference the socio-economic data characterising the territory, which is not available at the ZIP-code level. The link between socio-economic data and their corresponding polygons is made through the unique identifier of municipalities as defined by Swiss Statistics (www.statistics.admin.ch/).

2.1.2 Animal and breed information

The data used to characterise investigated animals are summarised in Table 1. They include pedigree information (dam, sire, sex, birthdate, and introgression) as well as the ZIP-code where the animal was born. The standard format is given in the third column (Type), but the user also has the option of specifying the format of the data if it does not meet the standard format (see remark in the fourth column).

Table II. 1: Description of the variables to be provided from the breeding organisations at the individual level.

Parameter	Input	Type	Remark
C	Animal ID	Text	
D	Sire ID	Text	
E	Dam ID	Text	
F	Sex	M/F	Or as specified
G	Year of birth	Eg. 2009	The whole date can also be specified
H	Introgression	Real [0;1]	Or Real [0;100]
I	ZIP code	e.g. 3096	

From Table 1, the animal, sire and dam ID as well as the sex and birthdate (parameters C to G) are used to run the pedigree analysis with the PopRep module (Groeneveld et al. 2009). This analysis calculates the mean inbreeding coefficient by year, the generation interval and the effective population size. The introgression of each animal (input H) is used to compute the mean introgression over the last generation interval.

The ZIP-code (input I) is used for two purposes. On the one hand it makes it possible to link animals with socio-economic variables at their respective locations, and on the other hand it enables the determination of the geographical concentration of breeds (see section Geographical concentration). This parameter is computed with the help of spatial SQL (structured query language). Given the lack of a precise position for each individual, the centroid of the ZIP-code polygons containing the animals is used as an approximation.

The introgression rate of each animal corresponds to the fraction of animals from other breeds in the pedigree. This individual rate is used to calculate the average introgression rate of the breed over the last generation interval. GENMON requires the introgression rate of each animal to be computed before being uploaded onto the database.

The Swiss Original Braunvieh data were provided by the Swiss Brown Cattle Breeders' Federation (Braunvieh Schweiz n.d.) and consisted of a pedigree file containing 94,099 animals born between 1923 and 2014 (56% during the last decade). For the Franches-Montagnes breed, data were made available by the Swiss Federation of the Franches-Montagnes horses (FM-CH n.d.) and the herdbook included 46,166 animals, born between 1831 and 2013, (67% during the last decade). For the Valais Blacknose, data were produced by Swiss sheep breeders organisation (SSZV n.d.) and contained 110,584 sheep born between 1910 and 2012 (86% during the last 10 years). The fourth simulated breed (SIM) was obtained from an existing pedigree in which we changed IDs, birthdates, introgression rates and geographic distribution of all animals. Some individuals have been removed while others have been shuffled in the pedigree. The final pedigree contained 65,664 simulated animals born between 1920 and 2016 (63% during the last ten years). Though this breed does not represent any specific species, it was entered in the system as a pig breed. Therefore, weights and thresholds are assigned to it as if it were a pig breed. The pedigree is available at http://doi.org/10.5281/zenodo.220887 .

Additionally, the breeding associations are also asked to provide general information about the breeding activities listed in Table 2.

Table II. 2: Description of the variables required to characterise the breed to be monitored, provided by the breeding association.

	Input	Type	Remark
J	Cryo-conservation management plan	Presence of frozen semen (yes/no) and of a real cryo-conservation management plan (yes/no)	
K	Number of farms	Integer	To compute the trend of the number of farms
L	Number of farms 5 years ago	Integer	
M	Cultural value	Does the breed have a cultural value (yes/no)	To compute the cultural value score
N	Past cultural value	Did the cultural value of the breed decrease in the recent past (yes/no)	

The information on the cryo-conservation (input J in Table 2) is used to compute the Cryo-index (see column criteria aggregation of Fig 2).

The current and past number of farms (input K and L) are used to compute the trend of the number of farms, while the current and past cultural value (input M and N) will give a cultural value score (0 if it has no cultural value, 0.5 if the value does exist but is decreasing and 1 if the value exists and is stable) and all four inputs are used in the BAS index. Cultural value include criteria such as antiquity, role of the agricultural system, farming techniques, role in landscape, gastronomy, folklore and artistic expression (Gandini and Villa 2003).

2.1.3 Socio-economic and environmental data over the Swiss territory

Data used for the socio-economic and environmental characterisation of breeding locations are summarised in Table 3. They include statistics on demographic facts (demographic balance, percentage of young and old people), the importance of agriculture (number of jobs in agriculture compared to the total number of jobs, the surface used for animal breeding compared to the total surface of the commune, past number of jobs in agriculture) as well as current and forecasted land use.

Table II. 3: Data input and characteristics for socio-economic and environmental assessment.

	Input	Type	Remark
O	Increase/decrease in population over the last 2 years	%	
P	Percentage of the population younger than 19 years	%	
Q	Percentage of the population older than 65 years	%	
R	Number of jobs in the primary sector		To compute the percentage of jobs in the primary sector
S	Total number of jobs (all three sectors)		
T	Total surface of the commune	km^2	To compute the percentage of grazing surface
U	Surface used for animal breeding	ha	
V	The number of jobs in the primary sector from a previous year		To compute the evolution of jobs in the primary sector considering two years
W/X	Current land use and land use forecast		WSL

Unless specified in the "Remark" column, these variables were obtained from Swiss Statistics (http://www.bfs.admin.ch/bfs/portal/en/index.html).

Socio-economic data used for the analyses (input O to V in Table 3) are provided by Swiss Statistics: they were extracted from the regional portraits of municipalities for the year 2014 (Swiss Federal

Statistical Office 2014) except the surface used for animal breeding (input U) which comes from STAT-TAB (2012), a dynamic interface specifically designed to download data from the Swiss Statistics Office(Swiss Federal Statistical Office 2012) as well as the previous number of jobs in the primary sector (input V) which was extracted from STAT-TAB (2010).

Land use GIS-layers (W and X) (Price et al. 2015) were also used to calculate the percentage of agricultural land that would be lost in the future, which gives an estimation of future regional sustainability of agriculture. In their original format, they consist of six files (current state of land use plus five scenarios for 2050) containing ASCII grids with cells of 1 ha spatial resolution. Projections according to five scenarios are available. Some of them are based on different policies as described by IPCC (Nakicenovic et al. 2000). With the intent of facilitating the use of GENMON, only one scenario (the "trend" scenario") is taken into account. It is calculated using a linear interpolation of the total area of each land use type based on the land statistics of 1985, 1997 and 2009. This scenario seems to be a reasonable assumption, given that unless very strict policies are set up, the trend that has been noticed over the last few decades is likely to continue. A comparison between the present and future states is carried out outputting a percentage of agricultural land loss per ZIP-code.

Note that the Local Agriculture Sustainability (LAS) Index is calculated for all municipalities (including those without breeding activities). A description of how the information at the municipality level is linked to information at the breed level is given in the section Multi-criteria aggregation.

2.1.4 Integration of geographic data

The data input described above shows various geometry types. Therefore, a succinct description of the method used to match the different spatial units is proposed in Fig 3. All components are brought to the ZIP code level: The links are done either by joining attributes (ZIP code number, commune unique identifier) or according to geometries.

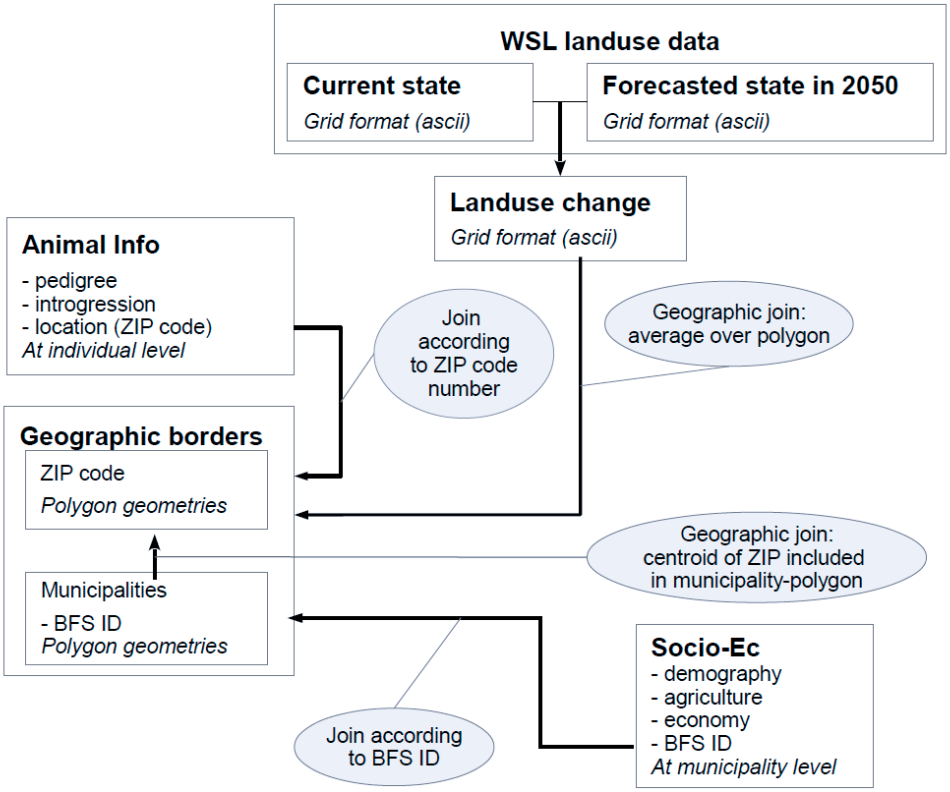


Figure II. 3: Link between the different geographic data types (point, polygons and grids). All components are brought to the ZIP code level. The links are done either by joining attributes (ZIP code number, BFS ID) or according to geometries. BFS ID: unique identifier from the statistical office.

2.2 Selection of relevant criteria

In this section we describe the criteria included in GENMON in order to quantify aspects that are important to evaluate and monitor the conservation status of the breeds under study. They belong to five categories: 1) pedigree-related issues and genetic diversity, 2) introgression, 3) geographical concentration 4) cryo-conservation plan and 5) agriculture sustainability. The selection of criteria has been discussed with a group of 12 experts, scientists and professionals whose activities are related to livestock breeding and management, within a workshop organised for this purpose. The size of the group is justified by the need to favor discussion and emergence of ideas between experts. The description of the discussion course is made available in the supporting information (S1 Appendix).

2.2.1 Pedigree Analysis

Pedigree information can be used to analyse the genetic structure of respective populations (Groeneveld et al. 2009). Since in most countries (including Switzerland) DNA-sampling in livestock, and especially in local livestock breeds, is not currently performed on a regular basis and is not available for marker-based genetic analysis, pedigree data is a valuable alternative to approximate genetic diversity within populations (Cunningham et al. 2001). Many researchers have already investigated the problem of pedigree analysis and there are several software solutions that can be used for this purpose (e.g. PopRep; Groeneveld et al. 2009). In pedigree-based methods, the inbreeding rate can be estimated from pedigree data, which is then used to estimate the effective

population size (Ne). The inbreeding coefficient (*F*) is a measure of the relatedness of ancestors (Rousset 2002), and the effective population size (Ne) is the number of breeding individuals in an idealised population (see for example Alderson 2003; Beissinger and McCullough 2002). The effective population size is an essential parameter in conservation genetics and population management because of its direct relationship with the level of inbreeding, fitness and the amount of genetic variation loss in populations (Caballero and Toro 2000). In consequence, these two parameters (*F* and Ne) are calculated and taken into consideration as criteria in GENMON. Given that the inbreeding coefficient is computed at the individual level, it is necessary to perform an average of this coefficient. In order to alleviate problems arising from missing or incomplete pedigree data in certain years, here we compute the average inbreeding coefficient over the last Generation Interval (GI, average age of parents when giving birth to their offspring, Groeneveld et al. 2009), as an alternative to the average the last year for which data is collected. Nevertheless, this choice has the disadvantage of reducing the relative importance of recent data resulting in a greater inertia of the coefficient, which makes it more difficult to identify sudden changes in the pedigree structure. Admittedly, the inbreeding coefficient has known drawbacks in specific situations (e.g. after a bottleneck, the inbreeding coefficient will not decrease even when the effective population size rises), but we decided to retain this coefficient, as it is a value that breeders are familiar with. Breeder's involvement into the process is essential to successfully undertake the whole GENMON process.

As regards the effective population size (Ne), several methods have been proposed and output different values with the same data (Gutiérrez et al. 2008). This is the reason why GENMON proposes four different Ne values based either on the evolution of inbreeding or on coancestry (see Groeneveld et al. 2009 for more details). Poprep calculates these values for every year, each time using animals in the last generation interval. Only the value of the last year is taken into account in the computation of the pedig-index, corresponding in fact to the last generation interval.

We also included the pedigree completeness as a criterion to maximize, in order to counterbalance the fact that breeds with incomplete pedigree will artificially achieve a low level of inbreeding. While it is true that such breeds are not necessarily endangered, GENMON will lower their final score to emphasise potential problems concealed by their incomplete pedigree. This criterion is computed as the average pedigree completeness at the sixth generation (output by poprep) over the last generation interval. It is necessary to take deep pedigree completeness because the first generations are usually almost 100% complete, which does not discriminate between complete versus incomplete pedigree.

In parallel to the PopRep analysis, the trend of the number of females and males over the last five years is also computed, giving an insight into the evolution of the breeding practices.

2.2.2 Introgression

The introgression rate is entered by the user for each animal. It corresponds to the percentage of foreign blood (based on the fraction of ancestors in the pedigree belonging to other breeds) per individual. Like the inbreeding coefficient, the average computation takes place over the last generation interval.

2.2.3 Geographical concentration

To quantify the geographical concentration of breeds, Alderson (Alderson 2010) proposed to compute the smallest circle containing at least 75% of the animals, centered on the centroid of animal positions.

2.2.4 Cryo-conservation plan

The presence of a cryo-conservation plan is also taken into account via a score bounded between 0 and 1. If such plan exists and follows the FAO guidelines for the given specie (FAO 2012) , the score for the breed is 1. Conversely, if the material to be frozen are collected in an uncontrolled manner, the score is lowered to 0.5. Indeed, if the individuals from which gametes will be collected are not chosen with care, close kinship between selected animals can possibly exists and the value of the cryo-conservation is lessened (Boettcher et al. 2005). Ultimately, if no gametes are cryo-conserved, the score is 0.

2.2.5 Breed Agriculture Sustainability

To calculate the agricultural sustainability for each breed, we take two components into account: statistics on regions where the animals are reared (which leads to a local agriculture sustainability index) and general breed information. The first component, i.e. local agricultural sustainability index (i.e. a score quantifying how sustainable agriculture is in a given municipality), is assessed by an approach inspired by Bertaglia *et al.* (Bertaglia et al. 2007) who quantified the marginality of a region (i.e. areas where possible land uses are relatively limited) combining land use, demographic and socio-economic data using a deliberative approach for variables selection. Here we replaced this deliberative step by a representative panel of experts (see Fadlaoui et al., 2006) with complementary skills (ecology, animal production, agricultural science, socio-demography). The goal is to assess sustainability, represented by the three constituent parts of sustainable development (social, economic and environmental, Rasul and Thapa 2004) and to include at best information directly characterising agriculture and breeding activities. On this basis, a series of discussions between the 12 experts (see S1 Appendix) involved in the present research resulted in the selection of seven variables. Three variables are related to the socio-economic situation of the municipalities: demographic balance, percentage of inhabitants younger than 19 years and older than 65 years respectively, while the other variables account for rural and farming-related features describing municipalities: percentage of active people in the primary sector, percentage of surface used for breeding activities, employment trend in agriculture in the past years (we used data from 2010 compared to the situation in 2012) and projected agricultural land loss (the percentage of agricultural land lost by 2050, as calculated by Price et al., 2015). The underlying concept represented by each variable is explained in the next paragraph.

Regions with a too negative demographic balance (i.e. demographic changes through natural change of population and migration) are often marginal regions where cattle breeding activities can be difficult to pursue (Bertaglia et al. 2007). Negative consequences are expected for breeders and farmers of the municipality since, if a trend of depopulation exists, it is likely that the region will be subject to a lack of manpower. Indeed, on top of the important problem represented by heirs (i.e. manpower) abandoning the family farm, regions facing man power shortage are usually economically not attractive and economic activities (including agriculture) will generally decrease. Regarding the age structure, a municipality with a high proportion of older and retired inhabitants

will face financial problems in the future if there is no migration of professionally active younger people. This could lead to a tax increase in order that municipalities are still able to pay the costs for infrastructure, pensions or for hospital care. Indeed, Alderson (Alderson 2010) proposed to directly use the age of the breeder as a criterion, which would be a valuable alternative in our case. However, this parameter is difficult to evaluate and the corresponding data is not available in Switzerland. Then, the percentage of farmers provides a trustable insight into the predisposition of the municipality for activities in the primary sector. In fact, such an indicator reveals the available manpower existing in a given region, a crucial condition for the establishment and preservation of farming activities. Furthermore, a high percentage of farmers in a municipality is a sign of well-developed agricultural practices. This is an indication that the sustainability of breeding activities is likely to be ensured in such regions. Similarly, the percentage of surface used for cattle breeding activities suggests which regions of Switzerland are suited for this kind of activities. Municipalities with a large number of cattle farms will be those where long-term sustainability is higher.

As regards the criterion used to translate employment trend in agriculture, the difference of the percentage of people working in agriculture among years (here between 2010 and 2012) informs about current employment dynamics in this field of activity. Finally, as a last criterion, the agricultural land loss projection from Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) for 2050 is used (Price et al. 2015). This is a relevant criterion to assess the sustainability of breeding activities. Two kinds of land losses are mainly considered: agricultural land abandonment with subsequent forest growing as well as urban sprawl. Besides, the projections also take into account the consequences of the predicted climate change.

The second component at the breed level assesses breeding practices including the evolution in time of the number of farms in which animals are reared. This criterion adds another relevant dimension to the criterion considering the trend of the number of animals (described in section Pedigree Analysis). Indeed, if a large number of animals are still alive but are reared in a sole farm, the breed can be considered as endangered, since the future of the breed practically depends on the decisions of a single person. Additionally, the cultural value of the breed and its evolution is also considered, as by definition a breed being culturally significant will be better sustained by local people, so as to prevent its extinction (FAO 2007b).

In the following section, we explain how these two components (evolution of number of farms and of cultural values of the breed) are aggregated and how they relate to other criteria.

2.3 Multi-criteria aggregation

This section describes the process of criteria aggregation, necessary to allow comparison among breeds. As stated in the introduction, several methods exist to handle the problem of integrating various criteria, and the one chosen here is the MACBETH method, which is a weighted average using satisfaction thresholds (Costa et al. 1994). Using this approach, decision-makers have the possibility to establish the importance (weight) and a minimum and a maximum expected value for all criteria considered. For each criterion x_j (Fig 1 lists which criteria are included in which (sub)-index), we defined a null satisfaction threshold tn_j and a complete satisfaction threshold tc_j (S1 Appendix). Then, the values of the municipalities for all criteria falling within the defined range have been linearly scaled between 0% satisfaction and 100% satisfaction (0 and 1 respectively), while scores exceeding the limits have been bound to these minimal and maximal threshold values. A partial satisfaction score s_j per criterion is therefore obtained. Equation 1 synthesises the transformation while Fig 4 illustrates the scaling procedure applied to one of the variables.

$$s_j = \begin{cases} 0 & if\ x_j < tn_j \\ 1 & if\ x_j > tc_j \\ \dfrac{x_j - tn_j}{tc_j - tn_j} & otherwise \end{cases}$$

Equation II. 1 : Criterion scaling with the MACBETH method

The approach results in the attenuation of extreme values since we cut the tails of the distribution of the variables.

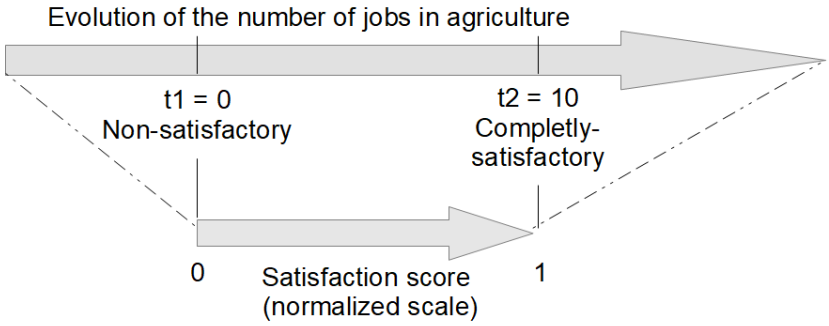


Figure II. 4: Criterion scaling with the MACABETH method for the variable "Evolution of the number of jobs in agriculture". Evolution of the number of jobs in agriculture represents the difference of the number of jobs in agriculture between 2010 and 2012 (in %); the upper arrow indicates the initial values while the lower arrow represents the values that have been normalised between 0 and 1 (s_j). The satisfaction thresholds for this criterion is defined as being 0 and 10 %. The values of the satisfaction thresholds are given in S1 Appendix.

In a successive step, a weighted sum of the scaled scores s_j using a weight w_j associated with each of the criteria has been performed, providing a global satisfaction percentage S for each municipality, ranging from 0% to 100% and interpreted as a sustainability index (Eq. 2).

$$S = \sum_{j=1}^{m} w_j s_j \quad where \quad \sum_{j=1}^{m} w_j = 1 \quad (2)$$

As shown in Figs 1 and 2, the MACBETH method is used at two levels to compute: 1) the sub-indices (fifth column in Fig 2: criteria aggregation (sub-indices); e.g. Pedig-index) containing themselves several criteria (i.e. inbreeding coefficient and effective population size for the Pedig-index) and 2) the global-index (sixth column: ranking based on the global index). With the purpose to decide where to place satisfaction thresholds and weights and to validate these choices, a spatial and statistical exploratory analysis of the different criteria has been carried out as a preliminary step (S2 Appendix). Subsequently and on the basis of the preliminary analysis, satisfaction thresholds and weights were defined for each variable based on the opinion of the 12 experts involved (the chosen values of the weights and thresholds are given in S1 Appendix).

A special note is required as regards the Breed Agriculture Sustainability index. Indeed, it contains statistical components produced at the municipality level to calculate the Local Agriculture Sustainability Index, LAS, and information produced at the breed level. Once the LAS is computed over all municipalities, a mean over the regions where investigated animals are reared is calculated (weighted by the number of animals per region). This value is then aggregated (according to the MACBETH method) with the criteria at the breed level (evolution of the number of farms, and the

cultural value). This aggregated parameter is then named “Breed Agriculture Sustainability” (BAS) Index. This score gives an insight into the sustainability of breeding conditions for a specific breed.

2.4 Web-portal implementation

In this section we describe the technology used to develop the GENMON application while we will argue the reasons supporting these choices in the discussion section. All technologies presented here are open-source.

The main part of the interface is built in Hypertext Markup Language (HTML) and Hypertext Preprocessor (PHP) language. The upload of a file (containing either animal information or socio-economic variables) is made through an HTML-form. The file is then stored in a database management system (DBMS). Here we use PostgreSQL (The PostgreSQL global development group n.d.) with its spatial extension PostGIS (The PostGIS Team n.d.). Once stored in the database, the pedigree analysis is completed by PopRep (Groeneveld et al. 2009). PopRep is a software coded in Perl, which performs a pedigree analysis, stores the results in the PostgreSQL database and produces three output reports about population structure, inbreeding and population size. Then, Structured Query Language (SQL)-queries are executed to aggregate values: averages and sums over the last generation interval for the whole breed as well as per ZIP-code. The indices (Pedig-, Introg-, Geog-, Cryo-, LAS-, BAS- and the Global-index) are computed. The weights and thresholds used for this computation must be provided by the user via the interface before the upload and are stored in the database. The whole procedure is described in the tutorial section of the interface.

For the visualisation part, the interface is built on OpenLayers (Openlayers n.d.), a javascript library. The layer of the map is stored in PostGIS and published to the web with the use of Geoserver (GeoServer n.d.). The map is displayed as a Web Mapping Service (WMS) image and exhibits the attributes of the polygons as a choropleth map (for example the variable “mean inbreeding” of the municipality). To obtain more information on a given selected municipality (additional attributes), a WFS (Web Feature Service) request is run. This enables the display of statistics for a region, as for example the total number of animals in the selected zone or the mean inbreeding.

3 Results

In this section, we give an overview of the outputs produced by GENMON, illustrated by means of three Swiss breeds: the Valais Blacknose sheep (VBN), the Franches-Montagnes horse (FM) and the Swiss Original Braunvieh cattle (OBV).

3.1 Summary table

The main output of GENMON is a summary table enabling the comparison between breeds according to the different criteria mentioned above, which are then aggregated in a global index (see Fig 5). Each line displays the summary information of one breed, whereas each column represents each criterion individually and the global index. To facilitate a comprehension at first glance of endangerment status of all breeds under study, breeds are ranked according their global scores (most threatened on top). Furthermore, a color code for each criterion gives the level of satisfaction according to the thresholds defined by the user (red: not satisfactory, green: satisfactory).

Breed Name	# Animals Last GI	Mean F Last GI	Ne Range	Pedigree completeness	Trend males Last 5 yr	Trend females Last 5 yr	Pedig-Index	Introg-Index	Geog-Index	BAS-Index	Cryo-score	Global index
VBN	34,291	0.101	50-70	95.9	+4.3	-0.3	0.45	0	12.9	0.56	0	0.43
FM	26,877	0.067	50-70	99.7	-2.4	-2.1	0.41	0.12	57.6	0.64	0.5	0.51
SIM	18,301	0.031	50-70	85.6	+54	+5	0.5	0.005	86.8	0.54	0.5	0.66
OBV	50,632	0.036	70-500	99.5	+2.5	+4.6	0.66	0	58.3	0.59	1	0.78

Figure II. 5: Summary output table of the GENMON application. The global index and its different components for three Swiss breeds and a simulated breed. The colors give the degree of satisfaction of each criterion (red: not satisfactory, dark green: totally satisfactory). The breeds are ordered from the most threatened on top to the healthiest at the bottom. (GI: generation interval, F: inbreeding coefficient, Ne: population size. The description of the indices is given in section 2; VBN: Valais Blacknose sheep, FM Franches-Montagnes horse, SIM simulated breed,OBV Swiss Original Braunvieh cattle). The colors are assigned according to the following thresholds (expressed in satisfaction score): 10% is the limit between red and yellow; 50% defines the limit between yellow and light-green; 95% corresponds to the threshold between light and dark green

Fig 5 shows that according to the weights and thresholds that were applied, the VBN is the most endangered breed with a relatively low global index (0.36), while the OBV seems to be a healthy breed (global index = 0.69). Beside these observations, the specific problems of each breed can also be quickly identified from the sub-indices. For example the VBN suffers from a high mean inbreeding coefficient (mean F = 0.101) and is spatially very concentrated (Geog-index = 12.9). On the other hand, the FM is significantly introgressed (Introg-Index = 0.12), which lowers its global index. The SIM breed has an incomplete pedigree which lowers its pedig-index and finally its global score. It has to be noted that weights and thresholds of the pedig-index are set differently for each species (see S1 Appendix). We propose here across-species early warning but the user can easily separate species and rank the breeds accordingly.

3.2 Detailed investigations

Once the main problems are identified, more detailed information can be obtained on the breeds of interest. Here we show the details for the breed with the lowest global index, namely the VBN; it is of interest to visualise the geographical distribution of inbreeding coefficients (Fig 6) and its evolution over the last few years (Fig 7).

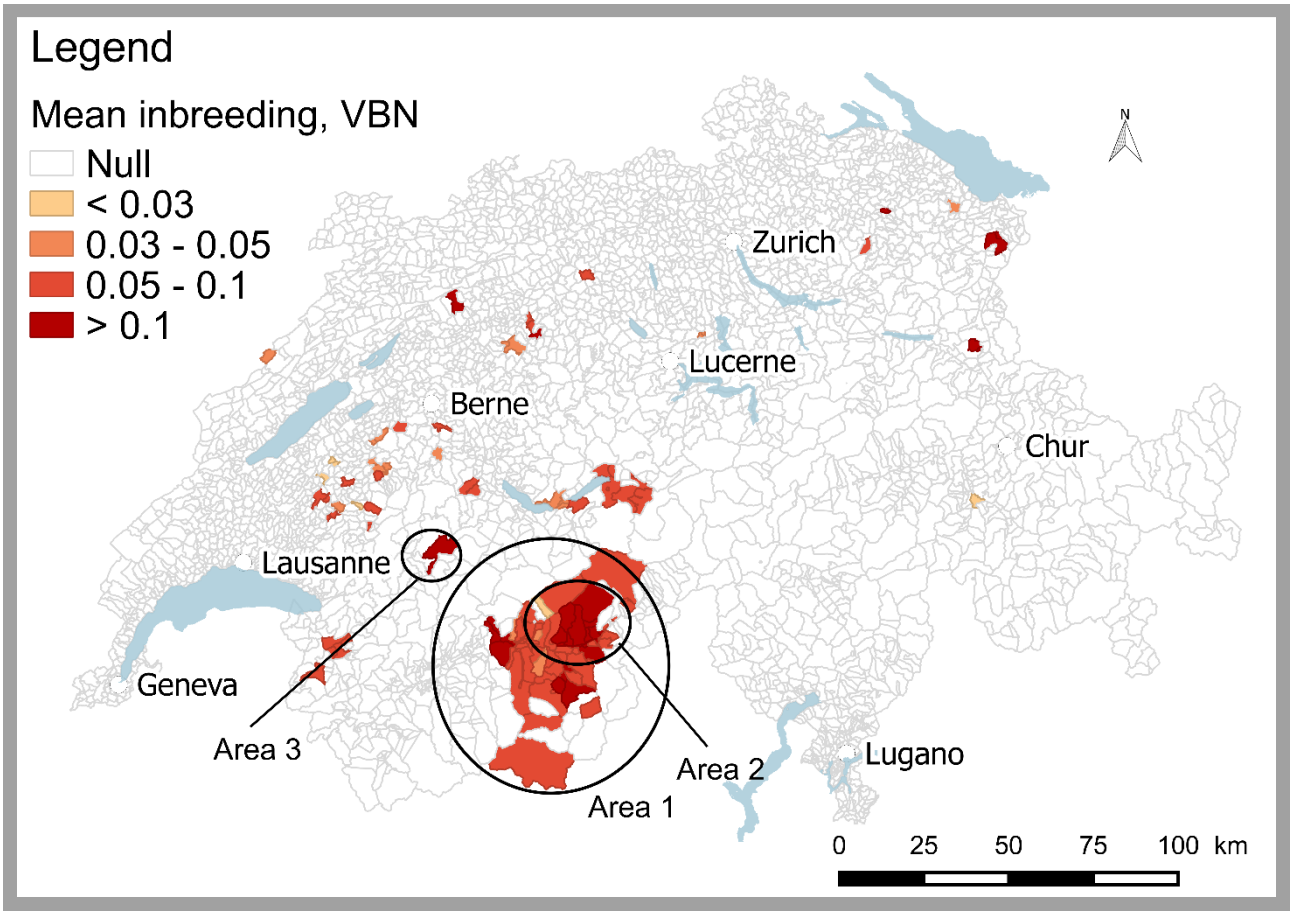


Figure II. 6: The geographical distribution of inbreeding coefficients per ZIP-code for the Valais Blacknose (VBN) sheep. The mean inbreeding is computed over the last generation interval (2010-2012). Null areas correspond to regions where no VBN sheep is reared.

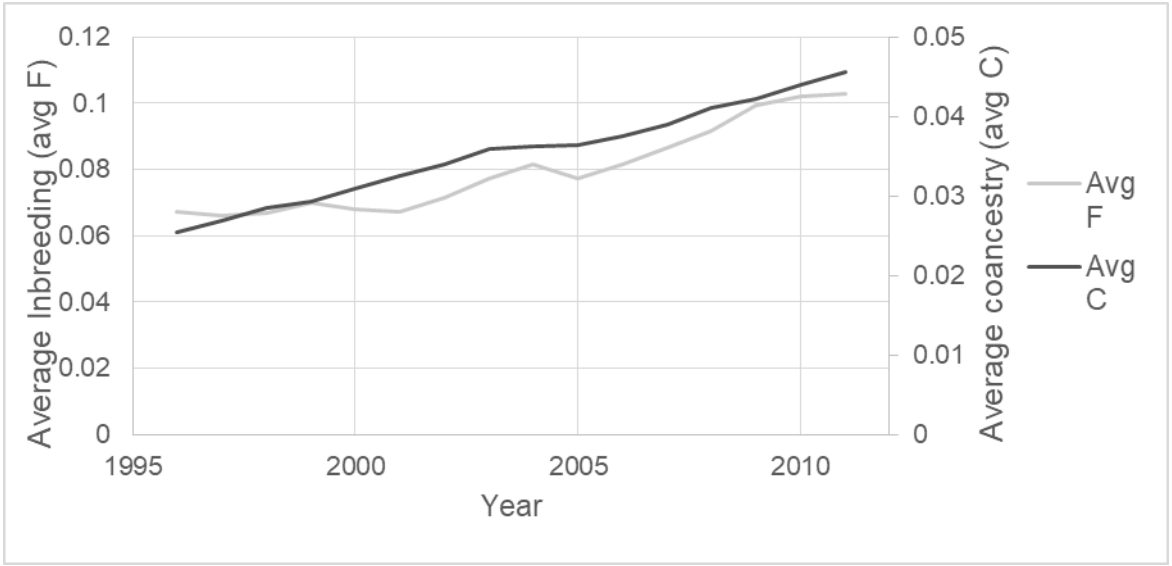


Figure II. 7: Inbreeding and coancestry coefficient for the Valais Blacknose (VBN) sheep breed between 1994 and 2012. Progression of more than 0.04 of the inbreeding and 0.02 of the coancestry in 15 years; the current average inbreeding is remarkable and exceeds 0.1 while the coancestry exceeds 0.04.

The VBN is very concentrated in a small portion of the Swiss territory (in fact 98% of animals are located within area 1 of Fig 6). The mean inbreeding coefficient is higher than 0.05 in almost all regions and some ZIP-codes area exceed 0.1. More precisely, 61% of the municipalities with VBN show a mean inbreeding higher than 0.03 (value that has been chosen for the satisfaction threshold) and 4% have a mean inbreeding higher than 0.1 (the value corresponding to the non-satisfaction threshold). Geographical distribution maps are essential to identify specific regions with high inbreeding coefficients that would require a more intense assistance in breed management (typically area 2 of Fig 6). Furthermore, to better identify regions which are critical for certain breeds and in order to assess the importance of a region for the breed, distribution maps of inbreeding coefficients can be compared with maps showing the number of animals per municipality, (see S1 Fig). This also allows the user to identify regions with only few animals, in which the inbreeding coefficient is likely to be overestimated (e.g. area 3 of Fig 6). With GENMON, other maps can be calculated as for example the geographical distribution of the introgression rate, shown here for the FM breed (S5 Fig).

3.3 Local Agriculture Sustainability Index

The spatial distribution of the local sustainability index, allowing a ranking of all Swiss municipalities (including those having no animals uploaded in the GENMON application) is shown in Fig 8. This cartographic representation of the sustainability index integrates social, demographic, economic and environmental characteristics of the different regions of Switzerland at the municipality level.

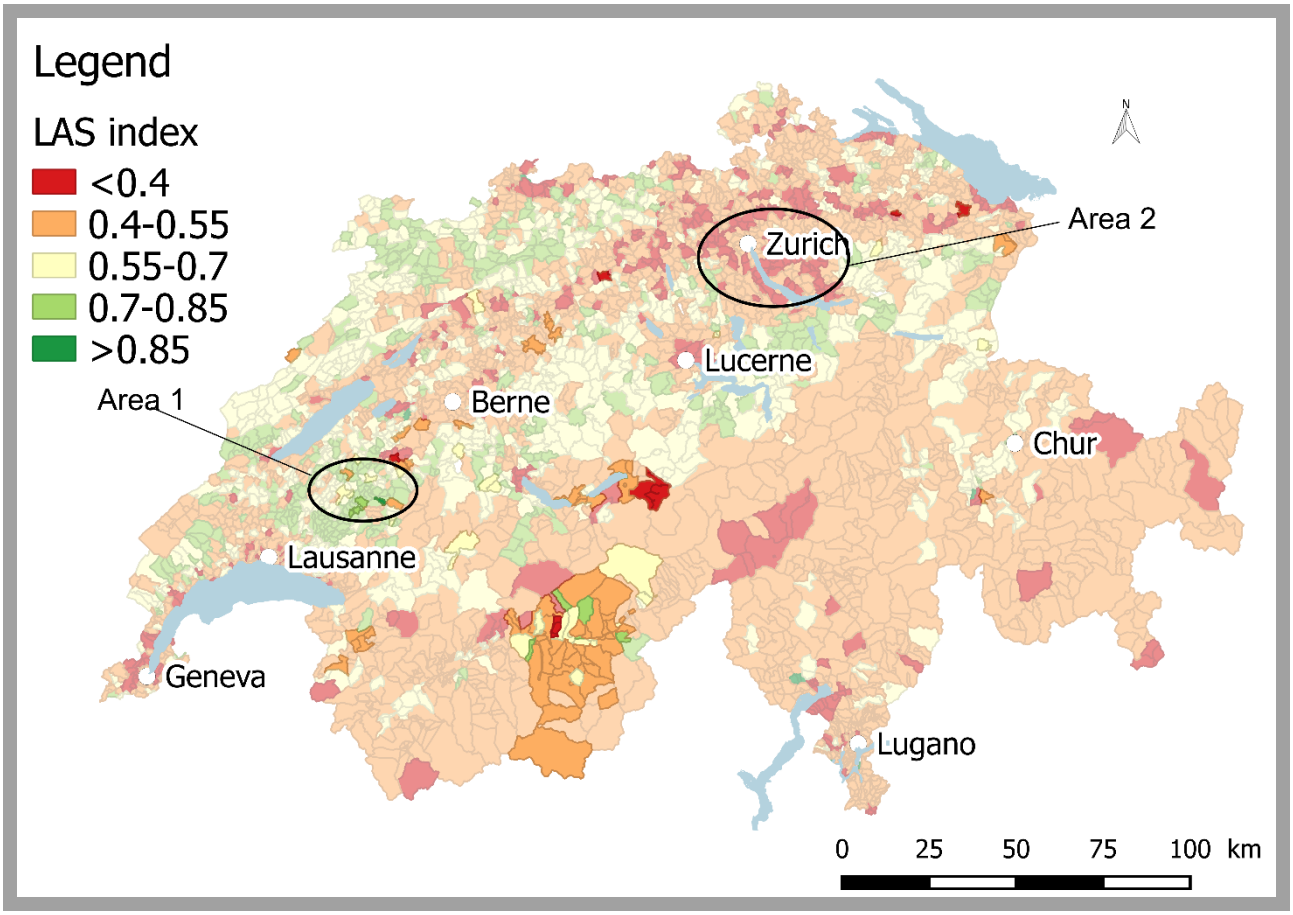


Figure II. 8: Geographical distribution of the local agriculture sustainability (LAS) index. The computation based on weights and thresholds described in S1 Appendix. The colors of the areas

that do not contain any Valais Blacknose sheep are faded. Sustainable areas are shown in green (e.g. area 1) while low sustainability is represented in red (e.g. area 2 situated in an urban low-land zone). The pale yellow shows intermediate values.

The Swiss territory shows different regional sustainability trends. Low sustainability values are found mainly in the Plateau region located between Lausanne, Berne and Zurich (Fig 8).

The best conditions for breeding activities are found primarily in lowland/mid-altitude areas located in the region between the Plateau and Jura (Fig 8). These municipalities are highlighted in green on the map (Fig 8). Actually, the States of Vaud and Fribourg (area 1 in Fig 8) as well as the central Switzerland have many regions with relatively high sustainability scores, potentially favorable for farm animal breeding activities. The Jura Mountains as well show high sustainability values, mainly in the States of Neuchâtel and of Jura. In order to identify if a specific breed is reared in sustainable conditions, the "Breed Agriculture Sustainability" index (BAS-index, Fig 1) can be used, which computes the average over the regions where animals are reared and weighted by the number of animals per region. In this case, the majority of municipalities with VBN are classified in the categories where the farming activity have low sustainability scores (only 48% of the ZIP-code areas containing VBN have a LAS index larger than 0.55).

4 Discussion

4.1 Performance of the GENMON application

GENMON offers a ranking evaluation of the level of endangerments based on a straightforward application. This service has not been implemented anywhere else and might serve as a good basis in order to initiate, support and supervise prioritisation and conservation programs as required by the FAO protocol (FAO 2007b). Being straightforward and easy-to-use, the application can be applied to a large number of breeds. Consequently, the Federal Office for Agriculture in Switzerland (BLW) intend to use GENMON to monitor local breeds in the future.

The use of georeferenced animal data as proposed in GENMON is also unique in FAnGR. Indeed, the English system of livestock endangerment scale only uses geography to assess geographic concentration of breeds. The integration of different data type such as socio-economic and environmental factors is made possible using georeferenced data that is often ignored in livestock conservation. The idea of assessing sustainability of breeding condition has already been proposed (Bertaglia et al. 2007), but has not been implemented in an automated pipeline until now.

Both in the English system and in GENMON, the estimation of the geographic concentration is a rough approximation. It is mainly used to assess the ability of diseases to spread between flocks, but does not take the barriers or paths of the environment into account. A better approach could be inspired by the assessment of the geographic range and the area of occupancy in wildlife conservation (see for example Gaston, 1991).

GENMON also offers flexibility, since the weights and thresholds can be adjusted, which enables an adequate modelling of the situation depending on the species and the country. The selection of variables, weights and threshold parameters used to build the LAS and BAS sustainability indices is part of a participatory process involving experts, in order to select the proper social, demographic and economic factors affecting the development of specific livestock breeds. Care

must be taken when deciding the weights and thresholds. Indeed these parameters will have a considerable influence on the final output. Here we undoubtedly face a heuristic problem with unknown properties, so that we have no way of assessing how good the final score is. As a result, the panel of experts must be representative and diverse enough to represent different backgrounds, breeding associations and professional activities related to the livestock sector. Indeed its role is to select a robust set of parameters translating the policy the government agency wants to apply. A key role will be played by specialists of the surveyed breeds to fine-tune these parameters. In the case we present here, the selection of parameters and the tuning of weights was carried out by a panel of experts working for the Swiss Federal Office for Agriculture (FOAG), and involved in the sector of animal production responsible for the monitoring of livestock breeds.

When confronted to GENMON, breeders reacted positively and actively gave their opinions to improve the application. They expressed their satisfaction in being able to step back and discover in more detail the context of breed monitoring that considers various criteria, to evaluate the status of their breed when compared to other and to diagnose which components of their breed could be improved.

GENMON has been designed for the specific case of Switzerland and relies on the data availability of this country. Nonetheless, the methodology could be used in other places, with inevitable modifications to adapt to the accessible data. For example, it is not mandatory to use ZIP codes, but one could use other administrative divisions. In addition, DNA instead of pedigree could be used to assess the level of inbreeding. The weights and thresholds should be discussed in each country individually, depending on the specific environmental conditions, breeding practices, policy implemented and data available.

Finally, GENMON quickly assesses the conservation status of breeds and to identify and prioritise vulnerable ones. Moreover, a rapid identification of factors affecting the conservation of breeds is possible through the detailed results (e.g. for the VBN Figs 6 and 7). This might serve as a good basis in order to initiate, support and supervise prioritisation and conservation programs.

4.2 Technology chosen

An important technological challenge was to use open source software only to develop GENMON. Indeed, open source technology have increased transparency due to the availability of code, offers a greater flexibility, given that source code can be modified according to the need of the application. (Ertz, Rey, and Joost 2014). Open source technology will also favor the implementation of this solution in other countries.

It has been chosen to build GENMON as a Web-service rather than a desktop application, so that a unique and central database will collect and store information coming from different sources. Moreover, given that the computations carried out with GENMON are intensive (especially the pedigree analysis), it is of interest to perform the computation on the server-side, which frees the user's computer for other tasks.

With regards to the software used in GENMON, the DBMS PostgreSQL was chosen since it is one of the most efficient open source DBMS (notably due to its capacity to handle georeferenced data efficiently, Steiniger and Hunter 2012) and because it easily communicates with an interface built in PHP. Moreover, the PopRep code used for the pedigree analysis uses a PostgreSQL database, which facilitates the data transfer. For the pedigree analysis, PopRep has been favored over other pedigree software like CFC (Sargolzaei et al. 2006) or ENDOG (Gutiérrez and Goyache 2005)

because it has already been successfully used by FOAG, and people from this institution are familiar with its outputs. Furthermore, PopRep also has the advantage of directly creating detailed reports, which can be useful for further analyses. Openlayers (Openlayers n.d.) has been selected for the cartographic environment, for it offers a large flexibility and is well documented.

5 Conclusions

GENMON was developed to satisfy requirements of the "Global Plan of Action for Farm Animal Genetic Resources" launched by FAO, which still needs to be set up in many countries. This application has been developed as a useful tool for FAnGR monitoring that could assist many countries in this task, provided they have sufficient data (including pedigree and socio-economic variables).

It is an easy-to-use WebGIS application relying on open source software solutions and provides a multi-criteria approach for monitoring endangered breeds based on subjective thresholds of a government agency. By means of geographic coordinates, the application integrates different types of criteria, including pedigree data, genetic introgression, socio-economic and environmental aspects and geographical concentration of the breeds under study to evaluate the local context in which they are reared. GENMON computes a global sustainability index for each breed, making it possible to compare the endangerment level of several species and/or breeds, while enabling the identification of the most important problems and their geographical location for breeds separately, on the basis of sub-indices. Based on these outputs, a detailed examination of the conservation status of breeds can be carried out, which might serve as a firm basis for proposing prioritisation policies. An important contribution of GENMON is to provide decision-makers with a clear identification of breeds, municipalities and corresponding breeders that should be supported with special policies in order to maintain a lively and sustainable breeding sector.

The GENMON application will be expanded with new features. A relevant example is the potential future use of genetic data following the methodology described by vanRaden *et al*. (2008), so as to complete pedigree information if it is not complete or to replace it to avoid the time-consuming pedigree analysis step and to assess inbreeding and effective population size in particular. However, current system developments will soon make it possible to process conservation indices based on marker-based genetic information as well.

The GENMON application is functional and can be accessed using the following link: lasigsrv2.epfl.ch/genmon-ch. A sample file is available for users interested in testing the upload of a file. The code is available on GitHub: https://github.com/SolangeD/GENMON.

6 Supporting information

S1 Appendix. Description of the workshop procedure to obtain thresholds and weights.

S2 Appendix. Descriptive statistics of the selected variables used in the local agriculture sustainability index.

S1 Fig. The geographical distribution of the number of individuals per ZIP-code for the Valais Blacknose (VBN) sheep.

S2 Fig. The geographical distribution of mean inbreeding coefficients per ZIP-code for the Original Braunvieh (OBV) cattle.

S3 Fig. The geographical distribution of the number of individuals per ZIP-code for the Original Braunvieh (OBV) cattle

S4 Fig. The geographical distribution of mean inbreeding coefficients per ZIP-code for the Franches-Montagnes (FM) horse.

S5 Fig. The geographical distribution of introgression per ZIP-code for the Franches-Montagnes (FM) horse.

S6 Fig. The geographical distribution of the number of individuals per ZIP-code for the Franches-Montagnes (FM) horse.

7 Author contributions

Conceptualisation: SJ SD CF FJ GM. Data curation: SD. Formal analysis: SD. Funding acquisition: SJ CF. Investigation: SD SJ CF GM. Methodology: SJ FJ SD CF GM. Project administration: SJ. Resources: CF. Software: SD. Supervision: SJ. Validation: SJ. Visualisation: SD. Writing – original draft: SD SJ FJ. Writing – review & editing: SD SJ CF IW FJ GM

8 Funding

GENMON was funded by the Swiss Federal Office for Agriculture (FOAG http://www.blw.admin.ch/), Division of Domestic Animals and Breeding.

9 Acknowledgments

We thank Gabriela Obexer-Ruff and Claude Gaillard of the Institute of Genetics, Vetsuisse faculty at University of Bern for their help and valuable suggestions. We thank Beat Bapst, Madeleine Berweger and Juerg Moll of the Swiss Brown Cattle Breeders' Federation for their assistance and for making available the list of breeders' addresses. We would also thank all breeding associations who kindly provided the data for the illustrative examples and who agreed to let us use them in this paper. We also thank all people having participated to the workshop.

TAKE HOME MESSAGE

- The monitoring of endangered breeds is essential to prevent further loss of genetic diversity.
- Our approach relies on a multi-criteria analysis accounting for biological, geographical and socio-economical indices.
- Geography plays a central role in the integration of various data at different levels.
- The proposed solution is in the form of a Web GIS platform.
- To help breeders with no *biogeoinformatic* background, emphasis is given on a straightforward use of the platform.
- The outcome is a ranking of the endangerment level of breeds which enables a rapid identification of problems.
- The case studies on three local Swiss breeds show that the endangerment level varies across species.

Chapter 3
Preserving locally adapted genetic variations

The previous chapter focused on the monitoring of the erosion of genetic diversity, so that actions can hopefully be taken in order to assist breeds at risk. However, given the current situation, erosion of genetic diversity seems unavoidable and in consequence, breeding programs should pay particular attention to the preservation of adapted traits (Bishop, 2012), for which molecular biology can offer interesting insights (Biscarini et al., 2015; Boettcher et al., 2015; Gibson & Bishop, 2005). In developed countries, breeding programs for example are now assisted by molecular tools, leading to genomic selection (Schefers & Weigel, 2012). Nevertheless this type of selection does not typically account for the environment in which the animal lives, thus ignoring the potential genome x environment interactions (Montaldo, 2001). Consequently, this selective breeding technique should be adapted to include candidate loci of local adaptation (Mwai et al., 2015), which, to our knowledge, has not yet been practically implemented. Possible genetic variations to include in such breeding programs are potential adaptations to disease resistance and heat tolerance (Hanotte et al., 2003; Kim & Rothschild, 2014), two traits that are of particular importance in the context of climate change. A more drastic way to preserve locally adapted traits is the use of gene-editing, successfully applied in cattle to confer heat tolerance found in Senepol cattle to Holstein animals (Dikmen et al., 2014).

The signature of local adaptation can be detected using landscape genomics (Manel et al., 2010), which studies the interaction between a genotype and the environmental conditions of the living organism at study. Landscape genomics is a perfect illustration of the use of *biogeoinformatics*, and has already been applied to livestock. A few years ago, the *Samβada* software (Stucki et al., 2017; provided in Appendix B) was developed to assess through logistic regressions, the relationship between genotype and environmental conditions, with special emphasis on high performance computing. However, this approach has been underexploited, one reason being that its use requires competences in too many fields.

As a result, we propose an integrated pipeline to perform landscape genomic studies, from the creation of an environmental file based on the geographic location of samples, up to the plotting of graphs to analyse the results. For livestock species, we take advantage of the fact that their annotated genome is usually available, establishing a link between a given mutation and a biological function. Two case studies on Moroccan sheep and Spanish cattle illustrate the use of the pipeline in FAnGR conservation. The data are drawn from two research projects: NEXTGEN, which aimed at using advanced molecular methods to preserve FAnGR and ClimGen, which was designed to take advantage of genomic tools to preserve farm animals resilient to climate change. Both projects are within the scope of *biogeoinformatics* in livestock conservation.

As the first author of this article, I completed most of the tasks, both in the implementation of the R package and in the writing of the manuscript. The other authors provided assistance and advises regarding the methods to be used and the writing of the paper, the complete list of contribution being available at the end of the chapter.

Duruz, S., Sevane, N., Selmoni, O., Vajana, E., Leempoel, K., Stucki, S., Orozco-terWengel, P., Dunner, S., The NEXTGEN Consortium, The CLIMGEN Consortium, Bruford, M.W., Joost, S. (2019). Rapid identification and interpretation of gene–environment associations using the new R.SamBada landscape genomics pipeline. *Molecular ecology resources,* 19(5), 1355-1365.
doi: 10.1111/1755-0998.13044

1 Introduction

Local adaptation implies the existence of advantageous alleles conferring a population living in its native habitat a higher fitness than any other allochthonous population living in the same habitat (Kawecki & Ebert, 2004). Landscape genomics methods, including genome-environment association (GEA) are among the approaches used to detect signatures of local adaptation and have become increasingly popular, mainly due to the decreasing cost of sequencing, but also because of the recent availability of fine-scale environmental datasets (Balkenhol et al., 2017; Rellstab et al., 2015). However, the massive amount of data that can be analysed due to these improvements have made the development of more efficient tools essential (Stucki et al., 2017).

To this end, *Samβada* was developed to perform large amounts of logistic regressions between genetic markers and multiple environmental variables (Stucki et al., 2017). *Samβada* computes uni- or multi-variate models between a binary genetic variable (e.g. the presence/absence of a genotype) and one or more environmental variables. Significance is assessed against a null model (i.e. constant model in the case of univariate or a parent model in the multivariate case). Population structure can be accounted for by treating one or several population variables as environmental variables in multivariate analyses. *Samβada* is written in C++ with a particular emphasis on high performance computing (HPC). Since its publication, *Samβada,* as applied alone or in combination with other methods, proved useful to target putative genomic regions underlying local adaptation in a wide variety of species, including domestic animals such as swine and cattle (Cesconeto et al., 2017; Vajana et al., 2018), wild animals such as the fresh water sculpin and European pond turtle (Lucek et al., 2018; Pereira et al., 2018), as well as many different plant species including the European beech and the cow-tail fir (Cuervo-Alarcon et al., 2018; Shih et al., 2018).

Despite its many advantages, *Samβada's* command-line format is sometimes laborious and the amount of pre- and post-processing represents an obstacle to its widespread use. Indeed, a typical processing chain, such as the one proposed by Stucki et al. (2017) includes (i) the use of a GIS software to retrieve environmental information at sampling locations; (ii) molecular data filtering by standard software such as *PLINK* (C. C. Chang et al., 2015); and (iii) the inclusion, whenever present, of population structure usually computed with a dedicated software such as *ADMIXTURE* (Alexander et al., 2009). Similarly, post-processing of results involves (i) the computation of p- or q-values (Storey, 2003) for the association tests involving each genotype; (ii) the production of maps and plots (typically Manhattan plots) in which the location in the genome (i.e. the position in base pair) of a point representing the result of a model is difficult to establish since the plot is rarely interactive; (iii) the formulation of queries to the Ensembl genome browser (Hubbard et al., 2002) to search for candidate genes adjoining the single-nucleotide polymorphisms (SNPs) highlighted.

However, the *R* software (R Core Team, 2018) provides an open source computing environment adapted to different fields in Biology, in which many of the above-mentioned pre- and post-processing tasks can be found in various *R*-packages. Further, *R* can be coupled with compiled

languages (such as C++) so as to be more efficient when processing large datasets (see e.g. the case of the software LFMM 2; Caye et al., 2019, p. 2).

In this context, we developed *R.SamBada*, an *R*-package designed to facilitate and enhance the whole data process described above by integrating multiple existing packages and building new functions into one easy-to-use pipeline. We present the use of the package by illustrating its benefits with two case studies for which driven signatures of selection were investigated as part of the ClimGen project (https://climgen.bios.cf.ac.uk/). The first dataset consists of 160 Moroccan sheep genotyped with whole genome sequencing (WGS) and characterised by no clear population structure, while the second one encompasses a Spanish Lidia Cattle population of 349 samples genotyped with 50K SNP chip, with one population variable. Both datasets are already published (see Data availability section) but have not yet been analysed with *Samβada*.

2 Materials and Methods

We first present *R.SamBada*, with an overview of its functions, and then describe its application to two case studies from the ClimGen (https://climgen.bios.cf.ac.uk/) project, detailing how the genetic data were collected and prepared for subsequent analyses. Both studies investigate climate-mediated selection at the genome level: the first analysis is carried out on a Moroccan sheep dataset using whole genome sequences, and the second one involves a Spanish cattle breed (Lidia) genotyped with the Illumina BovineSNP50 array.

2.1 Implementation

R.SamBada provides functions for (i) preparing the genetic (i.e. SNPs) and environmental information to be processed (Pre-processing), (ii) running *Samβada* directly into the *R* environment (Processing), and (iii) performing *post-hoc* analyses on the basis of *Samβada*'s output (Post-processing). The following sections detail these different steps (see Fig. 1).

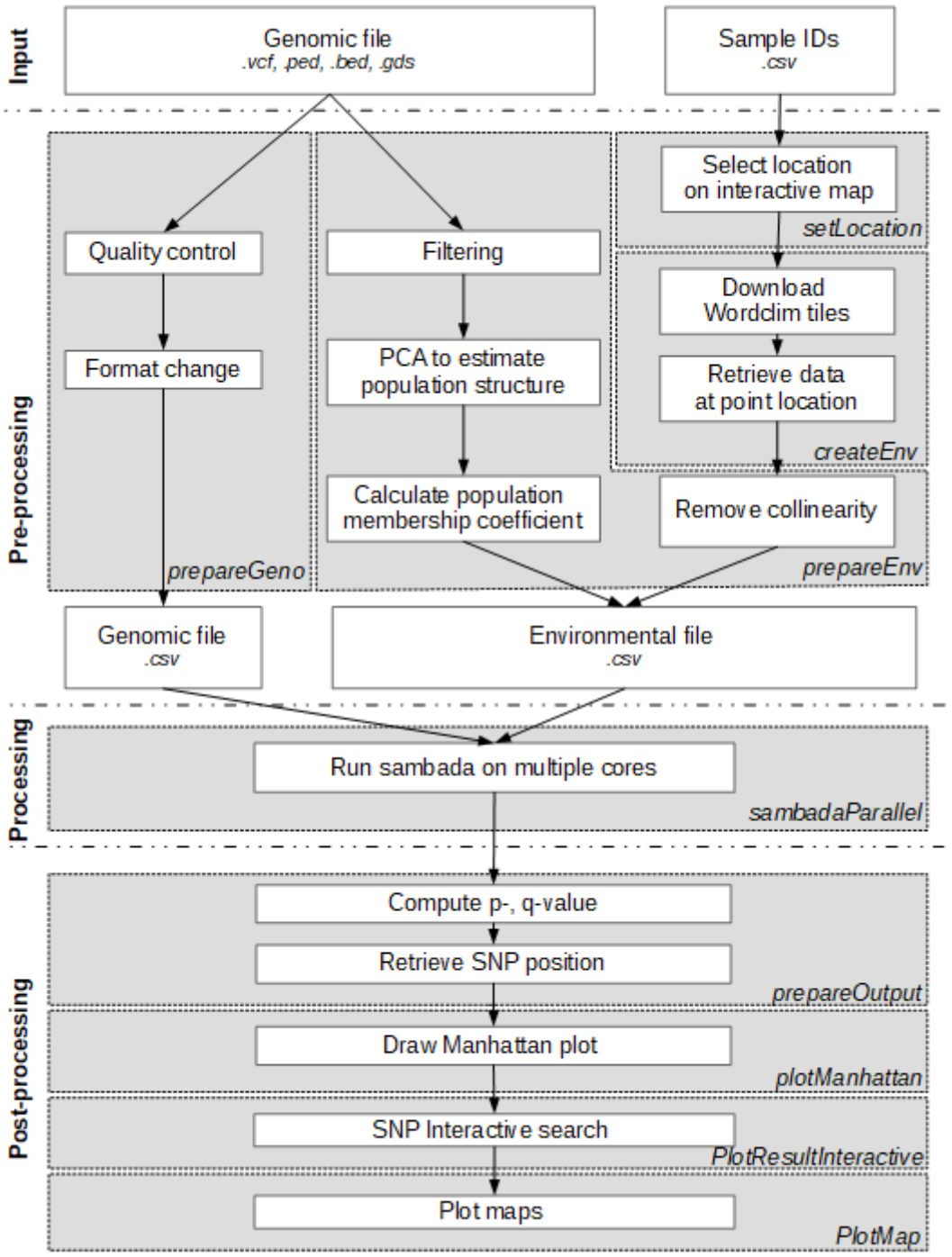


Figure III. 1: Overall functionalities and process in *R.SamBada*. Grey boxes with italic names indicate functions included in the package. The process starts with a genomic file and a file with sample locations or list of IDs. The preprocessing will format the genomic file and prepare the environmental file; *Samβada* is then run parallely on multiple cores; After computing of p-, q-values, Manhattan plots and maps can be drawn and Ensembl database can be queried.

2.1.1 Pre-processing

Three functions have been implemented to perform the main operations required before running *Samβada*. Firstly, *prepareGeno* is used to prepare the genomic file, by treating a SNP input dataset from various formats (.vcf, .gds, .ped or .bed) and generating a filtered file complying with

Samβada's input standards. *prepareGeno* relies on the *SNPRelate* package (Zheng et al., 2012) to perform standard quality control (QC) for minor allele frequency (MAF), linkage disequilibrium (LD) and missingness. In order to assist users in selecting adequate pruning levels, *prepareGeno* displays the frequency distributions of MAF, LD and missingness along with the proportion of SNPs discarded corresponding to the thresholds applied; in this way, QC can be tailored to avoid reducing the dataset too much while controlling for missing information.

Secondly, if coordinates are not available, *setLocation* can be used to open a local web page that assist users in defining sample locations using mouse-clicks on an interactive map. The projection system used is WGS84 (corresponding EPSG - European Petroleum Survey Group – code: 4326), a worldwide system with coordinates in degrees (longitude/latitude) (more information on projections in Leempoel et al., 2017).

Then, *createEnv* provides the user with a pipeline to produce an environmental dataset out of the file containing sample locations. If raster files representing environmental variables are available, then habitat information is directly derived at the sampling locations. However, if these files are not present, *createEnv* is able to use the samples' geographic coordinates to identify the correct tiles in the WorldClim (Hijmans et al., 2004) and SRTM (Shuttle Radar Topography Mission; Farr et al., 2007) databases and to download adequate climatic and altitudinal information. The WorldClim database contains monthly minimum, maximum and average temperatures and total precipitations together with a series of bioclimatic variables computed from these variables (e.g. precipitation of wettest quarter of the year, complete list available http://www.worldclim.org/bioclim), while SRTM only provides altitude. Coordinates can be given in any projection system (as long as the EPSG code of the projection is given as an input parameter of the function). A comma-separated value (.csv) file is then returned containing the sample IDs, their locations and the values of the corresponding environmental variables. The interactive mode shows maps of sample locations, so as to locate potentially misplaced points or erroneously-set projection systems. This function can save substantial effort, since one single command substitutes a long processing chain that typically includes the download of voluminous data for the entire globe, the import of both sample locations and raster environmental data into GIS software and the retrieval of environmental values at point locations.

Finally, the *prepareEnv* function produces a file containing the design matrix that *Samβada* will process. At first, highly-correlated environmental variables are removed according to a correlation-coefficient threshold defined by the user in order to keep only independent eco-climatic factors in the analysis. The interactive mode will show the graph of the number of variables discarded as a function of the chosen correlation threshold. Then the genetic structure of populations is assessed by means of a Principal Component Analysis (PCA) as implemented in *SNPRelate*. The user is provided with the possibility of further processing PCA output by a clustering algorithm, which calculates individual membership coefficients as a function of the distance from the clusters centroids (Lee et al., 2009). Changes in the clustering solution according to the chosen k-number of clusters can be interactively visualised. After ordering individuals according to their identifiers (as in the genomic file and necessary for *Samβada*'s analysis), a final .csv file is generated, containing the samples' IDs, retained environmental variables and either PCA score(s) or membership coefficient(s) representing population structure.

2.1.2 Processing

Samβada includes a useful module called *Supervision* that is designed to split the input file into several sub-files and merge the split result files, thus reducing drastically the computation time by allowing manual start of parallel sessions. This module has however rarely been employed to date, possibly due to its laborious and time-demanding preparation procedure. This limitation is overcome in *R.SamBada,* through the *sambadaParallel* function that implements Supervision by default, and relies on the doParallel *R*-package (Microsoft Corporation & Weston, 2017). Furthermore, unlike the previous version of *Samβada* (0.5.1 used in Stucki et al., 2017), version 0.8.1 (included in *R.SamBada*) makes it possible to directly assess the effect of population structure by comparing the full model (containing all population variables and one or more environmental variables) with the null model (containing only population variables).

2.1.3 Post-processing

Four *ad hoc* functions have been developed for obtaining and visualising *Samβada*'s outputs. In the post-processing pipeline the statistical significance of genotype-environment associations are derived since only G- and Wald-scores are calculated within *Samβada*, and no hypothesis testing is performed. Here, *R.SamBada* provides the function *prepareOutput*, which computes i) *p*-values by comparing the spread of G- or Wald scores from *Samβada* to a χ^2 distribution, and ii) *q*-values based on Storey's method (Storey, 2003). The visualisation of the position of outlier loci along the genome is possible using the *plotManhattan* function that generates Manhattan plots based on the *p*- or *q*-values as computed by *prepareOutput*.

Next, *plotResultInteractive* can be used to display interactive Manhattan plots. In particular, users can specify which chromosome(s) they want to visualise for which environmental variable, the *p*- or *q*-values being then plotted for each genotype as a function of their genomic coordinates. Marker name, position, *p*-value, functional relevance (e.g. intergenic-, non-synonymous-variants) as well as proximal genes – whenever present – can be then retrieved for each marker by directly clicking on the set of points of interest being displayed. Gene annotation and functional investigation are performed by internal calls to the Ensembl genome browser (Hubbard et al., 2002) and the Variant Effect Predictor (VEP) (Yates et al., 2015), respectively, while the whole interactive graphical interface relies on the *R*-package *shiny* (W. Chang et al., 2018). Additionally, a basic geographic map shows the geographic distribution of the marker, the environmental variable and the population structure (examples presented in Fig. S1, Supplemental information).

Finally, the *plotMap* mapping function makes it possible to represent the geographic distribution of i) the putative signature(s) of selection, ii) the environmental pressure associated (as a raster background if available), iii) the neutral population structure (see Fig. 5 for an example), and iv) the degree of genetic similarity among sampling sites for the target markers (i.e. its spatial autocorrelation, see Stucki et al. 2017). *plotMap* relies on the functionalities embedded within the *packcircles R*-package (Bedward et al., 2018) to shift nearby sampling points and prevent them from overlapping.

2.2 Case studies

2.2.1 Moroccan sheep

Sampling and genetic data. Moroccan sheep (*Ovis aries*) populations constitute an excellent case study to investigate potential local adaptation through landscape genomics, since they have experienced i) low anthropogenic selective pressure (Guessous et al., 1989), and ii) contrasted climatic conditions throughout the whole country, as imputable to presence of the Atlantic Ocean, the Atlas Mountain, and the Saharan desert in the South. WGS data from sheep (*Ovis aries*) populations in Morocco were produced and made available by the NextGen project (https://nextgen.epfl.ch) and are analysed for climatic selection signatures in the present study. A total of 164 individuals were sampled according to a grid composed of 162 cells of 0.5° of longitude/latitude each, so as to maximise the range of environmental conditions and geographical distribution (see Fig. 3). Detailed sequencing and genotyping information are described in Alberto et al. (2018).

Preprocessing. QC analysis was performed using the *prepareGeno* function with MAF<0.05 and SNP missingness<0.1, leading to a pruned dataset composed by 20'226'452 SNPs (corresponding to 60'679'355 genotypes). SRTM and Worldclim variables (56 in total) were downloaded with *createEnv*, and *prepareEnv* was run to check for variable correlation in order to exclude variables showing an r^2 higher than 90%, resulting in a final dataset consisting of 16 environmental variables (13 Bioclim variables, 2 raw WorldClim and altitude). No population variable was included in *Samβada*'s models (univariate mode) since no evidence of population structure emerged using the PCA method implemented in *SNPRelate* (Fig. S2, with genomic filter of MAF<0.05, SNP missingness<0.1 and LD threshold<0.2).

Post-processing. q-values based on G-scores were visualised with a Manhattan plot using a significance threshold of 0.05. *plotResultInteractive* was used to detect genes neighbouring the markers under selection as well as to identify variant functions (e.g. non-synonymous SNPs).

2.2.2 Spanish Lidia cattle

Sampling and genetic data. The Lidia cattle breed (*Bos taurus*) emerged during the XVIII century and evolved mainly in the *dehesas* ecosystems of the West/South-West Iberian Peninsula, composed of pasturelands interspersed with Mediterranean oaks *(Quercus* ilex) (del Barrio et al., 2014). Since its establishment, Lidia was prompt to isolation by preventing crossbreeding with allochthonous cattle (Eusebi et al., 2017), and became fragmented into reproductively isolated lineages (called encastes) with homogeneous morphology and behavior and genetics (Boletin Oficial del Estado., 2001). Such a peculiar evolutionary and cultural context boosted Lidia's population size to become the largest Spanish breed and made it one of the most inclusive intergrading bovine population, granting high level of genetic richness among encastes coupled with low average genetic diversity values within lineages (Cañón et al., 2008). 349 individuals were sampled among 61 different breeders evenly distributed across Southern Spain's dehesas region (see Fig. 4). Between one and seventeen animals per breeder were selected based on pedigree information to minimise the risk of kinship among individuals. Animals were genotyped using the Illumina BovineSNP50 array v.2 (Eusebi et al., 2017).

Preprocessing. LD decay was first analysed to ensure a sufficient coverage of the genome with the SNP chip used in this study (Fig. S4). QC analysis was performed using the *prepareGeno* function

with a MAF<0.05 and SNP missingness<0.1. The resulting molecular dataset consisted of 38'335 SNPs (i.e. 115'005 genotypes). SRTM and Worldclim variables (56 in total) were downloaded with the *createEnv* function, and *prepareEnv* was used to test for variable correlation resulting in only 15 variables (10 Bioclim and 5 raw WorldClim variables) kept which showed a r^2 lower than 90%. Due to the presence of population structure observed with *SNPRelate*'s PCA method (see Results section and Fig. S3), *Samβada* was run in bivariate mode by adding a variable to account for population structure (score of the first PCA). This variable is not correlated with other kept environmental factors (highest correlation: precipitation in April, r^2=0.25).

Post-processing. p-values based on GScores were corrected for multiple testing with Bonferroni method, and subsequently were displayed in a Manhattan plot (q-values were not conservative enough in that case), with a significance threshold of 0.05, and *plotResultInteractive* was then used to detect associated genes.

3 Results

3.1 Time efficiency

Besides the time saved during pre- and post-processing, *R.SamBada* is more time-efficient than using *Samβada's* command line (v. 0.5.1) for two reasons: firstly, *R.SamBada* automatically integrates *Supervision* to distribute the processing of models over several cores, which makes the analysis run x times faster (where x represents the number of CPU), to which we must add a few minutes to split and merge the dataset (e.g. 24 minutes to split and merge the sheep dataset, compared to 160h saved by parallel computing on the same 11 cores). Secondly, if population variables are included in the analysis, the new version of *Samβada* (0.8.1) will only focus on models including population variables. Here, the time saved will depend on the number of population variables (for the Lidia cattle analysis, with one population variable, it reduced the computing time from 53 to 9 minutes).

3.2 Moroccan sheep

Population structure. The variance explained by the first three PCA components was 0.0085, 0.0083 and 0.0082 (Fig. S2), respectively, indicating no clear population structure. Therefore, no variable translating population structure was retained for subsequent analyses.

Genotype-Environment associations. When investigating *Samβada's* results, a significant peak around position 4.38e7 was observed on chromosome 23 in association with annual precipitation (Fig. 2). Within this genomic region, two SNPs (i.e. ss1208941124 at position 23:43867891 and ss1208941157 at position 23:43869831) were found to be non-synonymous for the gene *MC5R* (melanocortin 5 receptor) and in strong LD (r^2=0.97). A complete list of genes associated to visible peaks in this plot (Fig. 2) is also available (Table S1).

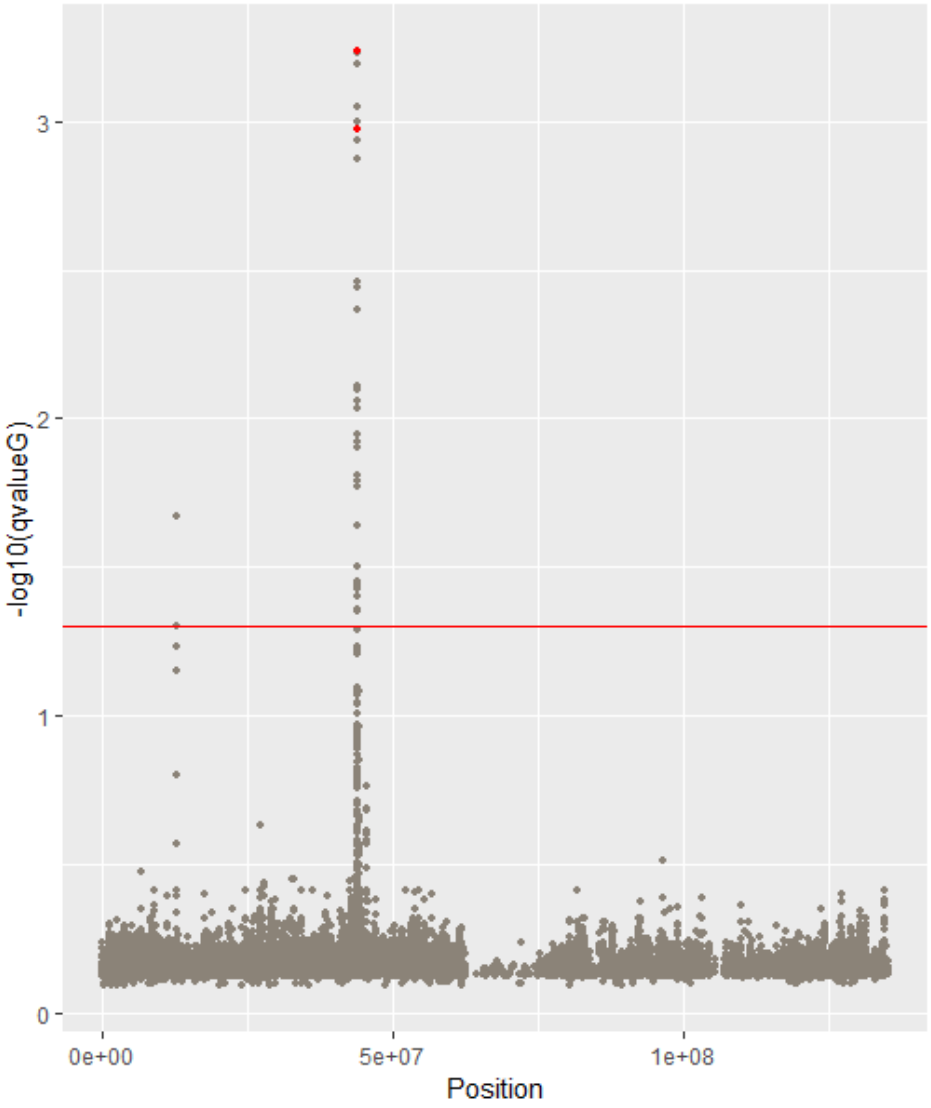


Figure III. 2: Manhattan plot of chromosome 23 of Moroccan sheep. Manhattan plot showing the q-values for each marker (with G- or Wald-Score>6) of chromosome 23 of Moroccan sheep associated with annual precipitation as calculated in *Samβada* in a univariate mode. Points in red correspond to models involving two non-synonymous SNPs (ss1208941124 and ss1208941157) in the MC5R gene (ss1208941124 having the lowest q-value of the two). The red horizontal bar shows a significance threshold of 0.05.

Given such a high LD, the spatial distribution of these markers is almost identical (except for one individual; data not shown), and only ss1208941124 is illustrated (Fig. 3). For this locus, genotype CC is very frequent in the Northern part of Morocco, where annual precipitation is on average high (reaching values of 1000 mm/year), while being almost absent in the South (at the Sahara Desert's gate where precipitation is as low as 50 mm/year).

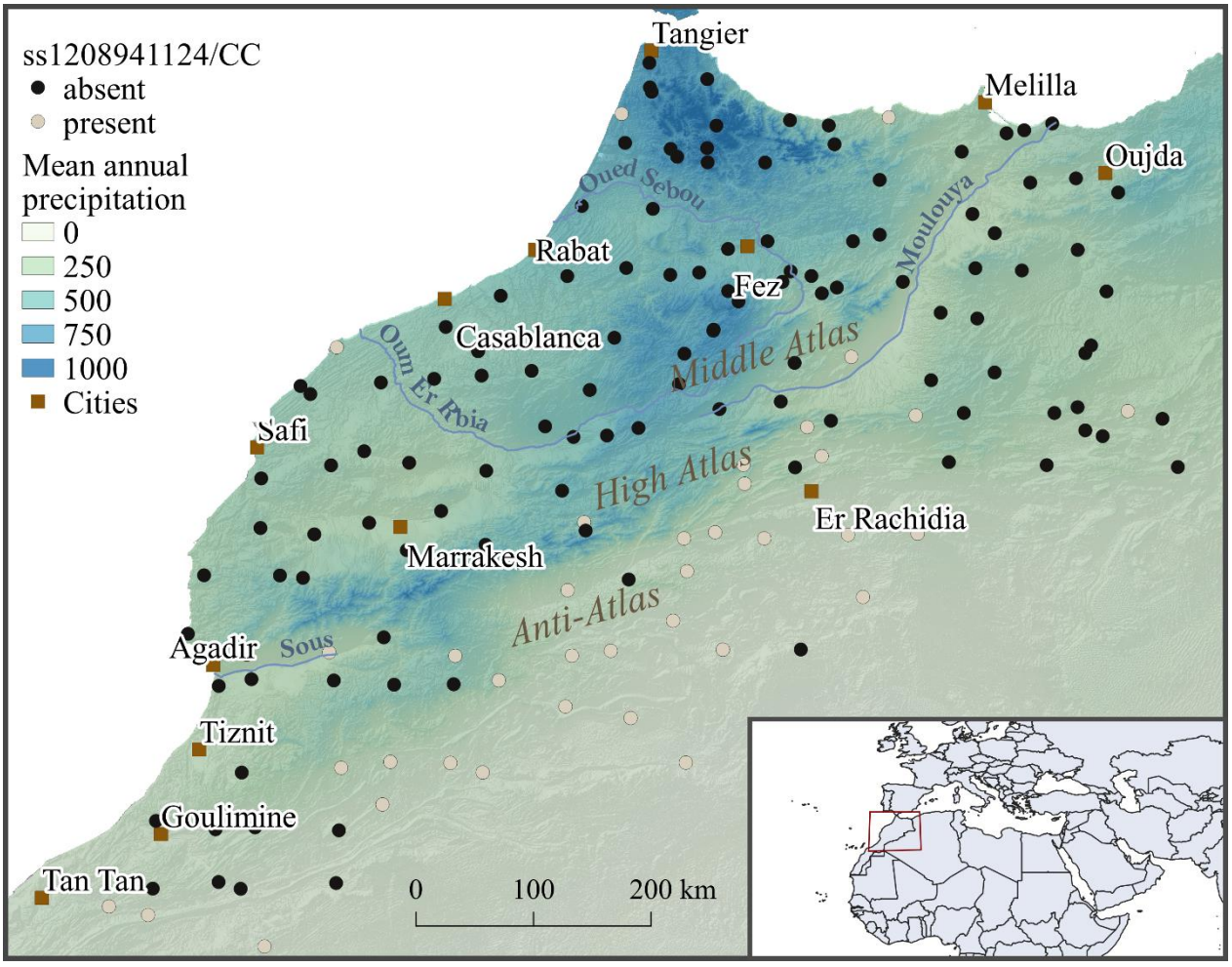


Figure III. 3: Spatial occurrence of the CC genotype for SNP ss1208941124. In the background, the shaded topography with mean annual precipitation (given in [mm/yr]) is displayed.

3.3 Lidia cattle in Spain

Population structure. The variance explained by the first three components of the PCA was 0.049, 0.029, 0.024, respectively (Fig. S3). In this case, the first principal component is likely to represent population structure, given the difference in variance observed between PC 1 and 2, and in accordance to what has been previously observed in between European cattle breeds (see e.g. Orozco-terWengel et al., 2015). Geographically, genetic clusters composed of either single or groups of proximately located farms, were identified (e.g. South from Badajoz), although no wider spatial pattern was evident (e.g. North-South gradient, Fig. 4).

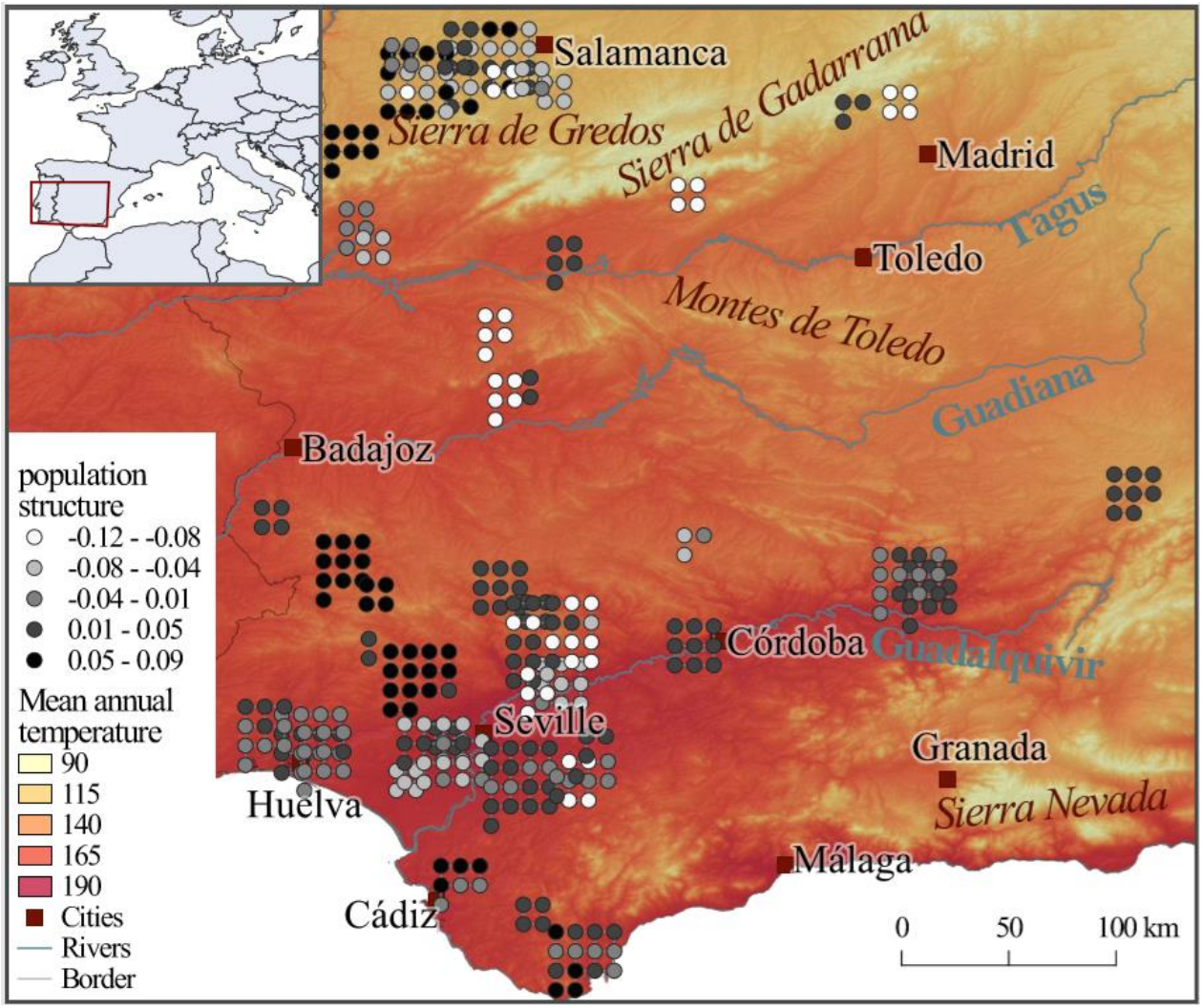


Figure III. 4: Spatial distribution of the Lidia cattle population structure. According to the scores of the first Principal Component, with a shaded relief and mean annual temperature [°C * 10] as background, as provided in the WorldClim database. Due to overlaps, close points are scattered around the farm.

Genotype-Environment associations. Several narrow peaks were observed in the models involving mean annual temperature (i.e. bio1 bioclim variable, Fig. 5). A complete list of genes associated to these peaks is available (Table S2). In particular, the Ensembl query revealed the SNP ARS-BFGL-NGS-106879 (at position 17:56127482) to be located ~30000 base pairs from the gene *HSPB8* (heat shock protein family B (small) member 8).

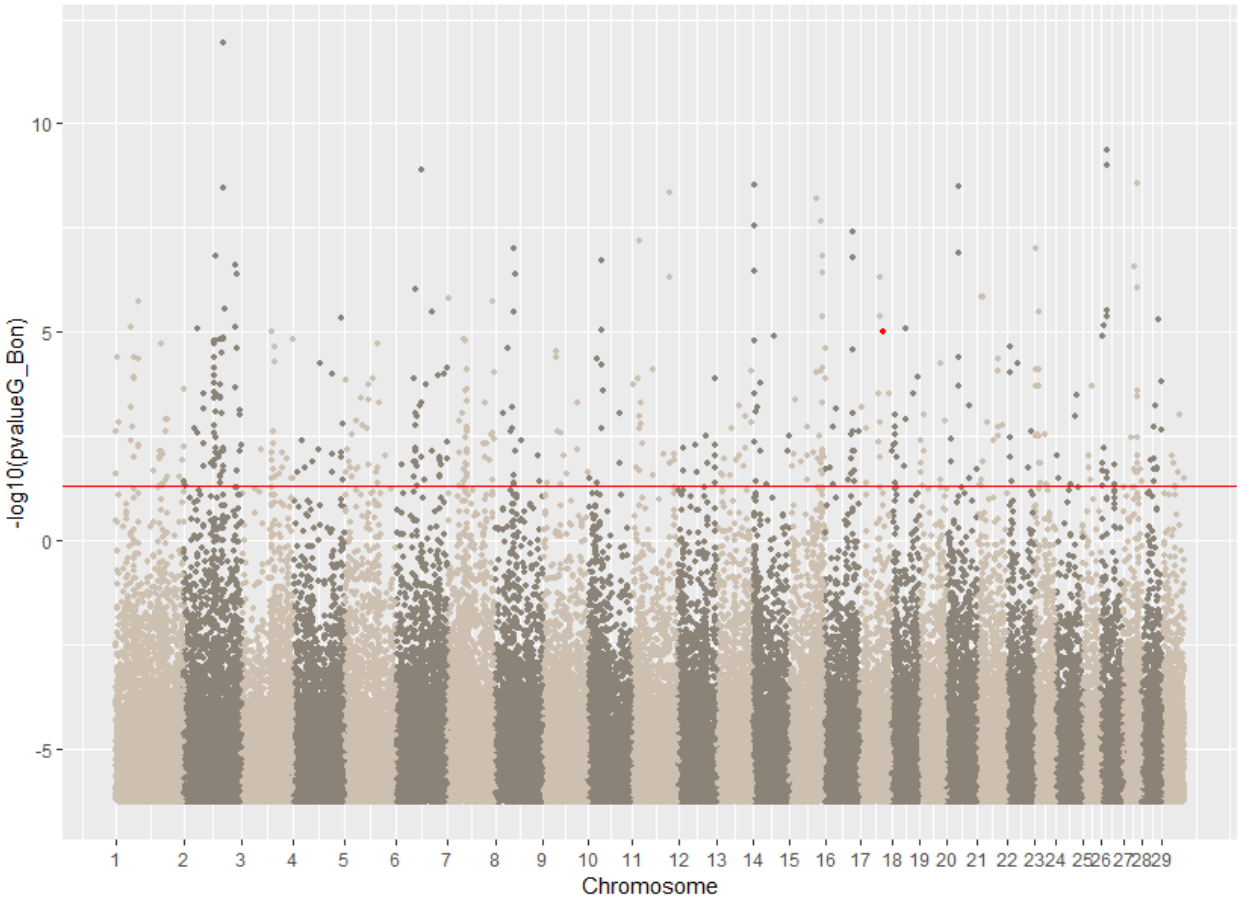


Figure III. 5: Manhattan plot of the Lidia cattle study. Showing the p-values with Bonferroni correction as derived from the *Samβada* models involving mean annual temperature and one population variable. The red point corresponds to SNP ARS-BFGL-NGS-106879, located 30000 base pairs apart from the HSPB8 gene.

Spatial occurrence of genotype AA from ARS-BFGL-NGS-106879 appears to be related to mean annual temperature (Fig. 6). More specifically, this genotype is geographically widespread in the study area, except for 23 individuals found in different farms from the Guadalquivir valley, a region with temperature reaching 36°C during the hottest month of the year. Importantly, however, when comparing Figs. 4 and 6 it can be seen that the genotype distribution does not match the prevailing population structure, hence this result is independent of the calculated population structure present within the breed.

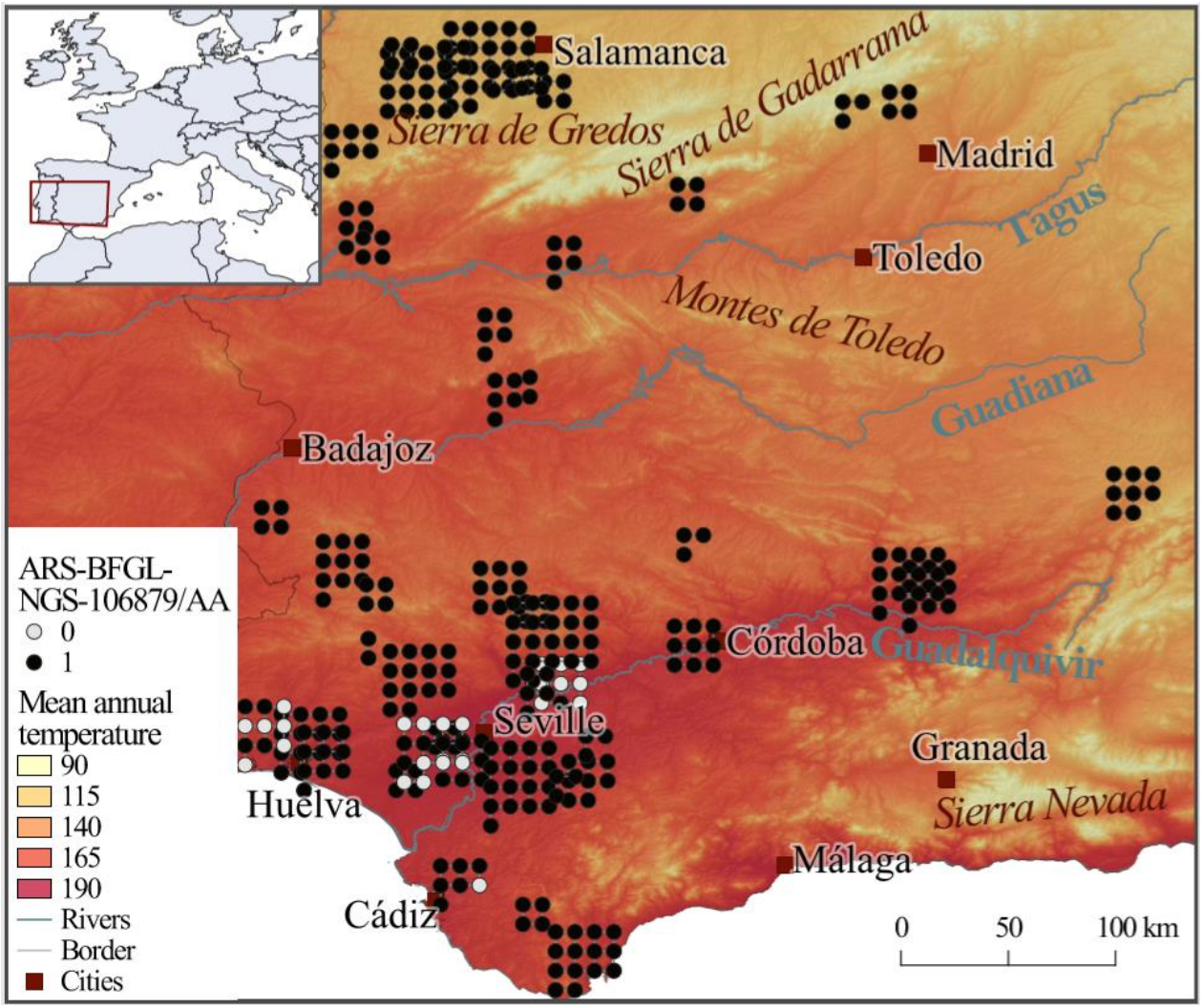


Figure III. 6: Presence-absence of the AA genotype of SNP ARS-BFGL-NGS-106879. Reported with shaded relief and mean annual temperature [°C * 10]) as background. Due to overlaps, close points are scattered around the farm.

4 Discussion

4.1 Role of the package

We have provided a demonstration of *R.SamBada*, encompassing the entire pipeline analysis from *pre-* to *post-hoc* processing, following the classical *Samβada* analysis pathway, but much more efficiently. *R.SamBada* helps saving user's time for preparing input files thanks to newly built functions, as well as computing time through better integration of population structure and automated split of computations on parallel cores. Additionally, it provides a standardised processing chain, thus facilitating reproducibility.

Moreover, part of the pre- and post-processing chain can possibly be coupled with other software used in landscape genomics and more generally with software designed to detect signature of selection. For example, the post-processing function *plotResultInteractive* could be used with any type of outputs as long as its structure is similar to the returned value of *prepareOutput* (i.e.

columns indicating the position of the SNP as well as the p-value associated with the corresponding genotype; refer to the package documentation for more detail).

4.2 Case studies

Sheep in Morocco. Two of the SNPs on chromosome 23 associated with precipitation (ss1208941124 and ss1208941157) are non-synonymous variants located within the *MC5R* gene. Although understudied in sheep, his gene has been reported to be linked to a wide range of physiological functions in different mammal species, including regulation of food intake and sebum secretion (Switonski et al., 2013). Wax secretion is of particular interest with respect to precipitation; indeed, sebaceous secretions in Merino sheep have been found to hinder *Dermatophilus dermatonomous* infection (Roberts, 1963), a skin disease affecting many domestic and wild animal species that can be lethal in extreme cases. In the same breed, Dermatophilosis outbreaks have been found to be linked with exceptionally rainy years (Yeruham et al., 1995). Thus, the secretion of wax could play an important role in protecting sheep against rainy weather, consistent with its environmental relationship with annual precipitation here.

Lidia cattle. The SNP ARS-BFGL-NGS-106879 is associated with mean annual temperature and located in the vicinity of the gene *HSPB8*. This gene is thought to code for a chaperone protein, which is upregulated in presence of heat and other environmental stress, and exerts an important cytoprotective role (Verma et al., 2016). In cattle, this gene was found to be associated with heat tolerance in both crossbred and pure *Bos indicus* Sahiwal cattle in India (Sengar et al., 2018; Verma et al., 2016), that can suggest its putative involvement with adaptation to heat tolerance in Lidia cattle as well.

This SNP lies at ~30Kbp outside the *HSPB8* coding region, either suggesting the SNP to be in LD with some adaptive variant within the gene or to possibly have an important regulatory effect on transcription. However, considering the relatively low average LD between loci at 30Kbp-distance (computed r^2 in this region=0.2), the existence of a significant variant within the gene is unlikely. In contrast, such a distance would suggest more likely this SNP to be involved in regulatory processes; indeed, according to Brodie et al. (2016), large insertions/deletions with regulative roles can be found as far as 2Mbp around a gene and associated with nearby SNPs.

4.3 Perspectives

R.SamBada represents a step forward in facilitating the chain of processes required to implement a landscape genomics study. However, several further improvements could implemented in the future. For example, the query base on the Ensembl database requires a reference genome for the species under investigation, which remains relatively uncommon for non-model species. It would therefore be very useful to further develop functions performing a BLAST alignment (Johnson et al., 2008) and see if any match can be found with orthologous genes from related species where genomes have been produced.

In addition, functionalities could be augmented to help the user define ad hoc QC thresholds. For instance, a function allowing species-specific estimation of LD in order to better calibrate the pruning applied before computing the PCA would be useful. Furthermore, *R.SamBada* currently only implements basic QC of genetic data (MAF, LD, missingness) and does not test for other useful checks (e.g. Identity By Descent – IBD – or Hardy-Weinberg Equilibrium – HWE). However, such controls can easily be performed with dedicated software like PLINK (C. C. Chang et al., 2015) or

vcftools (Danecek et al., 2011) before entering *Samβada*'s R-pipeline. Moreover, *Samβada* is one among several software solutions to detect selection signatures in a spatial context and can be used in combination with other packages like LFMM (Caye et al., 2019), BayEnv (Günther & Coop, 2013) or both (Stucki et al., 2017) in order to compare the results obtained. Further functionalities could be developed to ease the computation and comparison with those methods.

Finally, it is important to keep in mind that landscape genomic approaches such as *Samβada* implements an explanatory analysis which allows rapid identification of candidate genes, but lacks a validation procedure, meaning that derived hypotheses need to be further tested (e.g. through investigation of variant effect on protein tertiary structure and function or through lab experiments).

5 Supporting information

Figure S1: Interactive plot as a result of the function plotResultInteractive.

Figure S2: Proportion of variance explained for the first 100 axes of the PCA on molecular markers of Moroccan sheep.

Figure S3: Proportion of variance explained for the first 100 axes of the PCA on molecular markers of the Spanish cattle.

Figure S4: Average Linkage Disequilibrium (LD) per distance class in the Spanish cattle molecular data.

Table S1: List of genes and corresponding description within a window of 25000 bp around the two major peaks of the manhattant plot on Moroccan sheep.

Table S2: List of genes and corresponding description within a window of 50000 bp around SNPs belonging to major peaks of the manhattan plot of Spanish cattle.

6 Resources

6.1 Software availability

R.SamBada package is available in the *R* CRAN package repository and on GitHub (github.com/SolangeD/R.SamBada).

6.2 Data accessibility

The Moroccan sheep dataset is available https://projects.ensembl.org/nextgen/ population MODA. The Lidia cattle dataset is accessible from FigShare: https://doi.org/10.6084/m9.figshare.5394895.v4 (only Spanish samples included in the analysis).

7 Author Contributions

SDur wrote the major part of the R-package with the help of OS, EV and SS on specific points. In particular the new functionalities of the C++ code were developed by SS. NS wrote the sections of

the manuscript dedicated to the Lidia cattle case study. KL performed most of the analysis related to the Moroccan sheep case study and elaborated part of the related text. SDur wrote the rest of the manuscript with the help of all authors. SJ conceived and supervised the project. PO, MWB and SJ revised the manuscript.

8 Funding

This work was supported by the European Union 7th framework project NEXTGEN (Grant Agreement no. 244356, coordinated by P.T.) and the FACCE ERA-NET Plus project CLIMGEN (grant ANR-14-JFAC-0002-01). MWB and POTW were funded by BBSRC through the FACCE-JPI ERA-NET Climate Smart Agriculture project CLIMGEN (BB/M019276/1). NS is a recipient of a Marie Skłodowska-Curie Individual Fellowship funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No DLV-655100.

TAKE HOME MESSAGE

- Landscape genomic studies associate genetic variation with the environment in which the animal lives and is a concrete example of the application of *biogeoinformatics*.
- Applied to livestock species, this approach can highlight SNPs that are important to preserve and could therefore be included in selective breeding programs.
- *Samβada* is an (already existing) program written in C++ to perform these studies, with special emphasis on High Performance Computing (HPC).
- The full processing chain is time consuming and requires knowledge in biology, GIS and computer science. Typically it involves the following steps
    - Filtering of the molecular data
    - Retrieval of the environmental conditions
    - Assessment of the population structure
    - Statistical testing of the association between genome and environment
    - Creation of plots and maps displaying the results of the study
    - Link between plots and biological databases storing the function of genes
- We developed an *R* package called *R.SamBada* that proposes an automated pipeline in which all the above-mentioned steps are included.
- Two case studies on Moroccan sheep and Spanish cattle highlighted SNPs associated with high temperature and precipitation, located in or next to genes with biological functions linked to these kinds of environment.

Chapter 4
Preserving a traditional farming technique suited for local breeds

While the first two chapters targeted means of preserving local breeds by monitoring their genetic diversity and inspecting locally adapted genetic variations to be safeguarded, this last chapter steps back and focuses on the production system in which they are involved. Indeed, the preservation of these systems is strongly connected to the conservation of local breeds, as well as being a way of saving the socio-cultural heritage associated to the breed. In this chapter we attempt to better understand a specific farming technique: the transhumance system, which we will call alping as a reference to the German word “Alpung” or “montée à l’alpage” in French. Indeed, many folkloric events are associated to this system, such as the “combat de reines” where cows brought up to the mountains fight among each other to determine the queen (i.e. herd leader, Valais Promotion n.d.) or the “désalpe” celebrating the return of the cattle to the lowland farm (Fribourg Région n.d.) and “Chästeilet” a ceremony in which the cheese produced in the mountain is shared among breeders (Schweiz Tourismus n.d.). While milk quantity is known to be reduced when the cow is brought to the mountains, milk characteristics in turn are altered, which will then be exploited to create specific cheese: eight Protected Designation of Origin exist for Swiss alp cheese, along with many other “unprotected” types (Schweizeralpkaese n.d.). Furthermore, mountain pastures are considered as one of the richest natural area of Switzerland and the alpine economy defines the specificity and identity of Switzerland (Lauber 2013). Since alping requires locally adapted cattle, there is no doubt that the preservation of the system will indirectly help in protecting local breeds.

A large amount of data to describe milk production of Swiss dairy cows have been collected routinely for decades, creating a huge and complete database. Essentially, milk production and characteristics of every dairy cow is recorded monthly, together with phenotypic description of the animal (preliminary summary statistics are available in Appendix D). These data are used in breeding programs but, considering the amount of information included, remain under-exploited, especially in the alping context. The drop in milk production experienced by mountain-pastured cows can be explained by the need to adapt to a new environment, the different forage composition and harsh weather condition. In this context, *biogeoinformatics* offers a unique way of investigating the exact influence of the climate. Provided precise weather condition are available on a daily basis, one can associate the observed milk production on a given day with the weather condition endured by the animal at the location of the alp. Coupled with biological knowledge on the metabolism of cows, mathematical modelling, statistical testing and efficient storage of the information in a database, this environmental information can provide relevant insights to better understand the impact of alping and of climate change on this system.

As the first author of this article, I completed most of the tasks, both in running the analyses and in writing the manuscript. The other authors provided advises about the methods to be used and the writing of the paper, the complete list of contribution being available at the end of the chapter.

Duruz, S., Vajana, E., Burren, A., Flury, C., Joost, S. (2020). Big dairy data to unravel the effect of environmental, physiological and morphological factors on milk production of mountain-pastured Braunvieh cows. *Royal Society Open Science*, 7(7).
doi: 10.1098/rsos.200638

1 Introduction

Transhumance, which consists in moving livestock to high mountain pastures in the summer months, provides both ecological and socio-cultural services to the human populations living in the mountainous regions of many European countries (Bunce, Pérez-Soba, and Smith 2009; Liechti and Biber 2016; Olea and Mateo-Tomás 2009). Indeed, transhumance-annexed grazing sustains and preserves endemic plant communities (Herzog et al. 2005), feed local cattle to produce traditional alpine cheese, and attract many tourism-related activities (Gellrich and Zimmermann 2007). Further, it counteracts land abandonment in mountain areas and therefore contributes towards preserving landscape against scrubs growth and vegetation encroachment (Gellrich et al. 2007), as well as natural hazards such as avalanches (Newesely et al. 2000) and wild fires (Gellrich and Zimmermann 2007). The term "alping" (a translation of the German word "Alpung" or its French equivalent "montée à alpage") will be used here to describe the approximately 100 days that dairy cattle spend on alpine pastures during the summer months. Similarly, animals brought to mountain pastures will be referred to as "alped" cows, and the alpine summer pastures will be called "alps".

Despite such ecological and social benefits, the surface dedicated to alping decreases each year (~2400 ha per year, Lauber 2013), and a questionnaire-based study revealed in 2010 that one third of the participating breeders intend to probably abandon the transhumance practice in the following decades. In summer 2018, 107'000 dairy cows were alped in Switzerland during approximately 100 days (BLW 2020). A steep drop in milk production is observed during this period, which hampered the evaluation of lactation curves through standard models that assume a linear decrease in production (Jeretina, Babnik, and Skorjanc 2013) after the maximum milk yield is reached (i.e. ~100 days after calving) (Wood 1967). Among the explanations proposed to interpret such a detrimental effect on productivity are the feed deficit intake due to the meagre grassland as found in high alpine pastures, as well as the need to tackle environmental stress due to new and sometimes harsh habitat conditions (Zendri et al. 2016). On the other hand, milk composition is known to change during alping (Cassandro et al. 2008; Jõudu et al. 2008) and results in the production of highly valuable milk products such as butter and alp cheese.

Milk production and quality is notoriously affected by a wide variety of environmental factors, including calving season, vegetation types composing animals' diet (Hahn 1999; Hayes et al. 2003; Tekerli et al. 2000; Wilmink 1987). Environmental temperature is also known to directly affect cattle productivity because of heat (Hayes et al. 2003) or cold (Bryant et al. 2007) stress. Furthermore milk quality and production of alped cows are expected to be indirectly affected by global warming, as forage quality and biomass productivity of alpine sites are likely to decrease with increasing temperature and decreasing precipitation (Gilgen and Buchmann 2009; Signarbieux and Feller 2008).

Despite the existence of huge databases storing monthly milk records for several European cattle breeds, no effort has been produced so far (at least to our knowledge) to exploit such an information and understand the ways alping affects milk productivity (Jurt, Häberli, and Rossier

2015). Indeed, most of the existing literature focuses on small experiments (with sample size <100) mainly restricted to compare two groups of animals in different environmental conditions, so as to investigate the potential effects of altitude (Gorlier et al. 2013), vegetation type (Gorlier et al. 2013; Leiber et al. 2006), supplemental feeding (Berry et al. 2001; Bovolenta, Ventura, and Malossini 2002), calving season (Horn et al. 2014) or breed (Horn et al. 2013, 2014; Zendri et al. 2016). Furthermore, no adaptation of general models of lactation curves (Wood 1967) have been proposed to account for alping, which hinders a straightforward comparison of lactation curves for alped cow. Last but not least, the overall impact of environmental factors and global warming on milk production during alping is also still unknown.

For these reasons, a better understanding and characterisation of the impacts of transhumance on milk production and the way production is influenced by environmental factors is needed. To fill this gap, we relied on over five million monthly test-day milk records collected between 2000 and 2015 from more than 200,000 Braunvieh cows, a local Swiss cattle breed well adapted to the alpine pastures. Then, we used this information to: 1) devise a new mathematical model to fit lactation during alping; and 2) investigate the influence of the environment on milk production during alping and compare it with the effect of physiological and morphological factors. This can be achieved thanks to biogeoinformatics which takes advantage of geo-referenced animal data in order to link biological and environmental information with the help of advanced informatics tools (Bertaglia, Joost, and Roosen 2007).

2 Data

2.1 Milk records and animal information

Milk records from all alped Braunvieh cows were provided for the period 2000-2015 by the Braunvieh Schweiz AG breeding association. Importantly, a direct comparison with non-alped cows was not possible because we did not have access to these data. However, as milk measurements of alped cows entail records from both the lowland farm and the alp, the estimation of milk production in both situations was feasible1000. The full dataset is composed of 5,681,498 test day records (methods A4 and AT4 according to ICAR-Guidelines (ICAR 2014)), including 616,081 lactations derived from a total of 245,313 cows. In line with national and international rules, milk records are taken approximately on a monthly basis, with the first record taken between the 5th and 42nd day after calving. Each test day record includes information on the following traits: Milk (kg), Fat (kg and %), Protein (kg and %), somatic cell count (1000 cells/ml). To keep the reader focused on the main thread of the article, our study specifically analyses milk production in terms of quantity (milk yield); however results from computations with protein and fat content and yield are also available in supplementary materials (Sup. Mat. S2-S5). Out of the total number of records, 1,481,387 were taken in the alps, whose altitude were systematically stored in the database, while their precise location were documented in 95% of the cases (Fig. 1). The first record in the alp is usually taken within the first four days after arrival, and is followed by three more records in the alp to encompass the entire alping period (typically 100 days). Moreover, to morphologically describe animals, linear type description and classification of cows are scored during the first lactation of all cows of the database. In our study we considered the body height at withers and the scores (1-9) for foot angle. In addition, insemination data for each lactation (date, sire's name) are also available.

A stringent data quality control procedure was applied prior to analysis to remove: 1) incomplete years (which resulted in removing beginning of 2000 as well as end of 2015 due to missing lactation records); 2) cows with average interval between first and last insemination longer than 100 days (as computed over the first three lactations); 3) cows that had their first calf while being younger than two years, or older than four years; 4) cows belonging to breeds different from the Braunvieh or Original Braunvieh; 5) cows with parents other than Braunvieh or Original Braunvieh; 6) lactations shorter than 270 days; 7) lactations with calving interval shorter than 290 days; 8) lactations with alps below 1100 meters above sea level (masl) or above 2600 masl; 9) lactations with calving happening between March and August; 10) lactations from cows that had already calved more than nine times; 11) lactations with the first record taken after the 42nd day after calving; 12) lactations with records taken before calving; 13) records taken before the 5th day and after the 500th day after calving; 14) the second alping season (i.e. final part of lactation curves) from animals that are alped twice in the same lactation. After filtering, we obtained a final dataset composed of 3,527,138 records over 371,696 lactations from 175,474 cows.

2.2 Factors influencing milk characteristics

Milk characteristics are known to be influenced by different factors. Meaningful predictor variables were then selected according to literature review, by assuming the same factors to be relevant in both lowland and mountain conditions. As a result, climatic and environmental indices (Bryant et al. 2007; Jonas et al. 2008) were taken into account together with physiological (lactation number, pregnancy stage, Hayes 2013; Olori et al. 1997) and morphological factors (Table 1).

Table IV. 1 : List of factors included in the present study with supposed influence on lactation during alping. Factor-specific cut-off values are reported in the last column. These values are used to assess factor-specific effects on lactation (see Methods for an exhaustive explanation).

	Name	Description	Group cut-off
Environmental	Temperature Humidity Index (3 days)	Climatic index based on temperature and humidity, averaged over 3 days before milk record at the alp. See section on climatic data	59.4-65.4 / >65.4
	Temperature Humidity Index (30 days)	Climatic index based on temperature and humidity, averaged over 30 days before milk record at the alp. See section on climatic data.	59.6-63.1 / >63.1
	Cold Stress Index (3 days)	Climatic index based on temperature, wind speed and precipitation, averaged over 3 days before milk record at the alp. See section on climatic data.	960.1-1045.9 / >1045.9
	Cold Stress Index (30 days)	Climatic index based on temperature, wind speed and precipitation, averaged over 30 days before milk record at the alp. See section on climatic data.	997-1042.9 / >1042.9
	Spring precipitation	Average monthly precipitation [mm] between April and July; computed for each year, at the location of each alp.	<120.6 / >155.3
	Biogeographical region	Only regions with sufficient sample size were retained, and therefore two categorical variables were created. See section on biogeographical regions. .	North Alp / East Alp
	Altitude	Altitude [m] of the highest alp during the lactation cycle	<1600 / >1900
	Altitude difference	Difference in altitude between the highest alp and the lowland farm.	<641 / >1021
	Aspect 100m	Aspect of the alp (North/South facing) as based on 100m-resolution DEM.	300-60 / 120-240
	Aspect 1km	Aspect of the alp (North/South facing) as based on 1km-resolution DEM.	300-60 / 120-240
Physiological	Lactation number	Number of lactations the cow experienced since birth (correlated with animal age).	1st lact / ≥3rd lact
	Pregnancy stage	Pregnancy stage [days] at the beginning of alping.	<73 days / >153 days
Morphological	Height of animal	Height at withers [cm]	<139 / >143
	Foot angle	A note between 1 and 9 (with 9 being the steepest).	<4 / >6

2.3 Climatic data

Climate has been observed to influence milk production (Ugurlu et al. 2014). Consequently, maximum and mean temperature (Gorlier et al. 2013) as well as daily rainfall (Leiber et al. 2006) were extracted from the meteoswiss Grid-Data products database (Meteoswiss n.d.). This dataset

is derived by interpolation of records from several weather stations across Switzerland, and consists of 2km-resolution raster files (1km-resolution from the year 2014 and on). Further, daily average wind speed and relative humidity were obtained from respectively 440 and 495 meteoswiss weather stations. We then interpolated these values between stations to obtain a continuous representation of the variables, with a squared inverse-distance weighting (IDW) (Luo, Taylor, and Parker 2008) within a maximum distance of 50 km.

On the basis of such environmental data, the Temperature Humidity Index (THI) and Cold Stress Index (CSI) were computed following Bryant et al. (Bryant et al. 2007). These indices assess all relevant climatic conditions for the evaluation of "hot"/"cold" sensation instead of focusing on temperature only:

$$THI = 0.8T + \left(RH/100 \cdot (T - 14.4)\right) + 46.4$$

Equation IV. 1 : Temperature Humidity Index (THI) formulation

with *T* being the maximum daily temperature [°C] and *RH* the relative humidity [%], and

$$CSI = \left(11.7 + (3.1 \cdot WS^{0.5})\right) \cdot (40 - T) + 481 + 418 \cdot (1 - e^{-0.04 \cdot rain})$$

Equation IV. 2 : Cold Stress Index (CSI) formulation

with *WS* being the daily mean wind speed [m/s], *T* the mean daily temperature [°C] and *rain* the daily precipitations [mm]. These indices were computed over a 3- and 30-day period to account for short/long heat waves/cold spells, respectively.

2.4 Digital Elevation Model

Due to the coarse spatial resolution of temperature data (2-km), a correction of -0.45°C/100m (i.e. the observed temperature gradient in the dataset) was applied to account for local variation in temperature due to topography. This correction was achieved using both the Digital Elevation Model DHM25 dataset produced by swisstopo (Swisstopo n.d.) and the recorded altitude of the alp available in the dataset. The digital model DHM25 is a tridimensional representation of the earth's surface in Switzerland, as based on the elevation data from the Swiss National Map 1:25,000 (NM25). A symmetric 25-m grid matrix model is then interpolated starting from the digitised contour lines and spot heights from NM25. Comparisons among control points shows an average accuracy of the produced model of 2-3 m for the pre-Alps and Alps, respectively.

2.5 Biogeographical Region

The Federal Office for Environment (FOEN) divided Switzerland into six biogeographical regions (FOEN 2001), obtained using fauna and flora data and aggregating areas with common species. Species distributions being strongly related to the relief, these regions reflect in fact the topography of the country. Most of the alps hosting Braunvieh cows appear to be located in the Northern and Eastern Alps biogeographical regions. More rainfall occurs in the Northern Alps when compared to the Eastern side (Fig. 1), because the mountain chain acts as a barrier to precipitations coming from the West and North (Meteoswiss 2018).

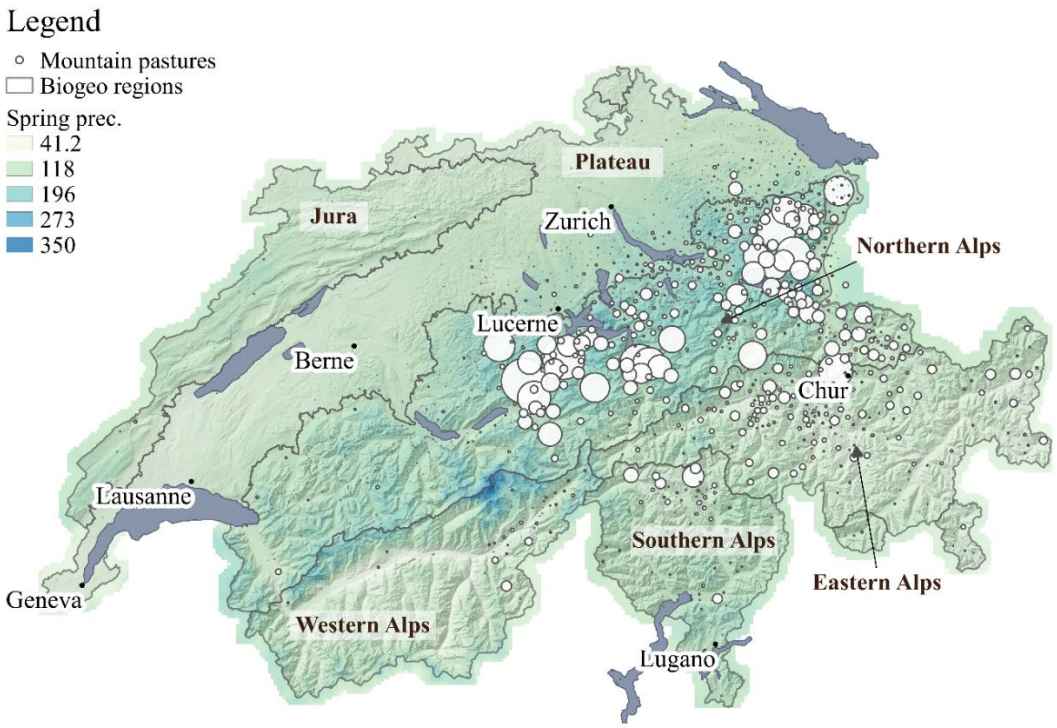


Figure IV. 1 Geographic location of the alps hosting Braunvieh cows (white circles), with average monthly precipitation in mm between April and July 2015 in the background (chosen as example year). Frontiers of biogeographical regions are also reported. The size of the circles is proportional to the number of milk records taken at a given alp, the biggest and smallest circles encompassing 28923 and 1 records, respectively. The majority of the alps hosting Braunvieh cows are located in Northern and in the Eastern Alps biogeographical regions.

3 Methods

3.1 Lactation curve modelling

A lactation curve is usually estimated from one single cow with repeated observations along a lactation cycle and with records taken on a daily/weekly basis (Wood 1967). Here, test-day milk records were collected monthly, making the individual-based estimates of lactation impossible because of the over-parameterisation issue faced when the number of observations is small (typically 10 monthly measurements during a whole cycle) with regards to the number of parameters to estimate, particularly when describing a complex curve like the one of alped cows (6 parameters, see below). Moreover, a measurement is highly influenced by local temporal variations linked to some momentary discomfort of the animal, so that the curve resulting from monthly records are exceedingly noisy. Therefore, we analysed averaged values by computing, for each test-day, the mean of all available records of that particular Day In Milk (DIM, or number of days after calving). Given that records from the same animal are one month apart but that they are not taken on the same DIM for all animals, the average of milk records for each test-day will constitute a smooth curve with daily values (as displayed in point observations of Fig. 2). As dates at which cows are alped or brought back to the lowland farm slightly differ among animals, records from cows remaining at the lowland farm during the alping season (between the 15th of May and the 31st of August) were excluded from this average computation, while only cows at the lowland

farm were considered in the average outside this time frame. Moreover, cows were grouped according to their calving month. Finally, when fitting the curve, each averaged milk yield was weighted according to the number of observations on that day.

Several models have been proposed to describe lactation curves (Val-Arreola et al. 2004), with the Wood, Wilmink, Ali-Schaeffer (AS) and Legendre polynomial formulations being the most popular (Macciotta, Vicario, and Cappio-Borlino 2005). Among these mathematical formulations, Wilmink proposes a linear equation that is retained in the present work given its inherent simplicity and good performance (Macciotta et al. 2005). This model is written as:

$$Y_t = a + b \cdot e^{-k \cdot t} + c \cdot t$$

Equation IV. 3 : Lactation curve modelling with Wilmink Model

where Y_t is the observed variable (milk yield), t is the DIM, and a, b, c and k are the parameters to estimate. However, k is usually set to 0.1 to make this equation linear (Macciotta et al. 2005). To validate this value with our data, non-linear regressions were also run with 6 test curves (one for each calving month) and the obtained values for k were between 0.05 and 0.37.

Here, we introduce additional terms to Eq. 3 in order to explicitly account for the transhumance effect. Particularly, alping has been observed to severely affect milk production, with alped animals showing a steeper linear decrease than before alping (Fig. 2). Further, alped cows usually experience a small yet rapid boost shortly after their return to the lowland farm, followed by a softer decline in milk production. Taking these observations into account, we then propose to adapt Eq. 3 as follows:

$$Y_t = a + b \cdot e^{-k \cdot t} + c \cdot t + d \cdot max(0, t - t_1) + f \cdot max(0, ceiling\ (t - t_2)/305) + g \cdot max(0, t - t_2)$$

Equation IV. 4 : Newly proposed equation for the modelling of lactation curve of alped cows

Where t_1 is the DIM at which the cow is alped, and t_2 is the DIM at which the cow is brought back to the lowland farm. Importantly, the expression d*max(0,t-t1) is the expected linear decrease during alping, so that the d-parameter reflects the effect of alping. The f*max(0,ceiling(t-t_2)/305 captures the expected boost in production after alping and g*max(0,t-t_2) represents the linear decrease in milk yield after alping; in the latter arguments, the max() term ensures the model to be only affected during and after alping respectively, while the ceiling expression (i.e. round to the upper integer) constructs a binary operator (0/1) to recreate the instantaneous boost after the return to the lowland farm. In our case, t_1 and t_2 were determined independently for each calving month. The proposed equation only works for a standard lactation period of 305 days.

The d-parameter enables the estimation of the loss in milk yield associated with alping over a given period of time. Indeed, the amount of milk lost during alping for a period of x days can be approximated with

$$Y_{loss} = \frac{d \cdot x^2}{2}$$

Equation IV. 5 : Estimation of milk yield loss during alping

However, it is essential that the model fits well the beginning of the curve for this equation to work, which can be achieved by artificially increasing the weight of point measurements before the transhumance. Thus, weights before alping were multiplied by 100 when investigating the d-

parameter depending on the calving month (Figs. 2 and 3). Furthermore, as older cows tend to calf later in the season, thereby creating a correlation between lactation number and calving month, the impact of alping according to the calving month is entangled with lactation number. Therefore, when examining milk production and the impact of alping for each calving month, only cows in their first lactation were considered (Fig. 3).

Ordinary linear regression models were then computed in R using the lm() function of the stats package (R Core Team 2018) to estimate parameters in Eq. 4.

3.2 Measuring the effect of influencing factors

For sake of interpretation, all influencing factors (i.e., explanatory variables) were grouped into environmental, physiological and morphological categories (Table 1). The effect of influencing factors was tested by comparing milk records produced in conditions as dissimilar as possible. Importantly, since the low number of measurements per animal imposed the use of averages, effect determination was not possible through classical regression models. Consequently, groups were created according to the first and third tertile of the distributions, in order to include animals from the most contrasted situations (environmental, physiological and morphological) while retaining enough observations to guarantee a sufficient statistical power. Since productivity is known to be optimised with mild weather conditions (Ugurlu et al. 2014), exceptions were made for THI and CSI where the second and the third tertiles were used as the two contrast groups instead of the first and third tertile.

Group membership was assessed through the creation of a dummy variable assuming the value of 1 if belonging to the group considered, 0 otherwise. Then, the impact of influencing factors was computed by adding an interaction term to Eq. 4 that allows chosen parameters to vary as a function of the group. The here defined environmental variables affect milk production during the alping stay only. Accordingly, lactation curves were modelled only until the end of the alping season (meaning the f and g parameters not to be estimated), with the sole *d*-parameter varying as a function of the group. In contrast, physiological and morphological factors influence the whole lactation cycle, so that all terms of Eq. 4 (coefficients a, b, c, d, f and g) are allowed to vary as a function of the group.

Within-group production was estimated both at the lowland farm and in the alps for physiological and morphological factors or during alping season only for environmental factors, by integrating the area under the lactation curve. The between-group difference was then assessed by computing the percentage of the difference in milk production with respect to the reference group, this group being arbitrarily chosen as the one with the highest milk production during alping. The difference in the *d*-parameter (Δd) between the two groups is then also displayed to show how differently the concerned groups were impacted by alping. As the response differs according to the calving months, results were computed for each calving month separately and the months of September and February were chosen as representative of autumn and winter calving, respectively.

3.3 Significance testing

Log-likelihood ratio tests were performed to investigate both the impact of adding the parameters d, f and g to the Wilmink model, and of the considered influencing factors. When testing the addition of parameters d, f and g to the Wilmink equation, Eq. 3 and 4 were considered as null and alternative models, respectively; when testing the influencing factors, the null model was

constructed by removing the interaction between the dummy variable group and the parameters of Eq. 4.

The resulting G-score test-statistics were then converted into p-values, which were further corrected for multiple testing by means of the Bonferroni's approach (Bonferroni 1936). G-scores are efficient ways of testing the performance of a nested model, and are slightly less conservative than Wald scores (Gourieroux, Monfort, and Trognon 1983). This seemed appropriate here since the applied correction for multiple testing is already sufficiently conservative. G-scores were evaluated using the lrtest() function from the lmtest R-package (Zeileis and Hothorn 2002).

4 Results

4.1 Lactation curve modelling

Overall, the proposed equation fits both the drop in milk production due to alping and the tail of the lactation curve, as illustrated here for the calving months of September and February (Fig. 2). In particular, the terms added to the Wilmink equation (Eq. 4) significantly increase the full model performance (p-value<10-16). In the case of autumn calving (Fig. 1a), the proposed equation fits the entire lactation cycle. For winter calving (Fig. 2b), the beginning and the end of the transhumance season appear to be the most challenging periods to be fitted because of a non-linear slope. The use of Eq. 5 can be illustrated with the autumn calving, with a *d*-parameter of -0.08, which is translated by a loss of 144 kg over 60 days. The modelling of protein and fat content curves are also available (Sup. Mat. S2-S3).

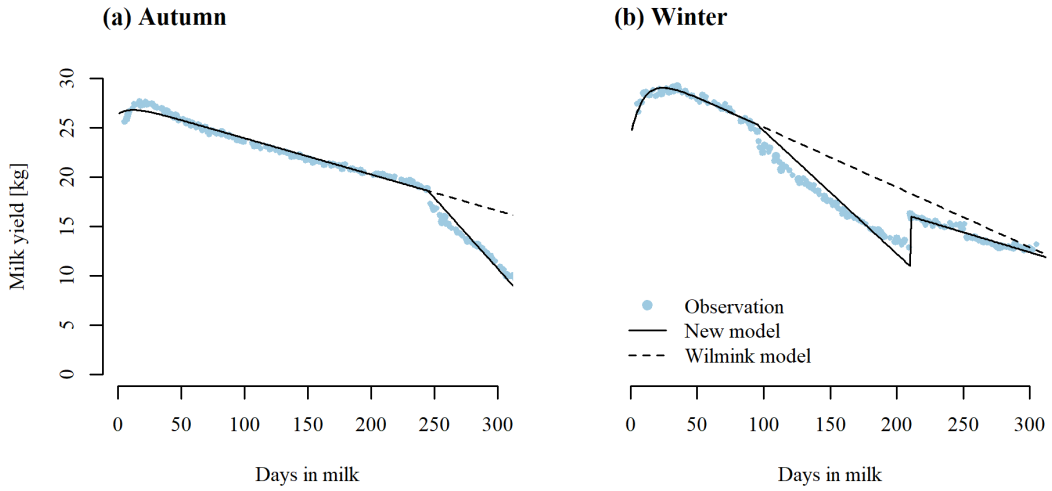


Figure IV. 2 : Lactation curves as derived from the proposed model (full line) and the Wilmink model (dashed line) for cows that calved in September (a) and February (b). The Wilmink model was fitted using points from the beginning of the curve only, i.e. before alping. . Each dot represents the average of milk records per day. When t>245 (a) and between 95 and 210 (b), records from the alp only are used to calculate the average, whilst records from the lowland farm only are included for the remaining time frame.

Total milk production and milk production during alping is reported for the calving months of September and February (Fig. 3). For the sake of comparison among months, only cows in their first lactation are considered in this graph, as lactation number and calving month are correlated.

Cows calving in autumn produce on average 6033 kg during their first lactation, among which 1320 kg are produced in the alp. In contrast, total milk production turns out to be lower for cows calving in winter (5155 kg during their first lactation), while milk production during alping is increased (1755 kg). The *d*-parameters for the two calving seasons being markedly different (-0.08 and -0.02 for autumn and winter calving respectively) indicates that productivity is more impacted by alping when calving occurs in autumn than when occurring in winter.

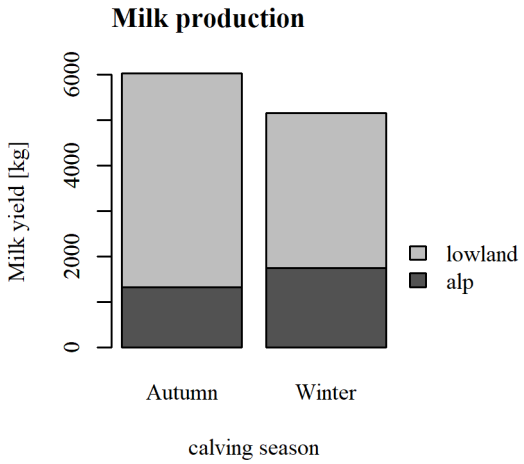


Figure IV. 3: Milk production during alping (black) and from the lowland farm (grey) is reported for autumn and winter calving, as represented by the months of September and February, respectively. Only cows in their first lactation are considered here.

4.2 Effect of influencing factors

The significance of the interaction between the group variable and the *d*-parameter is reported (Sup. Mat. S1). Hereunder, only factors with at least one calving month having a significant Δ*d* (i.e. a significantly different impact of alping between the two contrast groups) are presented. The influence of these factors on protein and fat yield is also computed (Sup. Mat. S4-S5).

Among environmental conditions, THI, spring precipitation, biogeography and altitude turned out to show a significant effect on milk production during alping (Fig. 4a-l). Particularly, precipitation in spring and the biogeographical region showed the most important difference on milk production during alping, followed by altitude and altitude difference. Further, calving period appears to interact with environmental conditions, with bigger differences between groups being present in autumn.

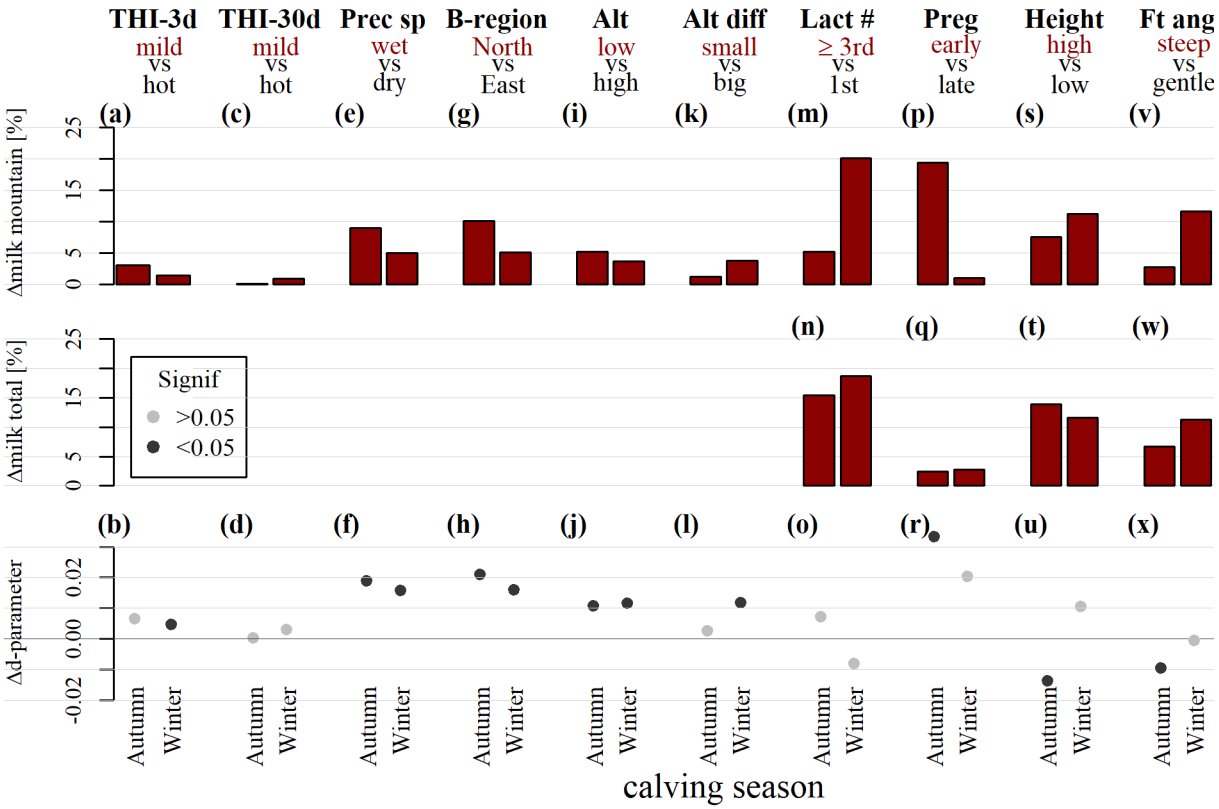


Figure IV. 4 Effect of influencing factors The effect of influencing factors is tested by investigating the difference in productivity between two groups of animals coming from contrasted conditions (first and third tertiles, except for THI where second and third tertile are chosen). Each factor is here reported in a separate column. At the top of each column, the factor name as well as the contrasted groups are reported; the group with highest milk yield during alping is chosen as the reference group, highlighted in red. In each barplot, the first bar shows the result for autumn calving, and the second for winter calving. The between-group difference in milk production during alping is displayed in the top panel, the between-group difference in milk production during the whole lactation in the intermediate panel, the change in the *d*-parameter at the bottom. The Δ*d*-parameter indicates how the reference group is impacted by alping compared to the other group, with positive values meaning lower negative impact (see Eq. 4). Significant Δd values are plotted in black, while grey indicates non-significance. Environmental factors affects production during alping only, making a comparison of the whole milk production redundant (which is why no graph is present in the intermediate panel of the concerned variables). To facilitate the understanding of this graph, the example of lactation (Lact #) is detailed here, where we refer to cows in their third or higher lactation as third lactation cows: third lactation cows produce 5% more milk during alping than first lactation cows when calving in autumn and even 20% more when calving in winter (m). When considering the whole lactation, third lactation cows produce 15% more milk than first lactation cows when calving in autumn and 19% when calving in winter (n). Third lactation cows calving in autumn are slightly less negatively impacted by alping (positive Δd) than first lactation cows; an inverse behaviour is observed for winter calving, although both relationships are not significant (o). THI: Temperature Humidity Index, as averaged over 3 (THI-3d) or 30 (THI-30d) days. Prec sp: precipitation in spring. B-region: biogeographical region. Alt (diff): (difference in) altitude, lact #: lactation number, Preg: pregnancy stage, height: height at withers and ft ang: foot angle.

The effect of environmental factors are small compared to those of physiological factors, where the biggest effect is found for pregnancy stage for winter calving with a difference in milk production during alping of 20%. Although third and higher lactation cows produce more milk during the whole lactation cycle including alping (Fig. 4m), they also appear to be more impacted by alping than the first lactation cows as highlighted by negative Δ*d*-values (Fig. 4o). The influence of pregnancy stage appears to affect milk production during alping, especially for cows calving in autumn (Fig 4p and 4r). Further, higher cows and/or with steeper foot angle produce more milk both before and during alping than lower ones with gentle foot angle (Fig. 4s, t, v, w). However, alping appears to negatively impact such cows, especially higher ones (Fig. 4u and 4x).

5 Discussion

5.1 The importance of calving season

The proposed model succeeded in quantifying the impact of alping on milk production by assuming a Wilmink pattern for non-alped cows (Fig. 2; Macciotta et al. 2005). This assumption is consistent with literature findings on the same breed [45] and was further validated with animals in our dataset that were alped at a very late stage in their lactation cycle. However, future studies with direct comparisons of lactation curves of alped versus non-alped cows could further corroborate this conclusion. As expected, total milk production resulted globally higher for cows with alping occurring at the end of the lactation, since the drop in production happens later in the cycle. Anyway, winter calving might still be financially attractive for farmers since milk produced in the alps will have a higher economic value on the market and productivity will be higher during alping (Fig. 3).

Calving season also influences the way an animal is prompt to respond to environmental stress, with a greater impact of transhumance (i.e. greater *d*-parameter in absolute value) for cows calving in autumn and therefore alping at the end of their lactation cycle. Increased feed intake is known to have distinct effects on milk production depending on the lactation stage (Johnson 1984), and from what we observe it appears that milk production at the end of the lactation cycle is more sensitive to environmental changes. Similarly, when studying the effect of the considered factors, we showed that the between-group difference in milk production during alping is almost always greater for autumn calving.

5.2 Effect of the environment and climate change

Climate change requires species to adapt quickly to new and extreme climatic conditions (Hayes et al. 2009). In this context, cattle survival and annexed services for humans are threatened because of the low adaptive potential observed for international transboundary breeds (Taberlet et al. 2008). Holstein Fresian cattle for example has been shown to be quite sensitive to heat, particularly with THI values above 65 (Bryant et al. 2007). In Switzerland, climatic conditions are becoming hotter and dryer (IPCC 2014), which exhorts to better understand the effects of climate on cattle welfare and production both at lowland farms and during transhumance. Here, we observe a sensible negative effect of precipitation in spring (Fig 4e-f), probably because of their influence on forage growth (Jonas et al. 2008). Interestingly, heat waves (which are known to highly affect cattle productivity, Kadzere et al. 2002) were found to have minimal impact on milk

production during alping, probably because temperatures at high altitude rarely reach problematic thresholds. To further test this hypothesis, several thresholds were tested with values spanning from 63 up to 75: the impact of higher THI remains low (always inferior to 3%), but values obtained from high thresholds should be taken with care as too few observations are found in these ranges. Similarly, cold spells seem to have an almost negligible influence (Sup. Mat. S1). The observed effect of biogeographical regions on production can be explained by the difference in spring precipitation between such regions (158mm/month versus 98mm/month for the Northern flank and the Eastern part, respectively). Altitude confirmed its effect on productivity (Gorlier et al. 2013), being intrinsically connected with climatic conditions and vegetation type.

5.3 Effect of physiological and morphological factors

Lactation number has long been known to strongly influence milk production (Strucken et al. 2012), and this also holds for milk production during alping (Fig 4m-o). Even more important, pregnancy stage was found to have a significant impact on milk production during alping, especially when calving occurs in autumn (Fig. 4p-r). In order to optimise milk yield, cows are generally inseminated a few months after calving, to reach a time span of one year between lactation cycles, implying pregnancy stage not to be considered in lactation models to avoid strong collinearity with calving season (Tekerli et al. 2000; Wilmink 1987). However, correlation among these variables was not extreme in the present case (r^2=0.8), most likely because of unsuccessful inseminations leading some cows to delay pregnancy. These results must be interpreted with care, as cows with an early pregnancy are prone to fertility problems.

Many recent research efforts focused on increasing yield in cattle, leading to augmented cattle size (Tsuruta, Misztal, and Lawlor 2004) but disregarding important side-effects such as the loss of adaptive traits through genetic erosion (Notter 1999). This phenomenon might become deleterious for transhumance. For instance, despite showing higher productive performances even at alping, higher cows appear to be more impacted when moved to high mountain pastures (Fig. 4s-u). As for foot angle, steep angle is associated with a smaller risk of developing hoof diseases (Rogers 2002). Cows with steeper foot angle were observed to produce more milk both in lowland farm and during alping, but this factor appears to have limited impacts on the *d*-parameter (Fig. 4v-x).

As further analyses, it would be interesting to determine the impact of the Estimated Breeding Value (EBV) of the animal, as it is a commonly used measure in the selection of better-performing animal (Tsuruta et al. 2004). This would assess how higher ranked animals (i.e. exhibiting better performances under normal conditions) are affected by alping and therefore indicate if the current selection is beneficial or damaging to alped cows.

5.4 Limitations

Traditionally, lactation modelling is performed on an individual basis, and usually relies on daily or weekly milk records (Olori et al. 1999). Here, we based our work on a database composed of monthly milk records, which required the transformation of the data into daily averages over thousands of cows to avoid over-parameterisation in the model. This averaging might have diluted the strength of the effect we investigated.

Moreover, the proposed approach still misses validation, which could be achieved by relying on individual observations recorded daily or weekly and belonging to different breeds from the one used here.

Next, the amount of observations among calving months was not constant in the dataset, which possibly made the estimates from the winter months less robust. Further, a hidden age effect – as older cows tend to calf later in the season– could have biased the observed differences in milk productions among groups.

Last but not least, the model does not explicitly take into account cow feeding during alping, which is likely to affect milk production (Leiber et al. 2006). Indeed, the use of concentrate feeding varies among alps and among cows of the same alp. Particularly, differences in milk yield with different calving season could be globally influenced by varying concentrates feeding, with cows at an early stage in the lactation cycle – and thus producing a substantial amount of milk – potentially receiving more concentrates. In a similar context, other studies estimate a herd effect by evaluating the difference among farms, to consider (among other) different management strategies (see for example Gacula, Gaunt, and Damon 1968; Hayes et al. 2009; Tsuruta et al. 2004). In our case, this was not possible, as animals are held in hundreds of farms and are then brought to hundreds of different alps, and no distinction exists to group these farms or alps into two distinct groups as done for other factors, where we compared the first versus the third tertile.

6 Conclusion

Transhumance is a traditional farming practice which supports the preservation of both agricultural biodiversity and the socio-cultural heritage of human communities. Nevertheless, a loss in productivity is typically linked with alped livestock, which might discourage farmers from pursuing transhumance and poses its beneficial side-effects on ecosystems under threat. Here, we combined biological, geo-environmental and computer science tools to better understand the influence of environmental, physiological and morphological factors on milk productivity during transhumance. We relied on high resolution meteorological data and five million georeferenced monthly milk records as collected from over 200,000 Braunvieh cows in Switzerland. We show that both environmental and morphological factors have limited influence on animal production, with dry conditions in spring being nevertheless the most affecting environmental factor. This evidence suggests that animal production during transhumance might become even more insecure in future years due to climate change, and stresses therefore the urgency of devising strategies to protect this practice. However, the effects of environmental variables are small compared to the ones of physiological factors that have long been known to influence lactation performances (lactation number, pregnancy stage); these factors indeed strongly impact milk production throughout the whole lactation cycle, including during the alping period.

7 Supporting information

Sup. Mat. S1: Reported between-group difference for each calving month and each criterion.

Sup. Mat. S2: Evolution of protein percentage over a lactation cycle.

Sup. Mat. S3: Evolution of fat percentage over a lactation cycle.

Sup. Mat. S4: Effect of influencing factors on protein yield.

Sup. Mat. S5: Effect of influencing factors on fat yield.

8 Data Accessibility

The data was provided from the Braunvieh-CH association, under the explicit conditions that they will not be shared nor used for other studies. However, a partial dataset is available https://datadryad.org/stash/dataset/doi:10.5061/dryad.z612jm68g with the average milk production during alping from 20000 cows, together with lactation information (calving date, lactation number) and environmental data at the location of alping. Cows were chosen randomly, with equal number of animals per year, lactation number and calving month. Furthermore, researchers interested in performing studies on these data may contact directly the association (see contact information homepage.braunvieh.ch).

Relevant code for this research work is stored in GitHub: https://github.com/SolangeD/lactModel and has been archived within the Zenodo repository https://www.doi.org/10.5281/zenodo.3889931.

9 Author Contributions

SD performed most of the analyses with the help of AB. CF and SJ supervised the work. SD wrote the first draft of the article and all authors contributed in improving this draft.

10 Acknowledgments

We are grateful to the breeding organisation Braunvieh Schweiz for extracting and distributing the full data set from their database

TAKE HOME MESSAGE

- The transhumance system (called “alping” here) is a unique farming technique playing key roles in preserving ecological niches and the socio-cultural heritage.
- This farming technique is indirectly linked to the preservation of locally adapted breeds.
- Alping has a great impact on lactation, but is understudied
- In a context of climate change, it is essential to quantify this impact and understand which factors influence lactation. To this end, environmental information was retrieved for the days before each record.
- We propose a mathematical function to model the lactation curve of alped cow.
- This enables the study of the impact on lactation of different factors, including physiological, morphological and environmental factors.
- This study requires knowledge in biology, statistics, mathematical modelling and GIS as well as expertise in handling large phenotypical and environmental dataset and therefore illustrates a successful *biogeoinformatic* study.
- Environmental factors, have a significant yet limited impact when compared to physiological factors.
- Particularly problematic climatic conditions are dry condition during spring.

Chapter 5
General discussion

1 Groping towards *biogeoinformatics*

The scientific community often praises transdisciplinary approaches, and *biogeoinformatics* is one successful example where scientists from biology, geographic information and informatics have been working together for a long time. Indeed, the Econogene project (Ajmone-Marsan 2005), established as early as 2001, already illustrated how tools from *biogeoinformatics* could be used to address issues of sheep and goats in marginal regions. This integrative approach has become increasingly essential at a time when huge datasets are available, whether genomic or environmental, but also socio-economic or socio-demographic. People mastering the different dimensions of *biogeoinformatics* are indispensable to extract relevant information from this phenomenal amount of data in different fields, as in livestock science treated in this thesis.

For newcomers to *biogeoinformatics*, Leempoel *et al.* (2017) describes "simple rules for an efficient use of geographic information systems in molecular ecology", one of which being to systematically record the geographical coordinates of samples. Indeed, while coordinates can easily be collected and open the door to a wide range of analyses, both in wild and farm animal species, we notice that their recording are often neglected. In our case studies, precise and complete geographic information was only available when the data collection was specifically designed to study the influence of the environment, as it is the case in the Moroccan sheep population of the NEXTGEN project. In contrast, when data were extracted from databases storing routinely-collected information, the **location** of the sampling is often **inaccurate or** simply **absent**; in chapter four for example, one third of the alps were not precisely georeferenced (mainly small alps with reduced number of animals). From a general point of view, while current biology increasingly relies on genetics as well as computer science and advanced statistical methods (Roy, Pantanowitz, and Parwani 2014), it rarely does so in the integrative framework of *biogeoinformatics*. In this context, geographic coordinates are key features to this integration and are therefore utterly essential.

As highlighted in the second chapter, Early Warning System (EWS) for the monitoring of livestock species rarely accounts for geography, and if they do, they only consider the geographic concentration of monitored breeds (Alderson 2003). In the field of landscape genomics treated in the third chapter, our experience has shown that biologists are reluctant to use GIS software to retrieve environmental conditions to complement their molecular data, and that they face various difficulties, linked for example to the existence of different projection systems. Finally, a huge quantity of data is available to describe lactation curves of Swiss cows as studied in the fourth chapter, both in terms of milk production and climatic conditions. Yet, as far as we know, no publication relies on the match of these two types of information, in order to retrieve environmental conditions from the sampling location and at the recording time. While this matching step calls for efforts and skills to be accomplished, the knowledge revealed is nevertheless extremely valuable.

This ignorance to the geographic component can either arise from a lack of awareness about the opportunity offered by derived variables or simply from the complexity of the tasks involved in the analysis, especially for people without geographical background such as biologist or breeders.

Based on this observation and in order to facilitate a wider uptake of *biogeoinformatics* in livestock conservation, this thesis proposed several **case studies** to **illustrate** the benefit of such an approach as well as two **automated pipelines** to monitor breed diversity and perform landscape genomic studies.

Importantly though, the spatial dimension of *biogeoinformatics* studies can be more intensely exploited than what we have performed here, in particular by means of spatial statistics (Storfer et al. 2007). While this thesis demonstrates how simple information such as environmental and socio-economic variables can be extracted by means of georeferenced studies, one should keep in mind that the analyses from chapters two to four represent only a small fraction of what the spatial dimension can reveal.

2 Developed software and methods

Efforts to preserve local breeds typically focus on increasing the number of animals and on monitoring their genetic diversity. While these measures are undeniably useful, attention should also be paid to also characterise local breeds and preserve the production system in which they are involved, as stressed out by the Global Plan of Action. This thesis attempts to follow these recommendations and addresses several facets of the problem, which in most cases requires the use of *biogeoinformatics*. As discussed earlier, the expertise required by this integrated approach is rarely found, which prompted us to create software solutions for a wilder use of the method.

The monitoring of FAnGR is an essential step for the development of a prioritisation strategy designed to preserve local breeds. This led to the development of the **GENMON** application integrating a **multi-criteria approach** to measure the **level of endangerment** of livestock breeds, considering demographic, geographic and sustainability criteria. Acknowledging the interest of this approach, the Federal Office for Agriculture (FOAG) has issued a call for tenders to bring the GENMON application to a **production phase** and to routinely monitor all local breeds of the country. This demonstrates how political stakeholders can benefit from the input of *biogeoinformatic* analyses, provided appropriate tools are developed.

Similarly, to facilitate the use of *biogeoinformatics* in the field of **landscape genomics** in order to characterise genetic resources, we created an **integrated R pipeline** which follows the whole treatment chain from pre- to post-processing. It handles different input formats, all the way to the creation of maps and graphs, which traditionally involved a wide variety of software in GIS, biology and statistics. The positive feedbacks received from many users around the world show that *R.SamBada* does indeed meet the expectations of biologists. In particular, two points have been raised repeatedly: the gain of time that this pipeline provides together with the facilitated access to landscape genomic studies within one single software environment (R) and more particularly the effortless retrieval of climatic conditions at sampling location from open environmental databases.

The study on alped cows concentrated on the production system suited for local breeds and used *biogeoinformatics* in a different way, as it did not rely on the processing of genetic data. This study particularly well illustrated how large amount of data can be processed (i.e. milk records and meteorological data) to retrieve useful information, such as the impact of climate change. Such an extensive geospatial database required advanced database management skills to efficiently retrieve the necessary information. From a statistical point of view, it led to the formulation of a **mathematic function** to describe the **lactation curve of mountain-pastured cows**, that will hopefully be used in future studies. While no dedicated software was created, documented scripts

are available to facilitate the use of this approach. It is definitely a good illustration of how *biogeoinformatics* can be used in the context of milk records.

3 Possible future developments and studies

While we tried to do the deepest possible analyses, and to construct software solutions as complete as feasible within a restricted timeframe, several improvements can be made to all chapters of this thesis.

The GENMON monitoring tool for example can be further developed, in order to include other types of data, such as molecular data. This would enable the computation of a molecular-based inbreeding coefficient, thus allowing the processing of breeds with non-existent or incomplete pedigree. Other countries will hopefully follow the Swiss example and adapt the tool to their needs to fit their constraints and interests.

On a different note, the sustainability of the breeding practice included in GENMON could also be further refined with different parameters, in particular during the evaluation of the cultural value of the breed. When describing the process, we showed that simplistic questions are asked to the user (i.e. if the breed has an important cultural value and if this value has decreased over recent years, Table II.2). This definition is undeniably a little coarse and deserves better attention. One way of improving this estimation could be achieved through interviews with farmer families in which they should describe their attitude towards local breeds. Importantly this attitude should be compared across generations to inspect if the importance attached to local breeds has been transmitted to younger generations. Furthermore, the cultural value of the breed could be assessed with more concrete indicators, in particular the existence of typical farming products, especially labelled ones such as GPI or DOP. Indeed, the existence of such branded products, besides being economically favourable for the producer also enhances the visibility of the breed, which may increase the chance of long-term sustainability of local breed farming.

Last but not least, while GENMON is versatile and allows the user to easily change weights and thresholds to investigate their consequences on the global index, no thourough sensitivity analysis has been run, as done for example by Wainwright et al. (2019) in a similar context. Future analyses could include such types of analyses.

Improvement can also be made to the *R.SamBada* pipline. Indeed, in her comment about the article, Manel (2019) endorses the advantages that the package offers but also highlights that the "most urgent extension [...] to consider would be its integration with other existing gene-environment associations algorithms already implemented in R", such as LFMM2 (Caye et al. 2019). This would allow the comparison of different methods and their corresponding results in order to focus on genetic variations that are found significant with different methods. Moreover, several users also contacted me, asking for additional functionalities in the post-processing of non-model species. Indeed, because these species do not have a reference genome, the identification of genes and associated biological functions close to genetic variations of interest cannot be achieved directly. A BLAST analysis (Basic Local Alignment Search Tool; Johnson et al. 2008) will then be adopted to compare the sequence containing the genetic variation of interest with the genome of a model species, under the hypothesis that similar sequences will have the same biological functions in both the investigated and the chosen model species. This type of analysis would include a whole set of new functions and software that were beyond the scope of the pipeline, but would nonetheless be definitely helpful.

Beyond the identification of candidate loci with this pipeline, further studies are urgently needed to validate the findings of landscape genomic studies, in order to obtain a solid theoretical basis to include specific mutations in breeding programs. Care should indeed be taken when using non-validated models as the resulting analyses might lead to inaccurate conclusions, and in particular, actions resulting from such conclusions should be taken with circumspection. However, in landscape genomic studies, no damage can be caused to local breeds when preserving targeted genetic variations, even if these mutations happen to be relatively neutral with regards to local adaptation. Possible validation procedures could be based for example on field experiments that would study the physiological impacts of an environmental change on a set of locally adapted animals. In fact, controlled-climate experiments on farm animals have already been conducted (see for example Zimbleman et al. 2009), but to our knowledge, they have never been applied to validate landscape genomic studies. It should be noted that such an approach is particularly difficult to put into practice, as animals are not that easily displaced, especially because of veterinary travel restrictions. Alternatively, validating study can also rely on functional proteomic tools to analyse the effect of a significant mutation on the shape of the protein, provided the genetic variation is located within a gene. Here again, while proteomics has been completed in various contexts for farm animals (Wang et al. 2017; Zhao et al. 2014), landscape genomic studies rarely rely on this field to validate their findings.

As for the alping case study, several additional studies could reveal important aspects. Special attention should be given to low precipitation inputs during spring, a climatic conditions that has proven challenging for alped dairy cows. In particular, landscape genomics methods using *R.SamBada* could be applied here to highlight SNPs conferring a benefit to animals living in dry environments, analogous to what Hayes *et al.* (2009) performed with resistance to high temperature.

To conclude, all case studies of this thesis investigated different breeds from various countries and focused on different levels linked to the conservation of local breeds (the breed itself, its genetic variations and the production system associated to it). In an ideal situation, when attempting to preserve a breed, analyses should encompass all topics presented in the chapters of thesis. This was however not feasible in our case studies, none of them had the required input data for all types of analyses, these data being either inexistent or inaccessible.

4 Outcomes of case studies

The case studies of this thesis were diverse and had specific goals, all related to the broad question of FAnGR management: i) the monitoring of FAnGR ii) the identification of locally adapted genetic variations and iii) the study of the production system, more specifically the impact of climate change on such a system.

The first datasets uploaded in the GENMON application showed that the **level of endangerment** greatly **differs** among native species of Switzerland. It also showed that the type of problems as well as the problematic regions differed among breeds. Valais black-nose sheep for example suffer from inbreeding, while Franches-Montagnes Horse is mainly threatened by introgression, consistent with results found in the literature (Pirault et al. 2013; Signer-Hasler et al. 2019).

Using the specifically-designed *R.SamBada* pipeline, we were able to identify **SNPs** in Spanish bovine and Moroccan sheep **associated to high temperature and precipitation** respectively, located in or next to annotated genes handling biological functions related to the survival in such environmental conditions. In more detail, one significant SNP of the Spanish cattle dataset is

located near the *HSPB8* gene thought to encode a chaperone protein secreted under heat stress (Verma et al. 2016). In the case of the sheep dataset, two of the SNPs are missenses located in the *MC5R* gene, known to be associated with sebum secretion (Switonski, Mankowska, and Salamon 2013), and thus potentially indicating an adaptive mechanism to increased rainfall.

The study of factors influencing milk production of alped cows revealed that environmental conditions have a relatively small impact as compared with physiological factors. Indeed while lactation number and pregnancy stage can account for as much as 20% difference over the whole lactation, low environmental factors led to a change usually inferior to 10% of the milk production during alping. However, the most **problematic environmental change** is **drier conditions during spring**, whereas temperatures have very limited effect. In the context of climate change where we expect both higher temperatures and fewer precipitation, it is crucial to determine which problems can arise from these changes.

5 The future of FAnGR

As highlighted in this thesis, the erosion of livestock **genetic diversity** is and will remain a **major challenge** (Bruford et al. 2015). In the long run, we should remember that monitoring genetic diversity is only the first step to prevent the erosion of genetic diversity, as identified by FAO, as this should serve as a basis to **establish conservation policies**.

When studying the **transhumance system** from lowland to mountain pastures, we were surprised by the lack of literature on this subject, leading to the conclusion that a **better understanding** of its impact on cattle is **necessary**. This knowledge alone will not be sufficient to preserve this farming technique, but could at least contribute towards the elaboration of **guidelines** to assist breeders performing transhumance with their cattle. Furthermore, the mathematical model proposed could be helpful in the context of breeding programs in which alped cows are typically considered together with lowland ones, resulting in a limited performance of the former, which are to be attributed to the environmental conditions in which they graze rather than to poor individual performance. Under these circumstances, it is crucial to estimate the exact impact of alping

Climate change also remains a major **concern** for the future of FAnGR and it is essential to study its consequences on livestock. Thereupon, we showed in chapter 4 that high temperatures were of small concern for alped cows, whereas the most problematic change was the lack of precipitation during spring. When studying the impact of climate change on cattle, most of the existing literature focuses on the impact of temperature (Hayes et al. 2009; Raible and CH2014-Impacts Initiative 2014) and says little about the effect of reduced precipitation. High alpine pastures are indeed specific cases where dangerous temperature ranges are rarely found.

Once again, the variety of challenges that the livestock sector has to face emphasises the importance of treating the problem of FAnGR management as a whole and of considering all aspects associated to it within an integrated framework, as provided by *biogeoinformatics*.

6 Conclusion

This thesis is an attempt to foster the use of *biogeoinformatics* for assisting FAnGR in dealing with current challenges, such as the erosion of genetic diversity, the pressure on traditional framing techniques and climate change. As highlighted in Joost *et al.* (2016), *biogeoinformatics* can be useful

both for wild species and farm animal management, to integrate various data types representing information from different topics. We indeed demonstrated in this thesis how *biogeoinformatics* can help to monitor FAnGR, identify locally adapted genetic variations to be preserved and determine which environmental conditions can become problematic in the future. To facilitate the use of this kind of approach for biologists or breeders with no geographic background, we implemented two integrated tools: a WebGIS monitoring platform for FAnGR conservation and an R pipeline to perform landscape genomics studies. The benefits of *biogeoinformatics* are illustrated with case studies from different species in several countries. In this way, we identified two genetic mutations potentially linked to temperature and precipitation in Moroccan sheep and Spanish cattle respectively, and found that one of the most problematic climate change for alping is the reduced precipitation during spring. However, several barriers still exist for a wider usage of *biogeoinformatics*, the evidence being the restricted number of studies in this field. If there is to be only one recommendation at the end of this thesis, then it would be to always record the location of samples when performing bioinformatics studies. Indeed, as we have demonstrated here, the information that can be derived from georeferenced samples and subsequent *biogeoinformatic* studies offer very interesting insights, notably aspects linked to current challenges of the FAnGR sector.

References

Ajmone-Marsan, P. 2005. "Overview of Econogene, an European Project That Integrates Genetics, Socio-Economics and Geo-Statistics for the Sustainable Conservation of Sheep and Goat Genetic Resources." Pp. 1–8 in *The role of biotechnology for the characterization and conservation of crop, forestry, animal and fishery genetic resources. International Workshop, Turin, Italy, 5-7 March 2005*. Food and Agriculture Organization of the United Nations (FAO).

Alberto, Florian J., Frédéric Boyer, Pablo Orozco-terWengel, Ian Streeter, Bertrand Servin, Pierre de Villemereuil, Badr Benjelloun, Pablo Librado, Filippo Biscarini, Licia Colli, Mario Barbato, Wahid Zamani, Adriana Alberti, Stefan Engelen, Alessandra Stella, Stéphane Joost, Paolo Ajmone-Marsan, Riccardo Negrini, Ludovic Orlando, Hamid Reza Rezaei, Saeid Naderi, Laura Clarke, Paul Flicek, Patrick Wincker, Eric Coissac, James Kijas, Gwenola Tosser-Klopp, Abdelkader Chikhi, Michael W. Bruford, Pierre Taberlet, and François Pompanon. 2018. "Convergent Genomic Signatures of Domestication in Sheep and Goats." *Nature Communications* 9(1):813.

Alderson, L. 2003a. "Criteria for the Recognition and Prioritisation of Breeds of Special Genetic Importance." *Animal Genetic Resources* 33:1–9.

Alderson, L. 2003b. "Criteria for the Recognition and Prioritisation of Breeds of Special Genetic Importance." *Animal Genetic Resources* 33:1–9.

Alderson, L. 2009. "Breeds at Risk: Definition and Measurement of the Factors Which Determine Endangerment." *Livestock Science* 123(1):23–27.

Alderson, Lawrence. 2009. "Breeds at Risk: Definition and Measurement of the Factors Which Determine Endangerment." *Livestock Science* 123(1):23–27.

Alderson, Lawrence. 2010. "Breeds at Risk." Pp. 16–17 in *Criteria and classification. Report from a seminar*.

Alexander, D. H., J. Novembre, and K. Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19(9):1655–1664.

Aulchenko, Y. S., D. J. De Koning, and C. Haley. 2007. "Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method for Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis." *Genetics* 177(1):577–585.

Avon, L. 1992. "Survey about Small Breeds of Cattle, Sheep, Goats."

Balkenhol, Niko, Rachael Y. Dudaniec, Konstantin V. Krutovsky, Jeremy S. Johnson, David M. Cairns, Gernot Segelbacher, Kimberly A. Selkoe, Sophie von der Heyden, Ian J. Wang, and Oliver Selmoni. 2017. "Landscape Genomics: Understanding Relationships between Environmental Heterogeneity and Genomic Characteristics of Populations."

Barker, J. S. F. 1999. "Conservation of Livestock Breed Diversity." *Animal Genetic Resources/Resources Génétiques Animales/Recursos Genéticos Animales* 25:33–43.

del Barrio, Jose M. Garcia, Rafael Alonso Ponce, Raquel Benavides, and Sonia Roig. 2014. "Species Richness of Vascular Plants along the Climatic Range of the Spanish Dehesas at Two Spatial Scales." *Forest Systems* 23(1):111–119.

Bedward, Michael, David Eppstein, and Peter Menzel. 2018. *Packcircles: Circle Packing*.

Beissinger, Steven R., and Dale R. McCullough. 2002. *Population Viability Analysis*. University of Chicago Press.

Berry, N. R., F. Sutter, R. M. Bruckmaier, J. W. Blum, and M. Kreuzer. 2001. "Limitations of High Alpine Grazing Conditions for Early-Lactation Cows: Effects of Energy and Protein Supplementation." *Animal Science* 73(1):149–162.

Bertaglia, M., S. Joost, and J. Roosen. 2007. "Identifying European Marginal Areas in the Context of Local Sheep and Goat Breeds Conservation: A Geographic Information System Approach." *Agricultural Systems* 94(3):657–70.

Bertaglia, Marco, Stéphane Joost, and Jutta Roosen. 2007. "Identifying European Marginal Areas in the Context of Local Sheep and Goat Breeds Conservation: A Geographic Information System Approach." *Agricultural Systems* 94(3):657–70.

Biscarini, Filippo, Ezequiel L. Nicolazzi, Alessandra Stella, Paul J. Boettcher, and Gustavo Gandini. 2015. "Challenges and Opportunities in Genetic Improvement of Local Livestock Breeds." *Frontiers in Genetics* 6:33.

Bishop, S. C. 2012. "Possibilities to Breed for Resistance to Nematode Parasite Infections in Small Ruminants in Tropical Production Systems." *Animal* 6(5):741–47.

BLW. 2020. "Agrarbericht 2019 - Sömmerungsbetriebe." Retrieved February 13, 2020 (https://www.agrarbericht.ch/de/betrieb/strukturen/soemmerungsbetriebe).

BMELV. n.d. *Genetic Resources in Germany*. Bonn: Federal Ministry of Food, Agriculture and Consumer Protection.

Boettcher, P. J., I. Hoffmann, R. Baumung, A. G. Drucker, C. McManus, P. Berg, A. Stella, L. B. Nilsen, D. Moran, M. Naves, and others. 2014. "Genetic Resources and Genomics for Adaptation of Livestock to Climate Change." *Frontiers in Genetics* 5.

Boettcher, P. J., A. Stella, F. Pizzi, and G. Gandini. 2005. "The Combined Use of Embryos and Semen for Cryogenic Conservation of Mammalian Livestock Genetic Resources." *Genetics Selection Evolution* 37(7):1–19.

Boettcher, P. J., M. Tixier-Boichard, M. a. Toro, H. Simianer, H. Eding, G. Gandini, S. Joost, D. Garcia, L. Colli, P. Ajmone-Marsan, and and 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.

Boettcher, Paul J., Irene Hoffmann, Roswitha Baumung, Adam G. Drucker, Concepta McManus, Peer Berg, Alessandra Stella, Linn B. Nilsen, Dominic Moran, and Michel Naves. 2015. "Genetic Resources and Genomics for Adaptation of Livestock to Climate Change." *Frontiers in Genetics* 5:461.

Boletin Oficial del Estado. 2001. “Boletin Oficial del Estado. REAL DECRETO 60/2001, de 26 de enero, sobre prototipo racial de la raza bovina de lidia.” 5255–61.

Bonferroni, Carlo E., C. Bonferroni, and C. E. Bonferroni. 1936. “Teoria Statistica Delle Classi e Calcolo Delle Probabilita’.”

Bovolenta, Stefano, Walter Ventura, and Franco Malossini. 2002. “Dairy Cows Grazing an Alpine Pasture: Effect of Pattern of Supplement Allocation on Herbage Intake, Body Condition, Milk Yield and Coagulation Properties.” *Animal Research* 51(1):15–23.

Braunvieh Schweiz. n.d. “Braunvieh Schweiz.” Retrieved (homepage.braunvieh.ch).

Brodie, Aharon, Johnathan Roy Azaria, and Yanay Ofran. 2016. “How Far from the SNP May the Causative Genes Be?” *Nucleic Acids Research* 44(13):6046–6054.

Bruford, M. W., D. G. Bradley, and G. Luikart. 2003. “DNA Markers Reveal the Complexity of Livestock Domestication.” *Nature Reviews Genetics* 4(11):900–910.

Bruford, Michael W., Catarina Ginja, Irene Hoffmann, Stéphane Joost, Pablo Orozco-terWengel, Florian J. Alberto, Andreia J. Amaral, Mario Barbato, Filippo Biscarini, Licia Colli, and others. 2015. “Prospects and Challenges for the Conservation of Farm Animal Genomic Resources, 2015-2025.” *Frontiers in Genetics* 6.

Bryant, J. R., N. López-Villalobos, J. E. Pryce, C. W. Holmes, and D. L. Johnson. 2007. “Quantifying the Effect of Thermal Environment on Production Traits in Three Breeds of Dairy Cattle in New Zealand.” *New Zealand Journal of Agricultural Research* 50(3):327–338.

Bunce, R. G. H., M. Pérez-Soba, and M. Smith. 2009. “Assessment of the Extent of Agroforestry Systems in Europe and Their Role within Transhumance Systems.” Pp. 321–329 in *Agroforestry in Europe*. Springer.

Burren, A., H. Signer-Hasler, M. Neuditschko, J. Tetens, J. Kijas, C. Drögemüller, and C. Flury. 2014. “Fine-Scale Population Structure Analysis of Seven Local Swiss Sheep Breeds Using Genome-Wide SNP Data.” *Animal Genetic Resources/Ressources Génétiques Animales/Recursos Genéticos Animales* 55:67–76.

C2SM, FDHA, CHN, NCCR Climate, and OcCC. 2011. “Swiss Climate Change Scenarios CH2011.”

Caballero, A., and M. A. Toro. 2000. “Interrelations between Effective Population Size and Other Pedigree Tools for the Management of Conserved Populations.” *Genetical Research* 75(03):331–343.

Calus, M. P. L., Y. de Haas, and R. F. Veerkamp. 2013. “Combining Cow and Bull Reference Populations to Increase Accuracy of Genomic Prediction and Genome-Wide Association Studies.” *Journal of Dairy Science* 96(10):6703–6715.

Cañón, J., I. Tupac-Yupanqui, M. A. García-Atance, O. Cortés, D. Garcia, J. Fernández, and S. Dunner. 2008. “Genetic Variation within the Lidia Bovine Breed.” *Animal Genetics* 39(4):439–445.

Cassandro, M., A. Comin, M. Ojala, R. Dal Zotto, M. De Marchi, L. Gallo, P. Carnier, and G. Bittante. 2008. “Genetic Parameters of Milk Coagulation Properties and Their Relationships with Milk Yield and Quality Traits in Italian Holstein Cows.” *Journal of Dairy Science* 91(1):371–76.

Caye, Kevin, Basile Jumentier, Johanna Lepeule, and Olivier François. 2019. “LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies.” *Molecular Biology and Evolution*.

Cervantes, I., J. M. Pastor, J. P. Gutiérrez, F. Goyache, and A. Molina. 2011. “Computing Effective Population Size from Molecular Data: The Case of Three Rare Spanish Ruminant Populations.” *Livestock Science* 138(1):202–206.

Cesconeto, Robson Jose, Stéphane Joost, Concepta Margaret McManus, Samuel Rezende Paiva, Jaime Araujo Cobuci, and Jose Braccini. 2017. “Landscape Genomic Approach to Detect Selection Signatures in Locally Adapted Brazilian Swine Genetic Groups.” *Ecology and Evolution* 7(22):9544–9556.

Chang, Christopher C., Carson C. Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *Gigascience* 4(1):1.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPerson. 2018. *Shiny: Web Application Framework Fro R*.

Charlesworth, D. 2003. “Effects of Inbreeding on the Genetic Diversity of Populations.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358(1434):1051–1070.

Chen, H., N. Patterson, and D. Reich. 2010. “Population Differentiation as a Test for Selective Sweeps.” *Genome Research* 20(3):393–402.

Colli, Licia, Stéphane Joost, Riccardo Negrini, Letizia Nicoloso, Paola Crepaldi, Paolo Ajmone-Marsan, ECONOGENE Consortium, and others. 2014. “Assessing the Spatial Dependence of Adaptive Loci in 43 European and Western Asian Goat Breeds Using AFLP Markers.” *PloS One* 9(1):e86668.

Costa, E., Carlos A. Bana, and Jean-Claude Vansnick. 1994. “MACBETH—An Interactive Path towards the Construction of Cardinal Value Functions.” *International Transactions in Operational Research* 1(4):489–500.

Craine, J. M., A. J. Elmore, K. C. Olson, and D. Tolleson. 2010. “Climate Change and Cattle Nutritional Stress.” *Global Change Biology* 16(10):2901–2911.

Crawford, R. D. 1990. “Poultry Genetic Resources: Evolution, Diversity, and Conservation.” *Developments in Animal and Veterinary Sciences (Netherlands)*.

Cuervo-Alarcon, Laura, Matthias Arend, Markus Müller, Christoph Sperisen, Reiner Finkeldey, and Konstantin V. Krutovsky. 2018. “Genetic Variation and Signatures of Natural Selection in Populations of European Beech (Fagus Sylvatica L.) along Precipitation Gradients.” *Tree Genetics & Genomes* 14(6):84.

Cunningham, E. P., J. J. Dooley, R. K. Splan, and D. G. Bradley. 2001. “Microsatellite Diversity, Pedigree Relatedness and the Contributions of Founder Lineages to Thoroughbred Horses.” *Animal Genetics* 32(6):360–364.

Curry, Mark R. 2000. “Cryopreservation of Semen from Domestic Livestock.” *Reviews of Reproduction* 5(1):46–52.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, and Stephen T. Sherry. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27(15):2156–2158.

Dikmen, S., F. A. Khan, H. J. Huson, T. S. Sonstegard, J. I. Moss, G. E. Dahl, and P. J. Hansen. 2014. "The SLICK Hair Locus Derived from Senepol Cattle Confers Thermotolerance to Intensively Managed Lactating Holstein Cows." *Journal of Dairy Science* 97(9):5508–20.

Edea, Z., H. Dadi, S. W. Kim, J. H. Park, G. H. Shin, Tadelle Dessie, and K. S. Kim. 2014. "Linkage Disequilibrium and Genomic Scan to Detect Selective Loci in Cattle Populations Adapted to Different Ecological Conditions in E Thiopia." *Journal of Animal Breeding and Genetics* 131(5):358–366.

Ertz, O., S. J. Rey, and S. Joost. 2014. "The Open Source Dynamics in Geospatial Research and Education." *Journal of Spatial Information Science* 2014(8):67–71.

Eusebi, P. G., O. Cortés, S. Dunner, and J. Cañón. 2017. "Genomic Diversity and Population Structure of Mexican and Spanish Bovine Lidia Breed." *Animal Genetics* 48(6):682–685.

Fabbri, Maria Chiara, Marcos Paulo Gonçalves de Rezende, Christos Dadousis, Stefano Biffani, Riccardo Negrini, Paulo Luiz Souza Carneiro, and Riccardo Bozzi. 2019. "Population Structure and Genetic Diversity of Italian Beef Breeds as a Tool for Planning Conservation and Selection Strategies." *Animals* 9(11):880.

Fadlaoui, Aziz, Jutta Roosen, and Philippe V. Baret. 2006. "Setting Priorities in Farm Animal Conservation Choices—Expert Opinion and Revealed Policy Preferences." *European Review of Agricultural Economics* 33(2):173–92.

FAO. 2007a. *Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration*. Rome: Commission on genetic resources for food and agriculture, FAO.

FAO. 2007b. *Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration*. Rome: Commission on genetic resources for food and agriculture, FAO.

FAO. 2007c. *The State of the World's Animal Genetic Resources for Food and Agriculture*. Food & Agriculture Org.

FAO. 2012. *Cryoconservation of Animal Genetic Resources*. Rome.

FAO. 2015. *The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture*. Food & Agriculture Org.

Farr, T. G., A. R. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, and L. Roth. 2007. "' The Shuttle Radar Topography Mission.' Reviews of Geophysics 45." *RG2004, Doi* 10.

Fautin, Daphne Gail, and Robert W. Buddemeier. 2001. *Biogeoinformatics of Hexacorallia (Corals, Sea Anemones, and Their Allies): Interfacing Geospatial, Taxonomic, and Environmental Data for a Group of Marine Invertebrates*. Citeseer.

FM-CH. n.d. "Fédération Suisse Du Franches-Montagnes." Retrieved (www.fm-ch.ch).

FOAG. 2002. *Les Ressources Génétiques En Suisse*. Federal Office for Agriculture, FOAG.

FOEN. 2001. *Die Biogeographischen Regionen Der Schweiz*. UM-137-D. Federal Office for the Environment (FOEN).

Fribourg Région. n.d. "Désalpe | La fête traditionnelle de montagne en sept. & oct." *Fribourg Région*. Retrieved April 10, 2020 (https://www.fribourgregion.ch/fr/Z9701).

Frichot, E., S. D. Schoville, G. Bouchard, and O. François. 2013. "Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models." *Molecular Biology and Evolution* 30(7):1687–1699.

FSO. 2019. *Food and Agriculture, Pocket Statistics*. Neuchâtel: Federal Statistical Office.

Gandini, G. C., L. Ollivier, B. Danell, O. Distl, A. Georgoudis, E. Groeneveld, E. Martyniuk, J. A. M. van Arendonk, and J. A. Woolliams. 2004. "Criteria to Assess the Degree of Endangerment of Livestock Breeds in Europe." *Livestock Production Science* 91(1–2):173–82.

Gandini, Gustavo C., and Emanuele Villa. 2003. "Analysis of the Cultural Value of Local Livestock Breeds: A Methodology." *Journal of Animal Breeding and Genetics* 120(1):1–11.

Gaston, K. J. 1991. "How Large Is a Species' Geographic Range?" *Oikos* 434–438.

Gellrich, Mario, Priska Baur, Barbara Koch, and Niklaus E. Zimmermann. 2007. "Agricultural Land Abandonment and Natural Forest Re-Growth in the Swiss Mountains: A Spatially Explicit Economic Analysis." *Agriculture, Ecosystems & Environment* 118(1):93–108.

Gellrich, Mario, and Niklaus E. Zimmermann. 2007. "Investigating the Regional-Scale Pattern of Agricultural Land Abandonment in the Swiss Mountains: A Spatial Statistical Modelling Approach." *Landscape and Urban Planning* 79(1):65–76.

GeoServer. n.d. *GeoServer*.

Gibson, J. P., and S. C. Bishop. 2005. "Use of Molecular Markers to Enhance Resistance of Livestock to Disease: A Global Approach." *Revue Scientifique Et Technique-Office International Des Epizooties* 24(1):343.

Gicquel, E., P. Boettcher, B. Besbes, S. Furre, J. Fernández, C. Danchin-Burge, B. Berger, R. Baumung, J. R. J. Feijóo, and G. Leroy. 2019. "Impact of Conservation Measures on Demography and Genetic Variability of Livestock Breeds." *Animal* 1–11.

Gilgen, A. K., and N. Buchmann. 2009. "Response of Temperate Grasslands at Different Altitudes to Simulated Summer Drought Differed but Scaled with Annual Precipitation." *Biogeosciences* 6(11):2525–2539.

Giuffra, EJMH, J. M. H. Kijas, V. Amarger, Ö. Carlborg, J. T. Jeon, and L. Andersson. 2000. "The Origin of the Domestic Pig: Independent Domestication and Subsequent Introgression." *Genetics* 154(4):1785–1791.

Glowatzki-Mullis, M. L., J. Muntwyler, E. Bäumle, and C. Gaillard. 2009. "Genetic Diversity of Swiss Sheep Breeds in the Focus of Conservation Research." *Journal of Animal Breeding and Genetics* 126(2):164–75.

Glowatzki-Mullis, M. L., J. Muntwyler, W. Pfister, E. Marti, S. Rieder, P. A. Poncet, and C. Gaillard. 2006. "Genetic Diversity among Horse Populations with a Special Focus on the Franches-Montagnes Breed." *Animal Genetics* 37(1):33–39.

Gorlier, A., M. Lonati, M. Renna, C. Lussiana, G. Lombardi, and L. M. Battaglini. 2013. "Changes in Pasture and Cow Milk Compositions during a Summer Transhumance in the Western Italian Alps." *Journal of Applied Botany and Food Quality* 85(2):216.

Groeneveld, E., Z. I. Duchev, M. Imialek, L. Soltys, M. Wieczorek, B. Scherf, O. Distl, G. Gandini, M. Jaszczynska, and A. Rosati. 2007. “FABISnet-A Web Based Network of Farm Animal Biodiversity Information Systems.” *Proc. GIL Jahrestagung* 91–94.

Groeneveld, E., B. D. Westhuizen, A. Maiwashe, F. Voordewind, and J. B. S. Ferraz. 2009. “POPREP: A Generic Report for Population Management.” *Genetics and Molecular Research* 8(3):1158–1178.

Groeneveld, E., B. D. Westhuizen, A. Maiwashe, F. Voordewind, J. B. S. Ferraz, and others. 2009. “POPREP: A Generic Report for Population Management.” *Genetics and Molecular Research* 8(3):1158–1178.

Guessous, F., I. Boujenane, M. Bourfia, and H. Narjisse. 1989. “Sheep in Morocco.” *FAO Animal Production and Health Paper (FAO)*.

Günther, Torsten, and Graham Coop. 2013. “Robust Identification of Local Adaptation from Allele Frequencies.” *Genetics* 195(1):205–220.

Gutiérrez, J. P., I. Cervantes, A. Molina, M. Valera, and F. Goyache. 2008. “Individual Increase in Inbreeding Allows Estimating Effective Sizes from Pedigrees.” *Genet. Sel. Evol* 40:359–378.

Gutiérrez, J. P., and F. Goyache. 2005. “A Note on ENDOG: A Computer Program for Analysing Pedigree Information.” *Journal of Animal Breeding and Genetics* 122(3):172–176.

Gutiérrez, Juan Pablo, Isabel Cervantes, Antonio Molina, Mercedes Valera, and Félix Goyache. 2008. “Individual Increase in Inbreeding Allows Estimating Effective Sizes from Pedigrees.” *Genet. Sel. Evol* 40:359–378.

Gutiérrez, Juan Pablo, and Félix Goyache. 2005. “A Note on ENDOG: A Computer Program for Analysing Pedigree Information.” *Journal of Animal Breeding and Genetics* 122(3):172–176.

Hahn, G. L. 1999. “Dynamic Responses of Cattle to Thermal Heat Loads.” *Journal of Animal Science* 77:10.

Hanotte, O., Y. Ronin, Morris Agaba, P. Nilsson, A. Gelhaus, R. Horstmann, Y. Sugimoto, S. Kemp, J. Gibson, and A. Korol. 2003. “Mapping of Quantitative Trait Loci Controlling Trypanotolerance in a Cross of Tolerant West African N’Dama and Susceptible East African Boran Cattle.” *Proceedings of the National Academy of Sciences* 100(13):7443–7448.

Hayes, B. 2013. “Overview of Statistical Methods for Genome-Wide Association Studies (GWAS).” *Genome-Wide Association Studies and Genomic Prediction* 149–169.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, K. Savin, C. P. Van Tassell, T. S. Sonstegard, and M. E. Goddard. 2009. “A Validated Genome Wide Association Study to Breed Cattle Adapted to an Environment Altered by Climate Change.” *PLoS One* 4(8):e6676.

Hayes, B. J., M. Carrick, P. Bowman, and M. E. Goddard. 2003. “Genotype$\times$ Environment Interaction for Milk Production of Daughters of Australian Dairy Sires from Test-Day Records.” *Journal of Dairy Science* 86(11):3736–3744.

Hedrick, Philip. 2011. *Genetics of Populations*. Jones & Bartlett Learning.

Herrero, Mario, Petr Havlík, Hugo Valin, An Notenbaert, Mariana C. Rufino, Philip K. Thornton, Michael Blümmel, Franz Weiss, Delia Grace, and Michael Obersteiner. 2013. “Biomass Use, Production, Feed Efficiencies, and Greenhouse Gas Emissions from Global Livestock Systems.” *Proceedings of the National Academy of Sciences* 110(52):20888–20893.

Herzog, F., R. Böni, S. Lauber, M. Schneider, and I. Seidl. 2009. “AlpFUTUR–an Inter-and Transdisciplinary Research Program on the Future of Summer Pastures in Switzerland.” Pp. 53–54 in *Proceedings of 15th meeting of the FAO-CIHEAM Mountain Pastures Network. AgroscopeChangins-Wadenswil Research Station ACW, Switzerland*.

Herzog, F., R. GH Bunce, M. Pérez-Soba, R. HG Jongman, A. Gómez Sal, and I. Austad. 2005. “Policy Options to Support Transhumance and Biodiversity in European Mountains: A Report on the TRANSHUMOUNT Stakeholder Workshop, Landquart/Zurich, Switzerland, 26-28 May 2004.” *Mountain Research and Development* 25(1):82–84.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2004. *The WorldClim Interpolated Global Terrestrial Climate Surfaces. Version 1.3*.

Horn, Marco, Andreas Steinwidder, Johann Gasteiner, Leopold Podstatzky, Alfred Haiger, and Werner Zollitsch. 2013. “Suitability of Different Dairy Cow Types for an Alpine Organic and Low-Input Milk Production System.” *Livestock Science* 153(1–3):135–146.

Horn, Marco, Andreas Steinwidder, Walter Starz, Rupert Pfister, and Werner Zollitsch. 2014. “Interactions between Calving Season and Cattle Breed in a Seasonal Alpine Organic and Low-Input Dairy System.” *Livestock Science* 160:141–50.

Huang, Ivy B., Jeffrey Keisler, and Igor Linkov. 2011. “Multi-Criteria Decision Analysis in Environmental Sciences: Ten Years of Applications and Trends.” *Science of The Total Environment* 409(19):3578–94.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, and others. 2002. “The Ensembl Genome Database Project.” *Nucleic Acids Research* 30(1):38–41.

ICAR. 2014. “ICAR Recording Guidelines Approved by the General Assembly Held in Berlin, Germany, on May 2014.”

IPCC. 2014. “Climate Change 2014 Synthesis Report; A Report of the Intergovernmental Panel on Climate Change.”

Jeretina, J., D. Babnik, and D. Skorjanc. 2013. “Modeling Lactation Curve Standards for Test-Day Milk Yield in Holstein, Brown Swiss and Simmental Cows.” *The J Anim and Plant Sci* 233:754–62.

Johnson, C. L. 1984. “The Effect of Feeding in Early Lactation on Feed Intake, Yields of Milk, Fat and Protein and on Live-Weight Change over One Lactation Cycle in Dairy Cows.” *The Journal of Agricultural Science* 103(3):629–637.

Johnson, Mark, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L. Madden. 2008. “NCBI BLAST: A Better Web Interface.” *Nucleic Acids Research* 36(suppl_2):W5–W9.

Jonas, Tobias, Christian Rixen, Matthew Sturm, and Veronika Stoeckli. 2008. “How Alpine Plant Growth Is Linked to Snow Cover and Climate Variability.” *Journal of Geophysical Research: Biogeosciences* 113(G3).

Joost, S. 2014. “Biogeoinformatics of Livestock Genomic Resources.” in *Livestock genomic resources in a changing world*.

Joost, S., A. Bonin, M. W. Bruford, L. Després, C. Conord, G. Erhardt, and P. Taberlet. 2007. “A Spatial Analysis Method (SAM) to Detect Candidate Loci for Selection: Towards a Landscape Genomics Approach to Adaptation.” *Molecular Ecology* 16(18):3955–69.

Joost, S., M. W. Bruford, Genomic-Resources Consortium, and others. 2015. “Editorial: Advances in Farm Animal Genomic Resources.” *Frontiers in Genetics* 6.

Joost, S., L. Colli, P. V. Baret, J. F. Garcia, P. J. Boettcher, M. Tixier-Boichard, P. Ajmone-Marsan, and The GLOBALDIV Consortium. 2010. “Integrating Geo-Referenced Multiscale and Multidisciplinary Data for the Management of Biodiversity in Livestock Genetic Resources.” *Animal Genetics* 41:47–63.

Joost, Stéphane, Solange Duruz, Estelle Rochat, and Ivo Widmer. 2016. *Open Computational Landscape Genetics*. PeerJ Preprints.

Jõudu, Ivi, Merike Henno, Tanel Kaart, Tõnu Püssa, and Olav Kärt. 2008. “The Effect of Milk Protein Contents on the Rennet Coagulation Properties of Milk from Individual Dairy Cows.” *International Dairy Journal* 18(9):964–67.

Jurt, C., I. Häberli, and R. Rossier. 2015. “Transhumance Farming in Swiss Mountains: Adaptation to a Changing Environment.” *Mountain Research and Development* 35(1):57–65.

Kadzere, C. T., M. R. Murphy, N. Silanikove, and E. Maltz. 2002. “Heat Stress in Lactating Dairy Cows: A Review.” *Livestock Production Science* 77(1):59–91.

Kawecki, Tadeusz J., and Dieter Ebert. 2004. “Conceptual Issues in Local Adaptation.” *Ecology Letters* 7(12):1225–1241.

Keeney, Ralph L., and Eric F. Wood. 1977. “An Illustrative Example of the Use of Multiattribute Utility Theory for Water Resource Planning.” *Water Resources Research* 13(4):705–712.

Kim, Eui-Soo, and Max F. Rothschild. 2014. “Genomic Adaptation of Admixed Dairy Cattle in East Africa.” *Frontiers in Genetics* 5:443.

Kohler-Rollefson, Ilse. 2004. “Farm Animal Genetic Resources: Safeguarding National Assets for Food Security and Trade.”

Lauber, S. 2013. “Avenir de l’économie Alpestre Suisse.” *Faits, Analyses et Éléments de Réflexion Issus Du Programme de Recherche AlpFUTUR. Institut Fédéral de Recherche Sur La Forêt, La Neige et Le Paysage (WSL), Birmensdorf*.

Lee, Chih, Ali Abdool, and Chun-Hsi Huang. 2009. “PCA-Based Population Structure Inference with Generic Clustering Algorithms.” *BMC Bioinformatics* 10(1):S73.

Leempoel, Kevin, Solange Duruz, Estelle Rochat, Ivo Widmer, Pablo Orozco-terWengel, and Stéphane Joost. 2017. “Simple Rules for an Efficient Use of Geographic Information Systems in Molecular Ecology.” *Frontiers in Ecology and Evolution* 5:33.

Leiber, Florian, Michael Kreuzer, Hans Leuenberger, and Hans-Rudolf Wettstein. 2006. “Contribution of Diet Type and Pasture Conditions to the Influence of High Altitude Grazing on Intake, Performance and Composition and Renneting Properties of the Milk of Cows.” *Animal Research* 55(1):37–53.

Liechti, K., and J. P. Biber. 2016. “Pastoralism in Europe: Characteristics and Challenges of Highland–Lowland Transhumance: -EN- -FR- Le Pastoralisme En Europe : Caractéristiques et Défis de La Transhumance de La Montagne Vers La Plaine -ES- El Pastoreo En Europa: Características y Problemas de La Trashumancia de Tierras Altas-Tierras Bajas.” *Revue Scientifique et Technique de l’OIE* 35(2):561–75.

Loftus, R., and B. Scherf. 1993. *World Watch List for Domestic Animal Diversity*. Rome: Food and Agriculture Organization.

Lucek, Kay, Irene Keller, Arne W. Nolte, and Ole Seehausen. 2018. “Distinct Colonization Waves Underlie the Diversification of the Freshwater Sculpin (Cottus Gobio) in the Central European Alpine Region.” *Journal of Evolutionary Biology*.

Luo, W., M. C. Taylor, and S. R. Parker. 2008. “A Comparison of Spatial Interpolation Methods to Estimate Continuous Wind Speed Surfaces Using Irregularly Distributed Data from England and Wales.” *International Journal of Climatology* 28(7):947–959.

Lv, Feng-Hua, Saif Agha, Juha Kantanen, Licia Colli, Sylvie Stucki, James W. Kijas, Stéphane Joost, Meng-Hua Li, and Paolo Ajmone Marsan. 2014. “Adaptations to Climate-Mediated Selective Pressures in Sheep.” *Molecular Biology and Evolution* msu264.

Macciotta, Nicolò Pietro Paolo, Daniele Vicario, and Aldo Cappio-Borlino. 2005. “Detection of Different Shapes of Lactation Curve for Milk Yield in Dairy Cattle by Empirical Mathematical Models.” *Journal of Dairy Science* 88(3):1178–1191.

Mack, Gabriele, Christian Flury, and others. 2008. “Wirkung Der Sömmerungsbeiträge.” *Agrarforschung (Switzerland)*.

Magarey, R. D., G. A. Fowler, D. M. Borchert, T. B. Sutton, M. Colunga-Garcia, and J. A. Simpson. 2007. “NAPPFAST: An Internet System for the Weather-Based Mapping of Plant Pathogens.” *Plant Disease* 91(4):336–345.

Manel, S., M. K. Schwartz, G. Luikart, and P. Taberlet. 2003. “Landscape Genetics: Combining Landscape Ecology and Population Genetics.” *Trends in Ecology & Evolution* 18(4):189–197.

Manel, Stéphanie. 2019. “Smoothing Technical and Computational Obstacles in Gene-Environment Associations.” *Molecular Ecology Resources* 19(6):1385–87.

Manel, Stephanie, Stephane Joost, Bryan K. Epperson, Rolf Holderegger, Andrew Storfer, Michael S. Rosenberg, Kim T. Scribner, Aurelie Bonin, and MARIE-JOSÉE FORTIN. 2010. “Perspectives on the Use of Landscape Genetics to Detect Genetic Adaptive Variation in the Field.” *Molecular Ecology* 19(17):3760–3772.

Martinez-Alier, Joan, Giuseppe Munda, and John O’Neill. 1999. “Commensurability and Compensability in Ecological Economics.” *Valuation and the Environment: Theory, Method and Practice* 37–57.

Masterman, A. J., G. N. Foster, S. J. Holmes, and R. Harrington. 1996. “The Use of the Lamb Daily Weather Types and the Indices of Progressiveness, Southerliness and Cyclonicity to Investigate the Autumn Migration of Rhopalosiphum Padi.” *Journal of Applied Ecology* 23–30.

Meteoswiss. 2018. “The Climate of Switzerland.”

Meteoswiss. n.d. “MeteoSwiss Grid-Data Products.”

Meuwissen, T. 2009. “Genetic Management of Small Populations: A Review.” *Acta Agriculturae Scand Section A* 59(2):71–79.

Microsoft Corporation, and Steve Weston. 2017. *DoParallel: Foreach Parallel Adaptor for the “parallel” Package*.

Milano, Marianne, Emmanuel Reynard, Nina Köplin, and Rolf Weingartner. 2015. "Climatic and Anthropogenic Changes in Western Switzerland: Impacts on Water Stress." *Science of the Total Environment* 536:12–24.
Montaldo, Hugo H. 2001. "Genotype by Environment Interactions in Livestock Breeding Programs: A Review." *Interciencia* 26(6):229–235.
Mwai, Okeyo, Olivier Hanotte, Young-Jun Kwon, and Seoae Cho. 2015. "African Indigenous Cattle: Unique Genetic Resources in a Rapidly Changing World." *Asian-Australasian Journal of Animal Sciences* 28(7):911.
Nakicenovic, Nebojsa, Joseph Alcamo, Gerald Davis, Bert De Vries, Joergen Fenhann, Stuart Gaffin, Kermeth Gregory, Amulf Griibler, Tae Yong Jung, Tom Kram, and others. 2000. "Emissions Scenarios."
Newesely, Christian, Erich Tasser, Peter Spadinger, and Alexander Cernusca. 2000. "Effects of Land-Use Changes on Snow Gliding Processes in Alpine Ecosystems." *Basic and Applied Ecology* 1(1):61–67.
Notter, D. R. 1999. "The Importance of Genetic Diversity in Livestock Populations of the Future." *Journal of Animal Science* 77(1):61–69.
OcCC. 2007. "Climate Change and Swizterland 2050. Expected Impats on Environment, Society and Economy."
OFS. 2013. *Actualité OFS: De l'herbe Au Lait. La Production de Lait En Suisse*.
Olea, Pedro P., and Patricia Mateo-Tomás. 2009. "The Role of Traditional Farming Practices in Ecosystem Conservation: The Case of Transhumance and Vultures." *Biological Conservation* 142(8):1844–1853.
Olori, V. E., S. Brotherstone, W. G. Hill, and B. J. McGuirk. 1997. "Effect of Gestation Stage on Milk Yield and Composition in Holstein Friesian Dairy Cattle." *Livestock Production Science* 52(2):167–176.
Olori, V. E., S. Brotherstone, W. G. Hill, and B. J. McGuirk. 1999. "Fit of Standard Models of the Lactation Curve to Weekly Records of Milk Production of Cows in a Single Herd." *Livestock Production Science* 58(1):55–63.
Openlayers. n.d. *Openlayers*.
Orozco-terWengel, Pablo, Mario Barbato, Ezequiel Nicolazzi, Filippo Biscarini, Marco Milanesi, Wyn Davies, Don Williams, Alessandra Stella, Paolo Ajmone-Marsan, and Michael W. Bruford. 2015. "Revisiting Demographic Processes in Cattle with Genome-Wide Population Genetic Analysis." *Frontiers in Genetics* 6:191.
Patil, Ganapati P., K. Sham Bhat, and S. W. Joshi. 2007. "Surveillance Geoinformatics of Hotspot Detection and Prioritization for Monitoring, Etiology, Early Warning and Sustainable Management." Pp. 302–303 in *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*. Digital Government Society of North America.
Pereira, Paulo, José Teixeira, and Guillermo Velo-Antón. 2018. "Allele Surfing Shaped the Genetic Structure of the European Pond Turtle via Colonization and Population Expansion across the Iberian Peninsula from Africa." *Journal of Biogeography* 45(9):2202–2215.
Pirault, Pauline, Sophy Danvy, Etienne Verrier, and Grégoire Leroy. 2013. "Genetic Structure and Gene Flows within Horses: A Genealogical Study at the French Population Scale." *PloS One* 8(4).
Price, Bronwyn, Felix Kienast, Irmi Seidl, Christian Ginzler, Peter H. Verburg, and Janine Bolliger. 2015. "Future Landscapes of Switzerland: Risk Areas for Urbanisation and Land Abandonment." *Applied Geography* 57:32–41.
R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Raible, C. C., and CH2014-Impacts Initiative, eds. 2014. *CH2014-Impacts: Toward Quantitative Scenarios of Climate Change Impacts in Switzerland*. Bern: Universität Bern, Oeschger Centre for Climate Change Research (OCCR).
Rasul, Golam, and Gopal B. Thapa. 2004. "Sustainability of Ecological and Conventional Agricultural Systems in Bangladesh: An Assessment Based on Environmental, Economic and Social Perspectives." *Agricultural Systems* 79(3):327–351.
Reist-Marti, S. B., H. Simianer, J. Gibson, O. Hanotte, and J. E. O. Rege. 2003. "Weitzman's Approach and Conservation of Breed Diversity: An Application to African Cattle Breeds." *Conservation Biology* 17(5):1299–1311.
Rellstab, Christian, Felix Gugerli, Andrew J. Eckert, Angela M. Hancock, and Rolf Holderegger. 2015. "A Practical Guide to Environmental Association Analysis in Landscape Genomics." *Molecular Ecology* 24(17):4348–4370.
Roberts, D. S. 1963. "Barriers to Dermatopilus Dermatonomus Infection on the Skin of Sheep." *Australian Journal of Agricultural Research* 14(4):492–508.
Rogers, G. W. 2002. "ASPECTS OF MILK COMPOSITION, PRODUCTIVE LIFE AND TYPE TRAITS IN RELATION TO MASTITIS AND OTHER DISEASES IN DAIRY CATTLE." 7.
Rousset, François. 2002. "Inbreeding and Relatedness Coefficients: What Do They Measure?" *Heredity* 88(5):371–380.
Roy, Bernard. 1991. "The Outranking Approach and the Foundations of ELECTRE Methods." *Theory and Decision* 31(1):49–73.
Roy, Bernard, and Philippe Vincke. 1981. "Multicriteria Analysis: Survey and New Directions." *European Journal of Operational Research* 8(3):207–218.
Roy, Somak, Liron Pantanowitz, and Anil V. Parwani. 2014. "Bioinformatics." Pp. 175–180 in *Practical Informatics for Cytopathology*. Springer.
Ruane, John. 2000. "A Framework for Prioritizing Domestic Animal Breeds for Conservation Purposes at the National Level: A Norwegian Case Study." *Conservation Biology* 14(5):1385–93.
Saaty, Thomas L. 1990. "How to Make a Decision: The Analytic Hierarchy Process." *European Journal of Operational Research* 48(1):9–26.
Sargolzaei, M., H. Iwaisaki, J. J. Colleau, and others. 2006. "CFC: A Tool for Monitoring Genetic Diversity." Pp. 27–28 in *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil, 13-18 August, 2006.* Instituto Prociência.
Sarkar, Sahotra, Robert L. Pressey, Daniel P. Faith, Christopher R. Margules, Trevon Fuller, David M. Stoms, Alexander Moffett, Kerrie A. Wilson, Kristen J. Williams, Paul H. Williams, and Sandy Andelman. 2006. "Biodiversity Conservation Planning Tools: Present Status and Challenges for the Future." *Annual Review of Environment and Resources* 31(1):123–59.

Schefers, Jonathan M., and Kent A. Weigel. 2012. “Genomic Selection in Dairy Cattle: Integration of DNA Testing into Breeding Programs.” *Animal Frontiers* 2(1):4–9.
Schweiz Tourismus. n.d. “Fromage d’alpage - Plaisir gastronomique venu des alpes.” *Suisse Tourisme*. Retrieved April 10, 2020 (https://www.myswitzerland.com/fr-ch/planification/vie-pratique/coutumes-et-traditions/fromage-dalpage-plaisir-gastronomique-venu-des-alpes/).
Schweizeralpkaese. n.d. “Variétés de fromage d’alpage.” *Alpkäse*. Retrieved April 10, 2020 (https://www.schweizeralpkaese.ch/fr/varietes-de-fromage-dalpage/).
Sengar, Gyanendra Singh, Rajib Deb, Umesh Singh, T. V. Raja, Rajiv Kant, Basavraj Sajjanar, Rani Alex, R. R. Alyethodi, Ashish Kumar, and Sushil Kumar. 2018. “Differential Expression of MicroRNAs Associated with Thermal Stress in Frieswal (Bos Taurus x Bos Indicus) Crossbred Dairy Cattle.” *Cell Stress and Chaperones* 23(1):155–170.
Shih, Kai-Ming, Chung-Te Chang, Jeng-Der Chung, Yu-Chung Chiang, and Shih-Ying Hwang. 2018. “Adaptive Genetic Divergence despite Significant Isolation-by-Distance in Populations of Taiwan Cow-Tail Fir (Keteleeria Davidiana Var. Formosana).” *Frontiers in Plant Science* 9:92.
Signarbieux, Constant, and Urs Feller. 2008. “Effects of an Extended Drought Period on Grasslands at Various Altitudes in Switzerland: A Field Study.” Pp. 1371–1374 in *Photosynthesis. Energy from the Sun*. Springer.
Signer-Hasler, Heidi, Alexander Burren, Philippe Ammann, Cord Droegemueller, and Christine Flury. 2019. “Extent of Genomic Inbreeding in Swiss Sheep and Goat Breeds.” *AGRARFORSCHUNG SCHWEIZ* 10(10):372–379.
Simon, D. L., and D. Buchenauer. 1993. “Genetic Diversity of European Livestock Breeds.” iv + 581 pp.
de Simoni Gouveia, João José, Samuel Rezende Paiva, Concepta M. McManus, Alexandre Rodrigues Caetano, James W. Kijas, Olivardo Facó, Hymerson Costa Azevedo, Adriana Mello de Araujo, Carlos José Hoff de Souza, and Michel Eduardo B. Yamagishi. 2017. “Genome-Wide Search for Signatures of Selection in Three Major Brazilian Locally Adapted Sheep Breeds.” *Livestock Science* 197:36–45.
SSZV. n.d. “Swiss Sheep Breeders Organization.” Retrieved (https://www.sszv.ch/).
Steiniger, S., and A. JS Hunter. 2012. “Free and Open Source GIS Software for Building a Spatial Data Infrastructure.” Pp. 247–261 in *Geospatial free and open source software in the 21st century*. Springer.
Storey, John D. 2003. “The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value.” *The Annals of Statistics* 31(6):2013–2035.
Storfer, A., M. A. Murphy, J. S. Evans, C. S. Goldberg, S. Robinson, S. F. Spear, R. Dezzani, E. Delmelle, L. Vierling, and L. P. Waits. 2007. “Putting the ‘Landscape’ in Landscape Genetics.” *Heredity* 98(3):128–42.
Strucken, E. M., R. H. Bortfeldt, D. J. De Koning, and G. A. Brockmann. 2012. “Genome-Wide Associations for Investigating Time-Dependent Genetic Effects for Milk Production Traits in Dairy Cattle.” *Animal Genetics* 43(4):375–382.
Stucki, Sylvie, Pablo Orozco-terWengel, Brenna R. Forester, Solange Duruz, Licia Colli, Charles Masembe, Riccardo Negrini, Erin Landguth, Matthew R. Jones, and NEXTGEN Consortium. 2017. “High Performance Computation of Landscape Genomic Models Including Local Indicators of Spatial Association.” *Molecular Ecology Resources* 17(5):1072–1089.
Swiss Federal Statistical Office. 2012. “STAT-TAB Landwirtschaftliche Betriebe Nach Jahr Und Gemeinde.” Retrieved (http://www.pxweb.bfs.admin.ch).
Swiss Federal Statistical Office. 2014. “Portraits of Communes.” Retrieved April 1, 2014 (http://www.bfs.admin.ch/bfs/portal/en/index/regionen/02/key.html).
Swisstopo. 2014a. “Amtliches Ortschaftenverzeichnis Mit Postleitzahl Und Perimeter.”
Swisstopo. 2014b. “swissBOUNDARIES3D.”
Swisstopo. n.d. “DHM25/200.”
Switonski, M., M. Mankowska, and S. Salamon. 2013. “Family of Melanocortin Receptor (MCR) Genes in Mammals—Mutations, Polymorphisms and Phenotypic Effects.” *Journal of Applied Genetics* 54(4):461–72.
Tabachnick, W. J. 2010. “Challenges in Predicting Climate and Environmental Effects on Vector-Borne Disease Episystems in a Changing World.” *The Journal of Experimental Biology* 213(6):946–954.
Taberlet, P., A. Valentini, H. R. Rezaei, S. Naderi, F. Pompanon, R. Negrini, and P. Ajmone-Marsan. 2008. “Are Cattle, Sheep, and Goats Endangered Species?” *Molecular Ecology* 17(1):275–84.
Tekerli, M., Z. Akinci, I. Dogan, and A. Akcan. 2000. “Factors Affecting the Shape of Lactation Curves of Holstein Cows from the Balikesir Province of Turkey.” *Journal of Dairy Science* 83(6):1381–1386.
The PostGIS Team. n.d. *PostGIS*.
The PostgreSQL global development group. n.d. *PostgreSQL*.
Thrupp, L. A. 1997. “Linking Biodiversity and Agriculture: Challenges and Opportunities for Sustainable Food Security.” *WRI Issues and Ideas (USA)*.
Tsuruta, S., I. Misztal, and T. J. Lawlor. 2004. “Genetic Correlations Among Production, Body Size, Udder, and Productive Life Traits Over Time in Holsteins.” *Journal of Dairy Science* 87(5):1457–68.
Ugurlu, M., B. Teke, F. Akdag, and S. ARSLAN. 2014. “EFFECT OF TEMPERATURE-HUMIDITY INDEX, COLD STRESS INDEX AND DRY PERIOD LENGHT ON BIRTH WEIGHT OF JERSEY CALF.” *Bulgarian Journal of Agricultural Science* 20(5):1227–1232.
Vahidi, S. M. F., M. O. Faruque, Anbaran M. Falahati, F. Afraz, S. M. Mousavi, P. Boettcher, S. Joost, J. L. Han, L. Colli, K. Periasamy, and others. 2016. “Multilocus Genotypic Data Reveal High Genetic Diversity and Low Population Genetic Structure of Iranian Indigenous Sheep.” *Animal Genetics*.
Vajana, Elia, Mario Barbato, Licia Colli, Marco Milanesi, Estelle Rochat, Enrico Fabrizi, Christopher Mukasa, Marcello Del Corvo, Charles Masembe, and Vincent B. Muwanika. 2018. “Combining Landscape Genomics and Ecological Modelling to Investigate Local Adaptation of Indigenous Ugandan Cattle to East Coast Fever.” *Frontiers in Genetics* 9.
Valais Promotion. n.d. “Combats de Reines | Valais Suisse.” Retrieved April 10, 2020 (https://www.valais.ch/fr/activites/culture-patrimoine/combats-de-reines).

Val-Arreola, D., E. Kebreab, J. Dijkstra, and J. France. 2004. “Study of the Lactation Curve in Dairy Cattle on Farms in Central Mexico.” *Journal of Dairy Science* 87(11):3789–99.

VanRaden, Paul, George Wiggans, Curt Van Tassell, Tad Sonstegard, and Leigh Walton. 2008. “Genomic Prediction.” *Changes to Evaluation System (April 2008). Online: Http://Aipl. Arsusda. Gov/Reference/Changes/Eval0804. Html*.

Verma, Nishant, Ishwar Dayal Gupta, Archana Verma, Rakesh Kumar, Ramendra Das, and Vineeth M.R. 2016. “Novel SNPs in HSPB8 Gene and Their Association with Heat Tolerance Traits in Sahiwal Indigenous Cattle.” *Tropical Animal Health and Production* 48(1):175–80.

Verrier, E., A. Audiot, C. Bertrand, H. Chapuis, E. Charvolin, C. Danchin-Burge, S. Danvy, J. L. Gourdine, P. Gaultier, D. Guémené, and others. 2015. “Assessing the Risk Status of Livestock Breeds: A Multi-Indicator Method Applied to 178 French Local Breeds Belonging to Ten Species.” *Animal Genetic Resources* 57:105–118.

Wainwright, Warwick, B. Vosough Ahmadi, Alistair Mcvittie, Geoff Simm, and Dominic Moran. 2019. “Prioritising Support for Cost Effective Rare Breed Conservation Using Multi-Criteria Decision Analysis.” *Frontiers in Ecology and Evolution* 7:110.

Waltman, William J., Steve Goddard, P. E. Read, Stephen E. Reichenbach, Ian J. Cottingham, and Jeffrey S. Peake. 2004. “Digital Government: New Tools to Define Terroirs and Viticultural Areas in the Northern Great Plains.” P. 28 in *Proceedings of the 2004 annual national conference on Digital government research*. Digital Government Society of North America.

Wang, Qiangjun, Xiaowei Zhao, Zijun Zhang, Huiling Zhao, Dongwei Huang, Guanglong Cheng, and Yongxin Yang. 2017. “Proteomic Analysis of Physiological Function Response to Hot Summer in Liver from Lactating Dairy Cows.” *Journal of Thermal Biology* 65:82–87.

Wilmink, J. B. M. 1987. “Adjustment of Test-Day Milk, Fat and Protein Yield for Age, Season and Stage of Lactation.” *Livestock Production Science* 16(4):335–348.

Wood, P. D. P. 1967. “Algebraic Model of the Lactation Curve in Cattle.” *Nature* 216(5111):164.

Xiong, Jin. 2006. *Essential Bioinformatics*. Cambridge University Press.

Yates, Andrew, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R. S. Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. 2015. “The Ensembl REST API: Ensembl Data for Any Language.” *Bioinformatics* 31(1):143–45.

Yeruham, I., D. Elad, and A. Nyska. 1995. “Skin Diseases in a Merino Sheep Herd Related to an Excessively Rainy Winter in a Mediterranean Climatic Zone.” *Journal of Veterinary Medicine Series A* 42(1–10):35–40.

Zeileis, Achim, and Torsten Hothorn. 2002. “Diagnostic Checking in Regression Relationships.”

Zendri, Francesco, Maurizio Ramanzin, Giovanni Bittante, and Enrico Sturaro. 2016. “Transhumance of Dairy Cows to Highland Summer Pastures Interacts with Breed to Influence Body Condition, Milk Yield and Quality.” *Italian Journal of Animal Science* 15(3):481–491.

Zhao, Chunping, Linsen Zan, Yan Wang, M. Scott Updike, George Liu, Brian J. Bequette, Ransom L. Baldwin VI, and Jiuzhou Song. 2014. “Functional Proteomic and Interactome Analysis of Proteins Associated with Beef Tenderness in Angus Cattle.” *Livestock Science* 161:201–209.

Zheng, Xiuwen, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, and Bruce S. Weir. 2012. “A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data.” *Bioinformatics* 28(24):3326–3328.

Zimbleman, R. B., R. P. Rhoads, L. H. Baumgard, and R. J. Collier. 2009. “Revised Temperature Humidity Index (THI) for High Producing Dairy Cows.” *J. Dairy Sci* 92(E-Suppl. 1):347.

Appendix A Supporting information for article in chapter 2

1 Appendix S1: Description of the workshop procedure to obtain thresholds and weights

12 people participated to the workshop, which were scientists and professionals whose activities are related to livestock breeding and management. In Switzerland, decisions related to Farm Animal Genetic Resources are taken by a group constituted of 10-15 experts working for the Federal Office for Agriculture (FOAG). Here - as GENMON is funded by FOAG - we adopted the same operational mode and involved the same people. The panel of experts includes people from academia (4), breeding associations (4), government agencies (2) and livestock industry (2). The participants received one week in advance the questionnaire posted at the end of this appendix, in order to enable a better preparation to feed the discussion.

On the day of the workshop, the application was first described. Then the participants were asked to fill in the sub-mentioned questionnaire on thresholds and weights. The answers were synthetized during a break, and a discussion among participants followed, to reach a consensus.

In the end, the following weights and thresholds were retained (table S1).The list of criteria is slightly different from the one given in the questionnaire, as some criteria were removed while other were added:

Table S1: Weights and thresholds retained after the workshop.

Index	Criteria	weight	Threshold T1	Threshold T2
Global index	Pedig-Index	50		
	Introgression	15	15%	3%
	Geographic concentration	15	20km	50km
	BAS - Index	10		
	Cryo-conservation	10		
Pedig-Index	Mean inbreeding	15	10% (15% for pigs)	3% (5% for pigs)
	Effective population size	40	50	250
	Pedigree completeness	15	87	97
	Trend males (last 5 years)	15	-5%	0%
	Trend females (last 5 years)	15	-5%	0%
BAS-Index	Demographic balance	5	0	3
	% jobs in agriculture	10	1	16
	Evolution of jobs in agriculture (%)	10	0	10
	% areas for breeding	15	6	30
	% of predicted agricultural land change	20	94	100
	% >65 years	15	20	4
	% <18years	5	3	10
	Cultural value of the breed	10		
	Evolution of the number of farms	10		

The full description of criteria is given in the Materials and Methods section. The weights and thresholds of the pedig-index can be set differently for each species. However, the experts decided that only pigs should have different thresholds.

GenMon Workshop

Criteria, weights and thresholds values

Your name (optional): _____

GenMon relies on different criteria, weighting factors for their consideration within the (sub-)indices. Additionally thresholds have to be defined to indicate the users (breeding organisations, NGOs and public bodies) if action is required or not. The goal of this work is to define relevant criteria, weights and thresholds with experienced persons involved in the management of AnGR of Switzerland for the further development of GenMon.

Therefore we would be grateful if you could fill in the following form during the 30 next minutes. We will then collect these documents and compile the information provided so that after the lunch break we can discuss the outcome together, based on your comments, and elaborate a consolidated list of criteria with related weights and thresholds.

The table at the end of the document summarizes all pieces of information for which we would like you to give your opinion. Please read first the instructions to understand how to fill in this table.

- **Criteria**

  The upper mentioned table and the figure 1 below lists all of the 14 criteria we are considering so far. If you have any new criteria to consider in mind, please write them down in the empty lines at the end of the table. On the other hand, if a criterion does not make sense to you, please indicate that in the comment column.

- **Weighting**

  Weights are used to confer the relative importance of each criterion. The system works like this: let us assume you have 100 points to distribute among the different criteria; you will attribute more points to the most important criteria.

  Here we have to define the weight of the components of the final index (1 in the figure hereunder) and of two sub-indices (2 and 3 in the figure). In total you have 300 points to distribute (100 per index or sub-index).

  If you think that a criterion does not make sense, you can assign it a weight of 0. In that case, please indicate in the comments why you think so.

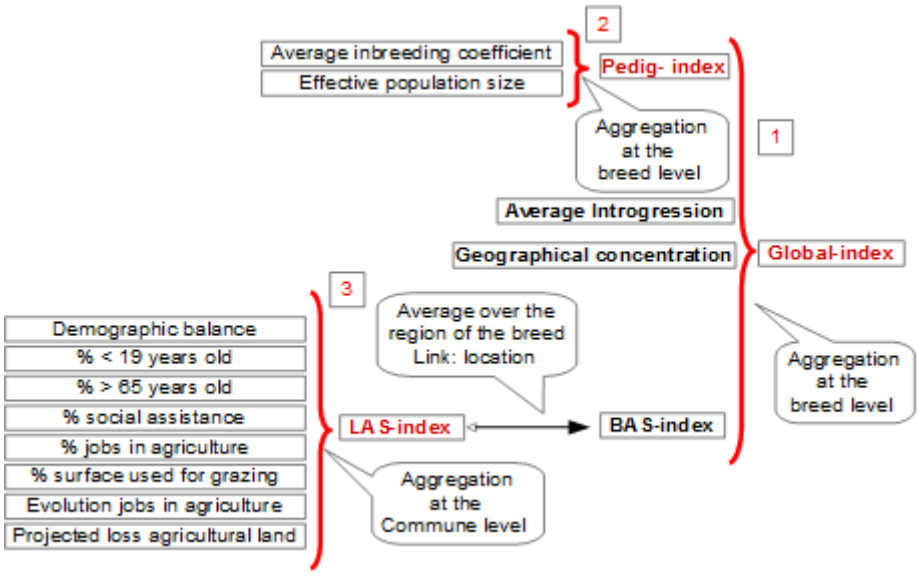


Figure 1: Criteria and their aggregation

- ## Thresholds (T1 and T2)

  For each of the following criteria, we have to define a non-satisfaction (T1) and a satisfaction threshold (T2). In the context of monitoring the thresholds will indicate if immediate action is required or not. Thereby the thresholds are the levels of the different criterions where the lights of the "monitoring-ample" are turning from red (not acceptable) to orange or to green (totally acceptable). It means that you have to decide for each criterion what is the lower limit value corresponding to a non-satisfactory situation, and what is the upper limit value corresponding to a satisfactory situation. The following figure provides an example of how thresholds work. This example is the criterion "average inbreeding". Here the satisfaction threshold is defined to 2% change, and any percentage below 2% will get the maximum satisfaction score (1=100% satisfactory). On the other hand, the non-satisfaction threshold is set to 10% and any breed having an average inbreeding higher than 10% have the lowest satisfaction score (0=0% satisfactory).

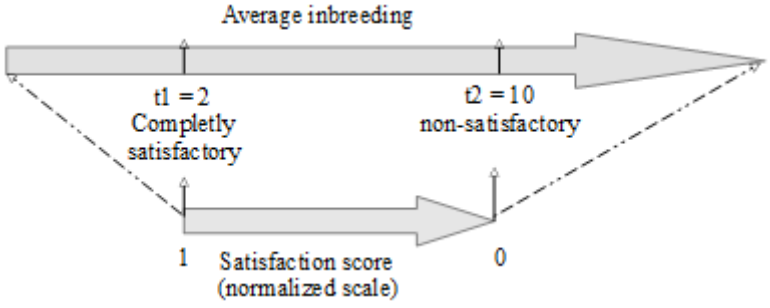


Figure 2: Scaling of a criterion using satisfaction thresholds

Note that there is no need to fill in gray cells in the table.
Moreover, if some thresholds apply only to specific species, please mention it in the comments column. You can also provide different thresholds for different species if you want (indicate it clearly in the comments column)

To help you determine the thresholds, we provide you with some ranges, explanations and statistics

- Introgression and Inbreeding: ranges between 0 and 100 %
- Geographic spread: This corresponds to a radius containing 75% of the animals. As an example, the Schwarznasenschaf has a radius of 13km, while the Braunvieh original has a radius of 59km.
- Agriculture sustainability: some statistics about Swiss communes are given in the table below. Also, the chosen non-statisfaction (T1) and satisfaction (T2) thresholds. Therefore, only indicate those thresholds (light gray in the table to fill in) if you do not agree with the proposed thresholds.

	Min	Max	Average	Standard deviation	Chosen T1	Chosen T2
Demographic balance	-18.2	38.8	1.8	3.4	0	3
Social assistance rate	0	11.4	1.9	1.7	5	2
% jobs in agriculture	0	100	15.8	16.1	1	16
Evolution of jobs in agriculture (%)	-100	1300	9.4	37.0	0	10
% areas for breeding	0	100	24.1	17.9	6	30
% of predicted agricultural land kept	78.5	100	98.4	2.5	94	100
% pop >65 years	0	36.4	21.2	3.5	20	4
% pop <18years	6.3	66.7	17.7	4.1	3	10

- **General comments?**

  If you want to give a general comment, you are encouraged to do so at the back of the page.

index	Criteria	Weight	T1	T2	Comments
Global 1	Genetic/pedigree (inbreeding, Ne)				
	Introgression				
	Geographic spread				
	Socio-economic criteria				
Genetics/ pedigree 2	Inbreeding				
	Population size				
Agriculture sustainability 3	Demographic balance		0	3	
	Social assistance rate		5	2	
	% jobs in agriculture		1	16	
	Evolution of jobs in agriculture (%)		0	10	
	% areas for breeding		6	30	
	% of predicted agricultural land kept		94	100	
	% pop >65 years		20	4	
	% pop <18years		3	10	
New criteria?					

General comments

2 Appendix S2: Descriptive statistics of the selected variables used in the local agriculture sustainability index

2.1 Descriptive statistics

The summary statistics of the 7 variables as well as the non-satisfaction (t1) and satisfaction thresholds (t2) applied in the context of this paper are displayed in table S2.

Table S2: Description of the variables used in the LAS-index

	Min	Max	Average	Standard deviation	t1	t2
% Landuse forecast 2050	78.5	100	98.4	2.5	94	100
Demogr. Balance	-18.2	38.8	1.8	3.4	0	3
Job agriculture	0	100	15.8	16.1	1	16
Grazing surface	0	100	24.1	17.9	6	30
Young	0	36.4	21.2	3.5	3	10
Old	6.3	66.7	17.7	4.1	20	4
Evolution jobs	-100	1300	2.6	37.0	0	10

t1 indicates the non-satisfaction threshold, t2 the total satisfaction threshold. Landuse forecast: percentage of agricultural land still used for agriculture in 2050; Demogr balance: difference in population between 2010 and 2012; Job agriculture: percentage of jobs in primary sector; grazing surface: the percentage of surface used for breeding activities; Young: percentage of people younger than 19 years; Old: percentage of people older than 65 years; Evolution jobs: difference in number of jobs in primary sector between 2010 and 2012)

The variables shown in table S2 have been subjectively selected by a group of 12 scientists as a priori significant and meaningful. The chosen t_{nj} and t_{cj} limits are always selected in the vicinity of the 1st and, respectively, the 3rd quartiles of the distributions.

For the demographic balance the null satisfaction was set to 0, i. e. all the municipalities having experienced a population decrease between 2012 and 2014 receive a satisfaction score of 0%. Complete satisfaction was obtained if the population of a municipality increase of 3%, a proportion deemed to be sufficiently large to bring concrete benefits.

The proportion of farmers deemed to yield null satisfaction has been set to 1% of the active population, a threshold avoiding municipalities with an underdeveloped agricultural sector to receive excessively bad satisfaction score. A symmetric logic is applied to the upper tail of the distribution, with rural municipalities exceeding 16% reaching a complete satisfaction level of 100%.

Then, for the percentage of surface used for breeding activities, 30% of grazing surface in a municipality was deemed sufficient to achieve the maximal satisfaction, whereas values below 6% were recognized as equally unsatisfying.

Finally, as regards the evolution of the number of jobs in agriculture, the null satisfaction was set to 0, so that municipality losing jobs in agriculture have a 0 satisfaction score. On the other hand, a positive evolution was judged completely adequate if it exceeded 10 full-time equivalents jobs.

The data range existing between null and complete satisfaction thresholds is comparable between the different variables: it always corresponds to the same partial satisfaction interval (0% to 100%).

2.2 Independence of selected variables

Table S3 shows the correlation matrix between the 8 variables selected, computed for the 2,564 Swiss municipalities. Selected criteria show a satisfactory independence with a largest correlation (in absolute value) of -0.67 between the percentage of people younger than 19 years old and those older than 65. This correlation was expected and the weights accorded to these two variables should be set accordingly, in order not to overweight the age structure of the population. All other correlation values are quite low (the second largest value being 0.41), which translates a sufficient level of non-redundant information for a proper assessment of the sustainability of the breeding activities in the Swiss municipalities.

Table S3 : Correlation matrix between the 8 variables included in the LAS-index

	Landuse forecast	Demogr. Balance	Job agriculture	Grazing surface	Young	Old	Evolution jobs
Landuse forecast		-0.10	0.19	-0.31	-0.13	0.29	0.06
Demogr. Balance			-0.15	0.00	0.19	-0.29	0.02
Job agriculture				0.29	0.20	0.05	-0.08
Grazing surface					0.41	-0.31	-0.17
Young						-0.67	-0.09
Old							0.05
Evolution jobs							

Landuse forecast: percentage of agricultural land still used for agriculture in 2050; Demogr balance: difference in population between 2010 and 2012; Job agriculture: percentage of jobs in primary sector; grazing surface: the percentage of surface used for breeding activities; Young: percentage of people younger than 19 years; Old: percentage of people older than 65 years; Evolution jobs: difference in number of jobs in primary sector between 2010 and 2012

3 Supplementary Figures

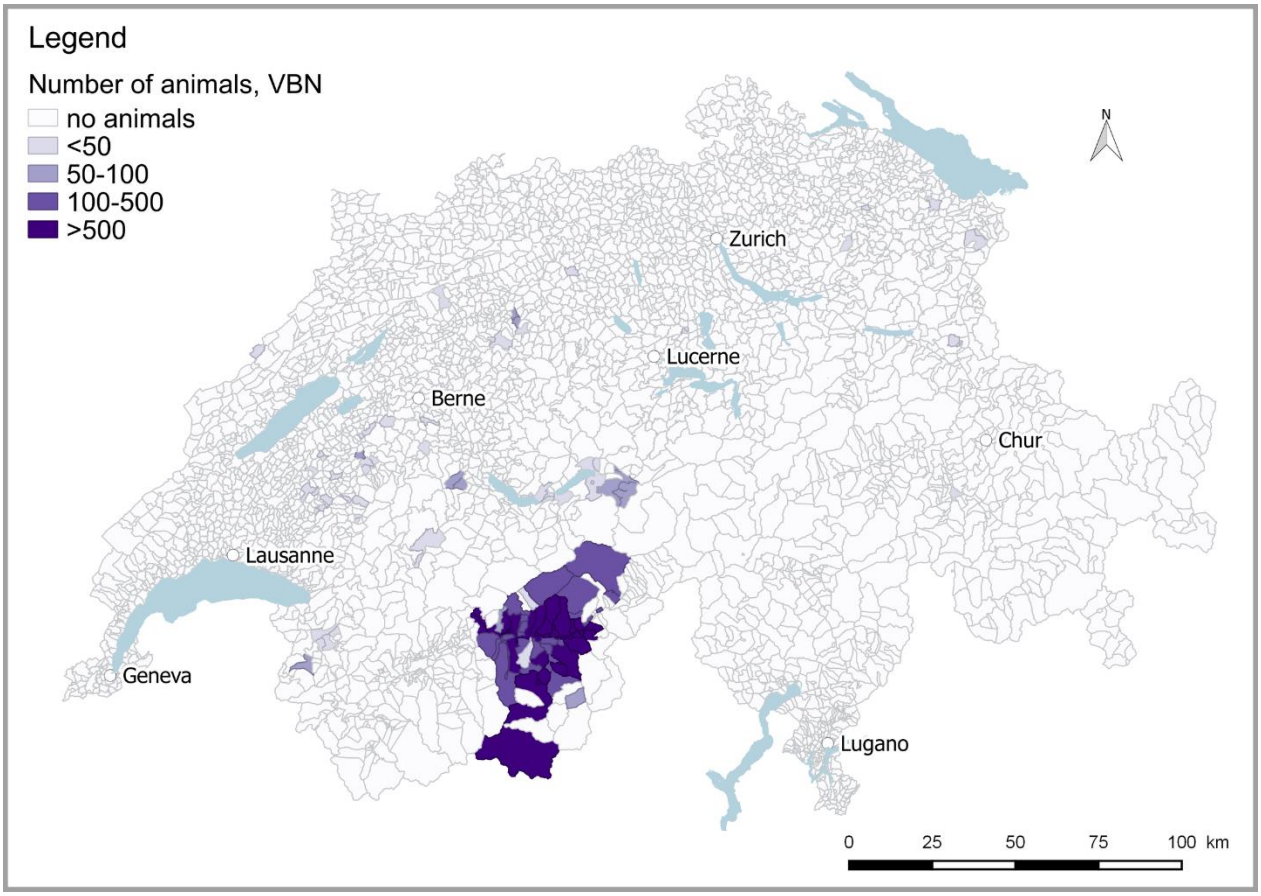


Figure S1: The geographical distribution of the number of individuals per ZIP-code for the Valais Blacknose (VBN) sheep Null areas correspond to regions where no VBN sheep is reared

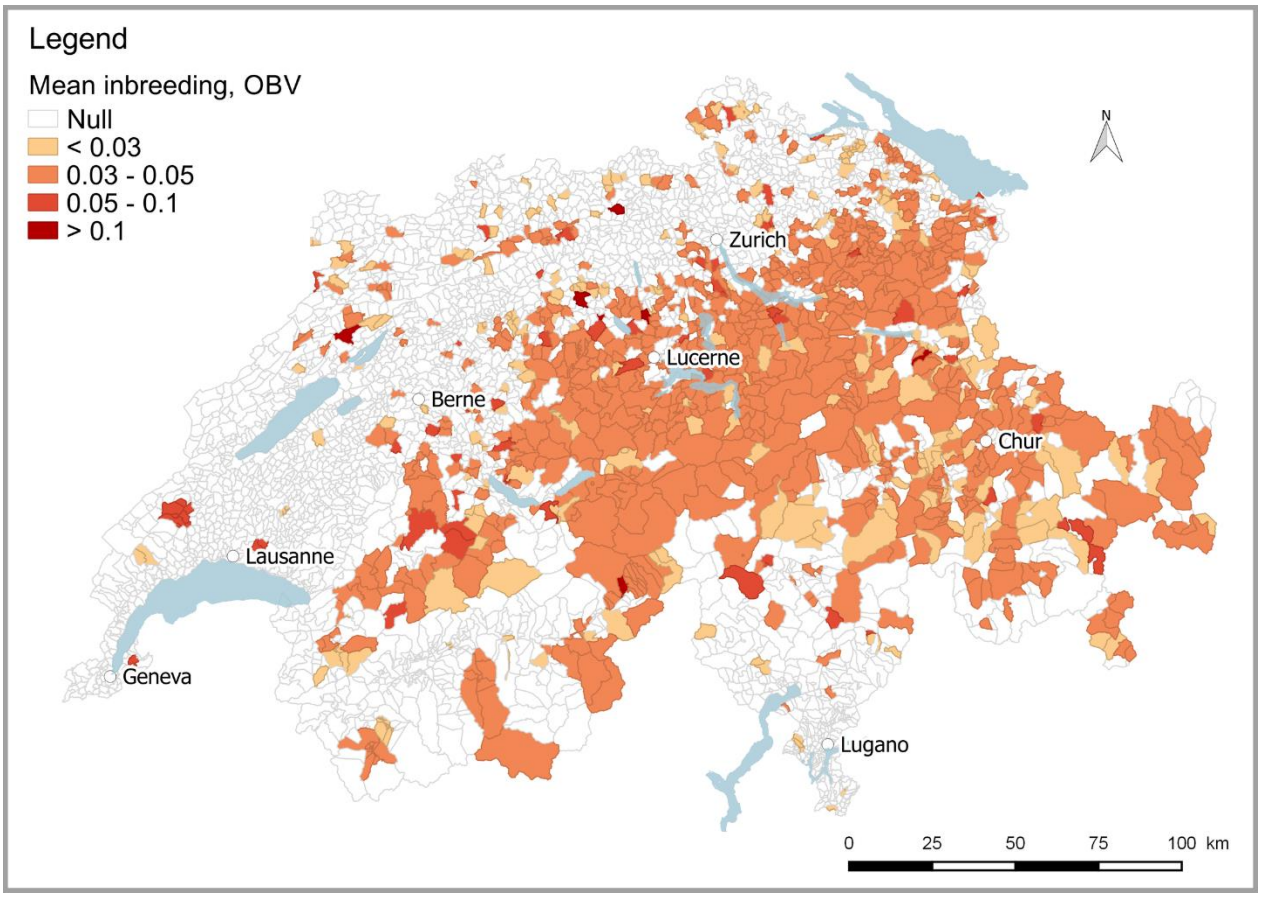


Figure S2: The geographical distribution of mean inbreeding coefficients per ZIP-code for the Original Braunvieh (OBV) cattle Null areas correspond to regions where no OBV cattle is reared.

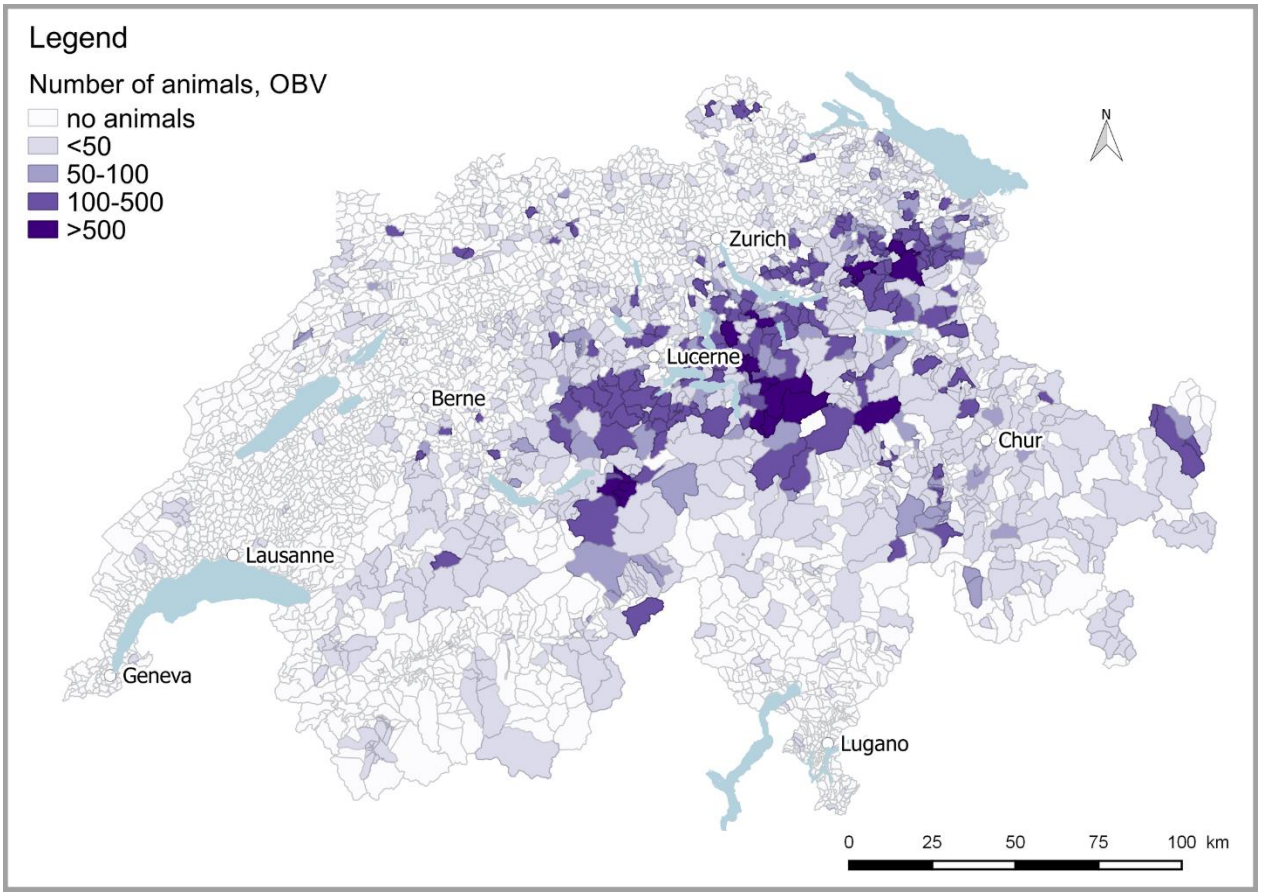


Figure S3: The geographical distribution of the number of individuals per ZIP-code for the Original Braunvieh (OBV) cattle Null areas correspond to regions where no OBV cattle is reared.

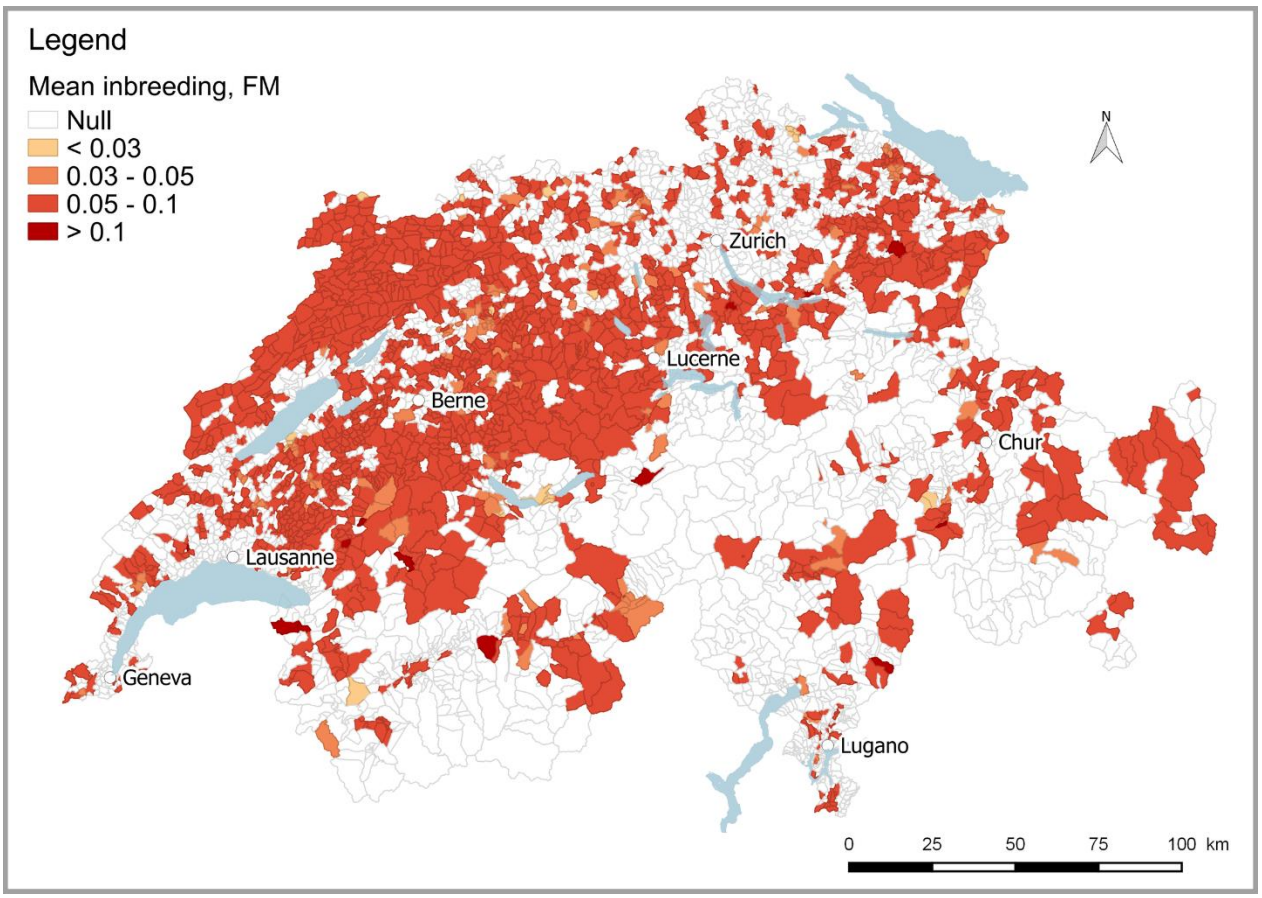


Figure S4 The geographical distribution of mean inbreeding coefficients per ZIP-code for the Franches-Montagnes (FM) horse Null areas correspond to regions where no FM horse is reared.

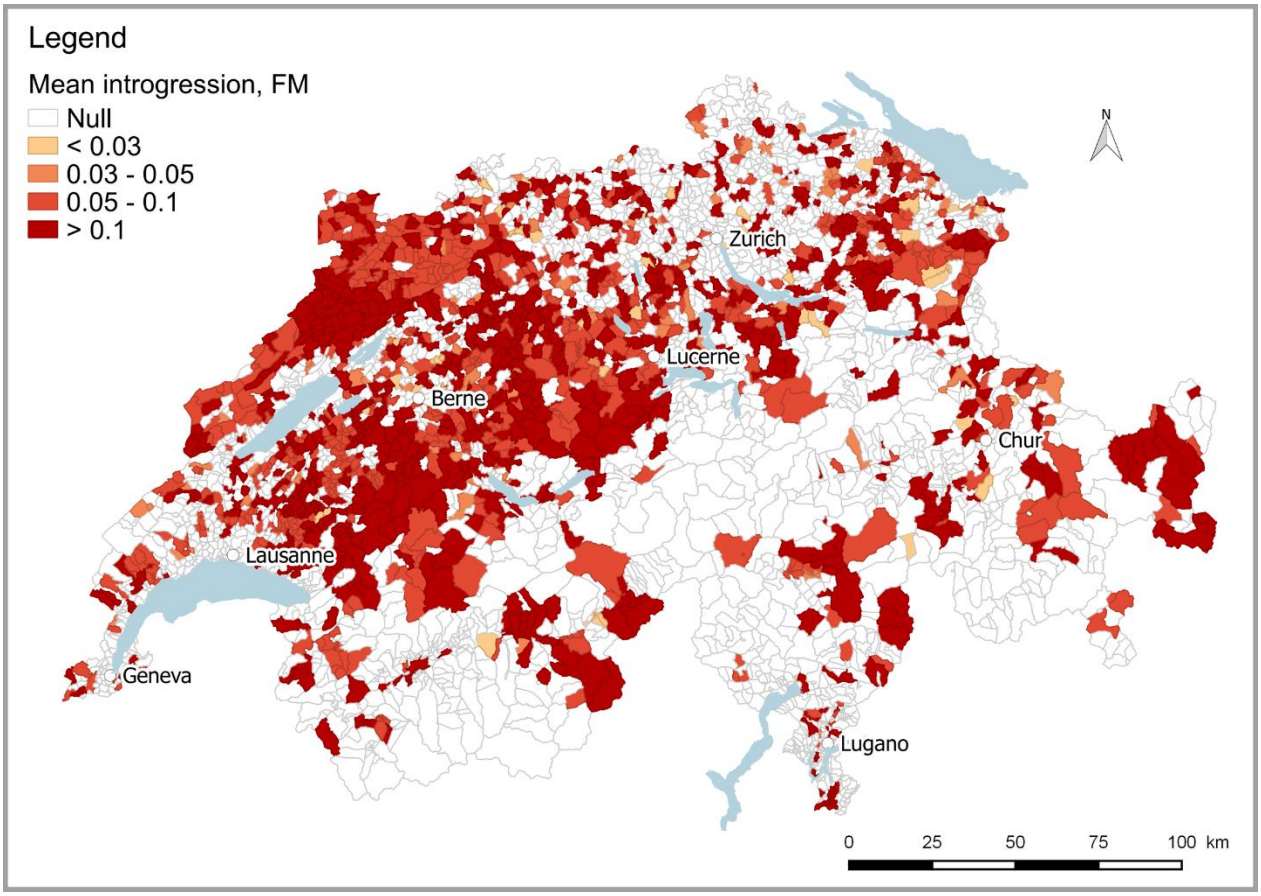


Figure S5: The geographical distribution of introgression per ZIP-code for the Franches-Montagnes (FM) horse Null areas correspond to regions where no FM horse is reared.

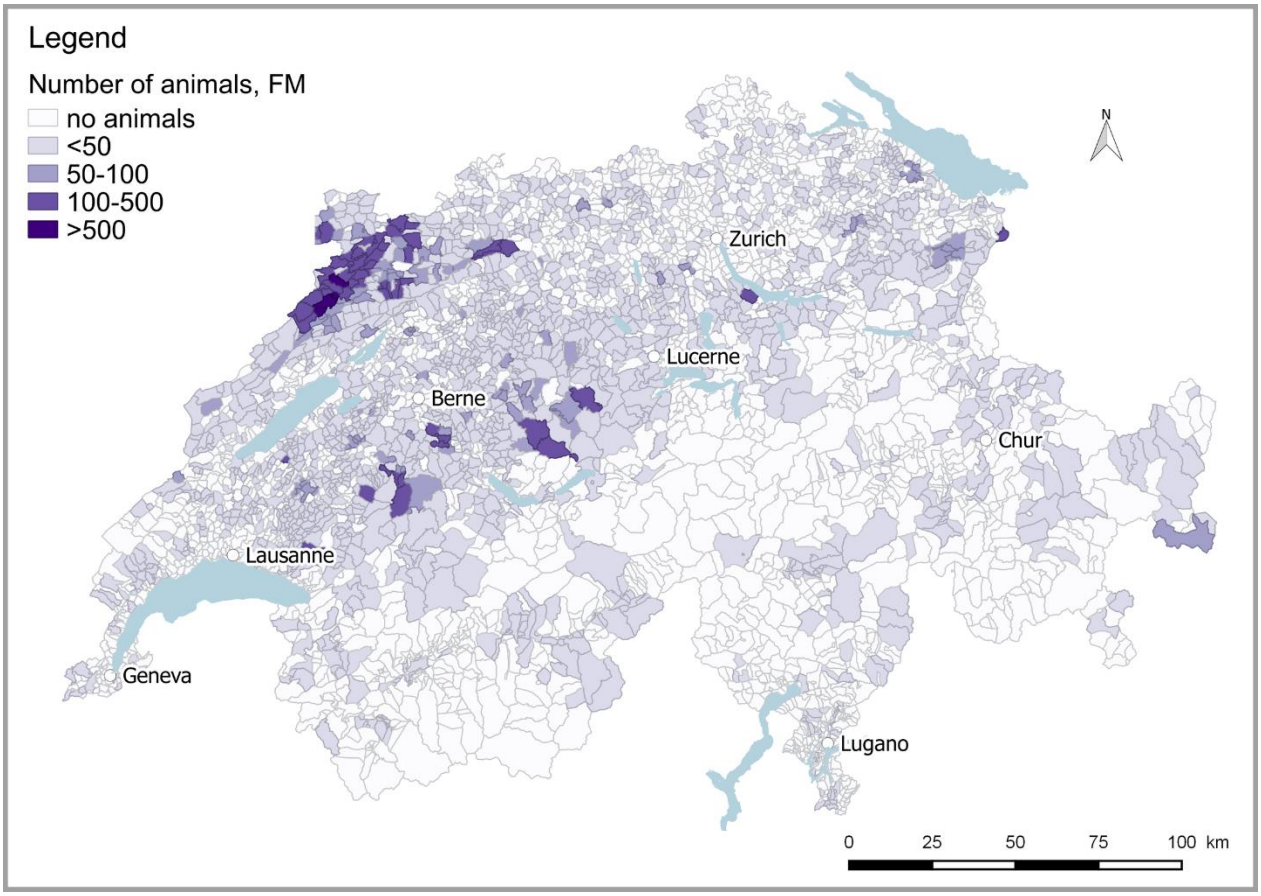


Figure S6: The geographical distribution of the number of individuals per ZIP-code for the Franches-Montagnes (FM) horse Null areas correspond to regions where no VBN sheep is reared.

Appendix B
Background information for chapter 3

Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., Negrini R., Landguth E., Jones M. R., The NEXTGEN Consortium, Bruford M. W., Taberlet P., Joost S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, *17*(5), 1072-1089.

Abstract

With the increasing availability of both molecular and topo-climatic data, the main challenges facing landscape genomics – that is the combination of landscape ecology with population genomics – include processing large numbers of models and distinguishing between selection and demographic processes (e.g. population structure). Several methods address the latter, either by estimating a null model of population history or by simultaneously inferring environmental and demographic effects. Here we present SAMβADA, an approach designed to study signatures of local adaptation, with special emphasis on high performance computing of large-scale genetic and environmental data sets. SAMβADA identifies candidate loci using genotype–environment associations while also incorporating multivariate analyses to assess the effect of many environmental predictor variables. This enables the inclusion of explanatory variables representing population structure into the models to lower the occurrences of spurious genotype–environment associations. In addition, SAMβADA calculates local indicators of spatial association for candidate loci to provide information on whether similar genotypes tend to cluster in space, which constitutes a useful indication of the possible kinship between individuals. To test the usefulness of this approach, we carried out a simulation study and analysed a data set from Ugandan cattle to detect signatures of local adaptation with SAMβADA, BAYENV, LFMM and an F_{ST} outlier method (FDIST approach in ARLEQUIN) and compare their results. SAMβADA – an open source software for Windows, Linux and Mac OS X available at http://lasig.epfl.ch/sambada – outperforms other approaches and better suits whole-genome sequence data processing.

1 Introduction

In the 1970s, several studies reviewed by Hedrick *et al*. (1976) implemented gene–environment associations to correlate the frequency of alleles with an environmental variable to look for signatures of selection (see also Mitton *et al*. 1977). Thirty years later, Joost *et al*. (2007, 2008) developed the concept to allow simultaneous processing of large numbers of logistic regressions to accommodate

the increasingly larger numbers of molecular markers in use since the introduction of PCR (e.g. ALFPs, microsatellites). Since then, correlative approaches have been used in parallel with population genetics outlier-detection methods (e.g. Beaumont & Nichols 1996; Vitalis *et al*. 2003; Foll & Gaggiotti 2008) as cross-validation (e.g. Jones *et al* . 2013; Henry & Russello 2013) to detect signatures of local adaptation, that is a region of the geographic landscape where a particular genetic variant occurs at higher frequency and is correlated with an environmental variable, potentially reflecting the higher fitness it confers to its carriers in that region (see a review in Vitti *et al*. 2013). Even though this kind of approach is still in vogue (Colli *et al* . 2014; Lv *et al*. 2014), there has been a recent revival in the interest of developing new statistical approaches for landscape genomics for use with genome-scale data sets, as such analyses enable the inference of environmental drivers of selection (Coop *et al* . 2010; Frichot *et al* . 2013; Günther & Coop 2013; Guillot *et al* . 2014; Frichot & François 2015; Gautier 2015; de Villemereuil & Gaggiotti 2015). For example, BAYENV (Günther & Coop 2013) implements a Bayesian method to compute correlations between allele frequencies and ecological variables taking into account differences in sample sizes and population structure. LFMM (Frichot *et al*. 2013; Frichot & François 2015) estimates the influence of population structure on allele frequencies by introducing unobserved variables as latent factors, while SGLMM (Guillot *et al*. 2014) extends the approach of Coop *et al*. (2010) by rooting it in a spatially explicit model and by implementing inference by means of the Integrated Nested Laplace Approximation and Stochastic Partial Differential Equation (SPDE) computational framework. Recently, Gautier (2015) introduces BayPass elaborating on the BAYENV model to capture some linkage disequilibrium information, among other important improvements, while de Villemereuil & Gaggiotti (2015) present BAYESCENV, an F_{ST}-based genome-scan method, which takes into account environmental differentiation between populations. It is based on the Beaumont & Balding's (2004) *F* model and similarly as implemented on BAYESCAN (Foll & Gaggiotti 2008), it considers that genetic variation at a given locus is affected by demographic processes that affect the entire genome (e.g. population expansions), selective events that change the allele frequencies at the locus as a response to an environmental variable (e.g. local adaptation to high temperature), and additional effects unrelated to the environmental variable tested. These methods aim at distinguishing between the effects of selection and those of demographic history; however, the increasing availability of large genomic data sets, has increased the computational intensity of this problem. In parallel, the geographic coordinates of samples are becoming frequently collected during field campaigns, enabling the computation of spatial statistics to shed an independent light on the interaction of selection and demographic signals.

Here we present the software SAMβADA, an extension of MATSAM (Joost *et al*. 2008), which offers an open source multivariate analysis framework to detect signatures of local adaptation in large-scale population genomics data sets. SAMβADA focuses on high performance computing to process whole-genome data and includes spatial statistics that measure indices of spatial autocorrelation to account for underlying patterns of spatial association in the data set due to population structure. The program is illustrated using two case studies: one in 5000 diploid individuals simulated for 100 SNPs in a heterogenous landscape, and the other one in 813 *Bos taurus* and *Bos indicus* individuals in Uganda genotyped for ~40 000 SNPs. Lastly, SAMβADA's performance is compared with other state-of-the-art software programs to detect signatures of selection.

2 Materials and methods

This section first presents SAMβADA's approach and implementation, with an overview of the accompanying modules. The second part introduces two case studies using simulation and a data set from Ugandan cattle, and how these data were collected and prepared for the subsequent analyses.

2.1 SAMβADA's approach

SAMβADA provides a locus-based approach to study local adaptation in a set of polymorphic markers using genome–environment associations. It aims at determining whether each investigated molecular marker is selected by one or a set of specific environmental variables (e.g. while multiple loci may be selected by the same environmental variable, it is also possible that different loci are affected by different environmental variables). As the analysis is performed independently for each locus, the number of possible combinations grows quickly with the size of both molecular (i.e. number of markers) and environmental data sets (i.e. number of variables) tested. To enable processing of large data sets, SAMβADA provides an automated procedure for selecting candidate loci associated with the environmental variables tested. For each locus, the set of predictor variables is kept parsimonious, because the main goal of the method is to detect which loci are potentially locally adapted rather than making predictions for the genotype of an individual based on its habitat. SAMβADA uses logistic regressions to model the probability of observing a particular genotype of a polymorphic marker given the environmental conditions at the sampling locations (Joost *et al*. 2007). As the state of a given genotype is considered as a binary presence/absence in each sample, SAMβADA can handle many types of molecular data (e.g. SNPs, indels, copy number variants and haplotypes), provided the user formats the input as required by SAMβADA and described in the software's documentation. Specifically, biallelic SNPs are recoded as three distinct genotypes (e.g. AA, AG and GG).

2.1.1 Univariate analysis

In the univariate case, each model involving a genotype and an environmental variable is compared with a constant model, in which the probability of the presence of the genotype is the same at each location in the landscape and is equal to its frequency in the data set. A maximum likelihood approach (Dobson & Barnett 2008) is used to fit the models. Significance is assessed with both log-likelihood ratio (*G*) and Wald tests (Joost *et al*. 2007). Bonferroni correction is applied for multiple comparisons (Bonferroni 1936; Shaffer 1995). To this end, the nominal significance threshold α is divided by the number m of hypotheses to be tested, that is the number of models that were fitted (e.g. if 10 000 SNPs are tested with five environmental variables, m = 150 000, as for each biallelic SNP there are three possible genotypes), to obtain the significance threshold $\alpha'(\alpha' = \alpha / m)$. The models having both P-values (computed from G and Wald scores) lower or equal to α' are considered as significant. To avoid numerous computations of P-values, the significance threshold α' is converted to a minimum score threshold using the quantile function of the χ^2 distribution. For each model, the property 'showing a score larger or equal to the score threshold' is equivalent to 'showing a P-value

lower or equal to the threshold *α″*. Thus, the significance assessment can be performed directly on the scores.

In comparison with MATSAM (Joost *et al*. 2008), SAMβADA proposes several improvements: faster processing (see SAMβADA's implementation and Table S8, Supporting information), multivariate analysis and measures of spatial autocorrelation.

2.1.2 Multivariate analysis

In the multivariate approach, several environment variables can be used at the same time to model the presence of each genotype. In this case, the selection procedure is similar to a forward stepwise regression (Dobson & Barnett 2008) and is adapted to assess the significance of multivariate models. Both *G* and Wald tests refer to a null model to build the null hypothesis. The current model could be compared to the constant model (the same as in the univariate case) using multivariate χ^2 statistics. While rejecting the null hypothesis in this configuration would indicate that at least one parameter in the model is statistically significant, it would not provide information about which parameter(s) is relevant to the model. Therefore, SAMβADA assesses parameter significance in multivariate models with either a Wald test applied to each parameter separately (except the constant parameter) or with *G* tests excluding a parameter at a time: model selection is based on simpler models nested in the current one (see Supporting information).

Multivariate models allow the inclusion of pre-existing knowledge, provided the data constitutes a continuous variable. In particular, if population structure was analysed beforehand and can be represented as a coefficient of membership for each individual, this information can be included in the modelling. For models involving both an environmental variable and this coefficient, the selection procedure will assess whether the environmental variable is associated with the genotype while taking into account the possible effect of admixture. In case there are many ancestral populations, several coefficients may be included in the analysis.

2.1.3 Spatial autocorrelation

Beyond the detection of selection signatures, SAMβADA quantifies the level of spatial dependence in the distribution of each genotype. This measure of spatial autocorrelation refers to similarities or differences in genotypes occurrences between neighbouring individuals that cannot be explained by chance. Assessing whether geographic location has an effect on allele frequencies is especially important in landscape genomics, because statistical models assume independence between samples. Thus, if individuals with similar genotypes tend to concentrate in space, spurious correlations may co-occur with specific values of environmental variables. On the other hand, spatial independence of data strengthens the confidence in the detections. Spatial autocorrelation is a well-known concern (Legendre 1993) when investigating local adaptation, but few software allow its measurement [e.g. GEODA – Anselin *et al*. (2006) – or the libraries PySAL for PYTHON – Rey & Anselin (2010) – or SPDEP in R – Bivand & Piras (2015)].

SAMβADA measures the global spatial autocorrelation in the whole data set with Moran's *I*, as well as the spatial dependence of each point with local indicators of spatial association (LISA) (see Moran 1950; Anselin 1995 and see Sokal & Oden 1978 for application in biology). In practice, LISAs are computed by comparing the value of each point with the mean value of its neighbours as defined by a specific weighting scheme based on a kernel function (see Supporting information). The sum of LISAs on the whole data set is proportional to Moran's *I* (Anselin 1995). Both a spatially fixed kernel type relying on distance only and a varying kernel type considering the number of points can be used. SAMβADA includes three fixed kernels (moving window, Gaussian and bisquare) and a varying one (nearest neighbours). Significant spatial autocorrelation indices are determined based on an empirical distribution of the indices: for Moran's *I* , values (genotype occurrences) are permutated among the locations of individuals in the whole data set and a pseudo *P* -value is computed as the proportion of permutations for which *I* is equal to or more extreme (higher for a positive Moran's *I* or lower for a negative Moran's *I*) than the observed *I*. For LISA, the pseudo *P* -value is separately computed for each point (individual), by keeping the individual of interest fixed and permuting the values of its neighbouring points with the rest of the data set.

2.2 SAMβADA's implementation

SAMβADA was developed as a standalone application written in C++, using the Scythe Statistical Library (Pemstein *et al*. 2011) which offers functions in matrix computation and probability distributions. SAMβADA is distributed under an open source GNU General Public License to ease its use for research and teaching.

2.2.1 Desktop and high performance computing

When the development started, the estimations of computational load showed that it could prove difficult to both provide the new features described above and analyse whole-genome sequencing (WGS) data sets with a single computer. Thus, SAMβADA is distributed with a module enabling High Performance Computing of large data sets.

Desktop version (SAMβADA): SAMβADA includes multivariate analyses and spatial autocorrelation computation. Many options are provided to facilitate formatting data and to customize analyses. For instance, the significance of models is assessed during the analysis and nonsignificant associations can be discarded on the fly. Moreover, models can be sorted out according to their scores before writing the results in order to facilitate their interpretation.

Parallel computing version (SAMβADA and Supervision): To speed-up the analysis of large data sets, Supervision enables parallel processing with SAMβADA by splitting data sets and merging results. The combination of SAMβADA and Supervision makes it possible to analyse large data sets: (i) univariate logistic models identify candidate loci exhibiting selection signatures; (ii) these loci may be then investigated in the light of spatial autocorrelation measures and multivariate models. The former step may point out whether the observed correlation is due to similarities between neighbours, while the

latter allows the inclusion of population structure, if any, in the model to assess the additional effect of the environmental variable after taking demography into account.

2.2.2 Modules

SAMβADA includes several modules that enhance interfacing with other programs.

Geovisualization of spatial statistics: SAMβADA provides an option to save spatial autocorrelation results as a shapefile (.shp), a common format for storing vector information in Geographic Information Systems (GIS). This feature relies on the shplib open source library (http://shapelib.maptools.org/), which is included and distributed with SAMβADA .

Recoding molecular data: SAMβADA is distributed with a utility for recoding molecular data into binary information, so that each genotype is considered on its own. Currently RecodePlink handles ped/map files, a standard format for SNP data used in genomics analysis (Purcell *et al*. 2007).

Supervision: For very large molecular data sets, SAMβADA provides a module to share workload between computers. Supervision splits the input data in several files that can be processed separately, even on independent computers. At the end of an analysis, Supervision merges the results to provide the same output as if the whole data set had been processed at once. This module enables the processing of WGS data sets with SAMβADA using a couple of desktop computers (see Table S9, Supporting information).

2.3 Alternative methods to detect selection

The performance of SAMβADA was compared with other software for detecting signatures of selection. These analyses involved two other correlative approaches [BAYENV – Coop *et al*. (2010) – and Latent Factor Mixed Models – Frichot *et al*. (2013); Frichot & François (2015)], and an F_{ST}-outlier-detection approach (Beaumont & Nichols 1996) included in ARLEQUIN 3.5 (Excoffier & Lischer 2010). Please note that these methods consider allele counts, whereas SAMβADA recodes them into genotypes. An overview of BAYENV, LFMM and ARLEQUIN is available in the supporting information.

2.4 Simulation study

As SAMβADA and LFMM (Frichot *et al*. 2013; Frichot & François 2015) share a similar correlative approach, simulated data were used to compare their performance in scenarios where the selected loci are known. The analyses used a subset of the simulation data generated by Forester *et al*. (2016) who included LFMM in their work.

2.4.1 Simulated data

The simulations were run using the program CDPOP v1.2 (Landguth & Cushman 2010), which models population genetic change across a landscape surface as a function of mutation, mating, gene flow,

drift and selection. Each simulation had 5000 diploid individuals with 100 bi-allelic loci, one of which was subject to selection. All loci experienced a 0.0005 mutation rate per generation, free recombination and no physical linkage. Ten Monte Carlo (MC) replicates of each simulation were run for a total of 1250 generations, discarding the first 250 generations as burn-in (no selection imposed) to establish a spatial genetic pattern prior to initiating the landscape selection configurations.

The simulations used a discrete landscape selection configuration generated using the neutral landscape model QRULE (Gardner 1999) to simulate binary landscape maps (1024 × 1024 pixels). Habitat fragmentation was controlled with the *H* parameter, which affects the aggregation of habitat pixels. A low value of *H* (*H* = 0.1) was used, resulting in less aggregated (more dispersed) habitat patches, and 10 landscape replicates were produced (one for each MC replicate) to average across stochastic variation among simulated landscapes. Discrete habitat types (type 'AA' or 'aa') represented habitat patches in which AA or aa genotypes were, respectively, favoured (see Fig. S3, Supporting information for an example of the landscape configuration).

The effect of varying selection strength was tested, mediated through density-independent (i.e. environment-driven) mortality (*s*) determined by genotypes of the selected locus. Selection strengths included *s* = 0.01 or '1%', *s* = 0.05 or '5%', and s = 0.10 or '10%'. AA individuals had no mortality in 'AA' habitat patches and experienced 1%, 5% or 10% mortality if they occurred in 'aa' patches. Individuals with 'aa' genotypes at the locus under selection experienced the opposite selection gradient. The Aa genotypes experienced uniform selection (*s* /2) across the entire surface.

Dispersal capacity for movement and mating was set to a maximum of 5% of the landscape surrounding an individual, with dispersal occurring once per generation. Mating pairs of individuals and dispersal locations of offspring were chosen based on a random draw from the inverse-square probability function of distance, truncated with the specified maximum distance. Mating parameters represented a population of unisexual individuals with females and males mating with replacement. The number of offspring produced from mating was determined from a Poisson distribution ($\lambda = 4$), which produced an excess of individuals each generation to maintain a constant population size of 5000 individuals at every generation. Carrying capacity of the simulation surface was 5000 individuals. Excess individuals were discarded once all 5000 locations became occupied, which is equivalent to forcing out emigrants once all available home ranges are occupied (Balloux 2001; Landguth & Cushman 2010). Combining the 10 landscape configurations and the three levels of selection strength, a total of 30 molecular data sets were analysed in this simulation study.

2.4.2 Simulation analysis

A set of 500 individuals were randomly selected from each simulation of 5000 individuals (the 500 individuals were chosen from the same position in the grid in each simulation and replicate) to carry out the selection analyses with SAMβADA and LFMM (see Fig. S3, Supporting information). Simulation data were filtered for a minimum allele frequency (MAF) of 1%; no simulation loci were found to have a MAF <1%. All analyses used three environmental predictor variables: the *x* -coordinate location of an individual ('*x* '), the *y* -coordinate location of an individual ('*y* ') and the location of an individual in an AA or aa patch ('habitat'). Two types of analyses were run with SAMβADA : (i) Univariate analysis with

the three environmental predictor variables; (ii) Multivariate analysis using the population structure to build the null models. For univariate analysis, the significance threshold was set to $\alpha' = 0.01/900$ (100 loci, three genotypes and three environmental variables) after Bonferroni correction. The second type of analyses was performed as follows for each replicate: Population structure was assessed with ADMIXTURE (Alexander *et al*. 2009) using the 99 neutral loci. ADMIXTURE (Alexander *et al* . 2009) estimates the maximum likelihood of individual ancestries from multilocus SNP genotype data sets and assumes that samples descend from a predefined number of ancestor populations that became mixed. ADMIXTURE estimates both the fraction of each sample coming from each population and the marker frequencies in these populations. The optimal number of populations K is assessed by a k -fold cross-validation procedure (see Table S4, Supporting information, for the value of K in each simulation). As the sum of the coefficients of admixture is 1.0 for each sample, only $(K - 1)$ values are required to specify the ancestry of each sample. Thus, $(K - 1)$ 'population variables' were created by computing a PCA on the coefficients of admixture and by taking the $(K - 1)$ first principal components. The set of predictor variables was composed by the three environmental variables ('*x* ', '*y* ' and 'habitat') and the $(K - 1)$ 'population variables'. The $(K - 1)$ 'population variables' were used to compute a 'null model' including the population structure for each marker, and then, the models to be tested were built by adding one environmental variable to the set of 'population variables'. In the current implementation of SAMβADA , this is performed by computing all the models from 1 to K variables (i.e. the total number of clusters in the data) before extracting the models of interest. As the models to be tested included one variable more than their corresponding null model, the total number of models considered for the Bonferroni correction was the same as for the univariate analysis.

For LFMM, K was determined using the Patterson method (Patterson *et al*. 2006) as suggested by Frichot *et al*. (2013) for simulation studies (see Table S5, Supporting information, for the value of K in each simulation). LFMM models were run with the package LEA (v. 1.2.0; Frichot & François 2015) in R (v. 3.2.3; R Core Team 2016) using the following parameters: 10 000 iterations with a burn-in of 5000 iterations, and five replicate runs. The median z -score and P -value were chosen from each set of five runs; significant outliers were detected as those loci with a P-value<(0.001/300) after Bonferroni correction. The significance thresholds α for SAMβADA and LFMM were estimated separately for each method.

For each of the three simulation scenarios, the following metrics were averaged across the 10 replicates: true-positive rate (TPR), false-positive rate (FPR) and a genotype–environment association index (GEA) that determines how effective a method is at identifying the predictor that is driving selection (Forester *et al*. 2016). The GEA index ranges from 3 (best performance) to 0 (worst performance) and is coded: 3 = correct identification of variable 'habitat'; 2 = 'habitat' is significant, but less than '*x* ' or '*y* '; 1 = 'habitat' is not detected but '*x* ' or '*y* ' are; and 0 = no variable is detected as significantly associated with the locus under selection.

2.5 Ugandan cattle

In addition to the simulated data set, we illustrate the use of SAMβADA with an empirical data set of Ugandan cattle, which is composed of two main populations. Ankole (or Ankole-Watusi) cattle are a

Sanga breed (taurine-zebu cross) that appeared in the Nile Basin around 2000 years BC. They migrated southward and are now found in southwest Uganda, Rwanda and Burundi (Ndumu *et al*. 2008; Ajmone Marsan *et al*. 2010). Shorthorn zebus were introduced in East Africa around the VIIIth century AD; they later spread as they were less affected than taurine and Sanga cattle by rinderpest, but their susceptibility to trypanosomiasis is presumed to have restrained their dispersion across Africa (Ajmone Marsan *et al*. 2010). Shorthorn zebus are now common in northeast Uganda and are being crossbred with Ankole cattle in the centre of the country.

2.5.1 Sampling design

In the context of the European Nextgen project (http://nextgen.epfl.ch), the sampling of Ugandan cattle was designed to cover the whole country, including each eco-geographic region, and to obtain a homogeneous geographic distribution of individuals across the country. To this end, a regular grid made of 51 cells of 70 × 70 km was produced. On average, four farms were visited in each cell and four unrelated individuals were selected from each farm, for a total of 917 biological samples retrieved from 202 farms. The sampling season took place between March 2011 and January 2012. Recorded information also included the location of the farm, the name of the breed, a picture and morphological information (e.g. withers height and horns length) for each individual. These elements were stored in a database accessible through a Web interface, enabling real-time monitoring of the sampling campaign.

2.5.2 Molecular data

Out of the 917 individuals, 813 samples were genotyped with a medium-density SNP chip (54 609 SNPs, BovineSNP50 BeadChip; Illumina Inc., San Diego, CA, USA). Only markers located on the autosomal chromosomes were considered in the analyses. The data set was filtered with PLINK (Purcell *et al*. 2007) with a call rate set to 95% for both individuals and SNPs, and a MAF set to 1%. The resulting data set after filtering contained 804 samples and 40 019 SNPs.

2.5.3 Population structure

Population structure was analysed with the software ADMIXTURE (Alexander *et al*. 2009) using a subset of 28 197 SNPs pruned for linkage disequilibrium as recommended in the manual. The SNPs were filtered with PLINK (option `–indep-pairwise`), $r^2 < 0.2$, sliding window of 10 SNPs, step size of 5 SNPs), and the number of populations K was chosen using the cross-validation index of ADMIXTURE. The best partition of the data set consisted of four populations, although the vast majority of the samples (96%) were allocated to one of two clusters on the basis of the ancestry coefficients (Fig. S1, Supporting information). Mapping these coefficients revealed that these two clusters (340 and 431 individuals of 804) occurred in the southwest and northeast of Uganda, respectively. Using pictures of sampled individuals, the first cluster was identified as Ankole cattle and the second one as zebu. These observations are in agreement with the known background of Ugandan cattle. The remaining two clusters (33 animals in total) possibly represent introgression from allochthonous gene pools. The

results of the population structure analysis were used to define the parameters needed by each method to detect selection signatures.

2.5.4 Environmental data

Habitat characteristics of sampling locations were described with the WorldClim data set containing monthly values of precipitation, minimum, mean and maximum temperature as well as 19 derived variables, at 1 km resolution (Hijmans *et al*. 2005). This data set provides appropriate data as it consists of representative climate information collected during 30 years (WMO standard climate normal, Arguez & Vose 2010) and its high resolution suits the scale of our study. These environmental variables were originally stored in four tiles (portions of map) which were pasted using the Geospatial Data Abstraction Library (GDAL Development Team 2013) and a customized Python script. The topography is described by the 90 m resolution SRTM3 (Shuttle Radar Topography Mission) digital elevation model (DEM) (Farr *et al*. 2007). SAGA GIS (www.sagagis.org) was used to paste the 36 tiles covering the country and to derive slope and orientation from the SRTM DEM. Longitude and latitude were also taken into account as a rough proxy for population structure. Finally, the values of the 72 environmental variables were extracted for each sampling locality using the 'Point Sampling Tool' extension (http://hub.qgis.org/projects/pointsamplingtool) in QuantumGIS (www.qgis.org).

Variable selection for univariate analysis: Considering all environmental variables in the computation of the multiple logistic regressions would have provided a comprehensive analysis with a low risk of missing detections. Nonetheless, some variables are highly correlated; thus, the corresponding models for a genotype are likely to represent the same phenomenon. To lower the dependency between models and spare computation time, we used the variance inflation factor (VIF) to control for multicollinearity (Dobson & Barnett 2008). A maximum VIF of 5 was chosen, corresponding to a coefficient of correlation of 0.9 between pairs of variables. The number of variables was reduced iteratively by randomly removing one of the two most correlated variables until the maximum correlation was lower than the threshold (0.9). This procedure led to a set of 23 environmental variables that were used for univariate landscape genomic analyses (Table S1, Supporting information).

Variable selection for multivariate analysis: The multivariate analysis with SAMβADA consisted in bivariate models along with their corresponding univariate and constant models. A maximum of two explanatory variables were considered to ease the interpretation of their respective effects. Moreover, SAMβADA's conservative approach to assess model significance tends to reject models including numerous environmental variables. In this study, the multivariate models were used to take population structure into account. The information on population structure was derived from the analysis of individual ancestries. To this end, a new variable 'population structure' was defined by performing a principal component analysis (PCA) on the coefficients of ancestry and was used to represent the population structure in SAMβADA analyses (see 'Protocol of analysis' for details). It was thus added to the set of 23 environmental variables and the correlation-based variable selection method was reapplied to limit the coefficient of correlation between pairs of variables to 0.81, which corresponds to limiting the VIF to 2.9. On this basis, 15 predictor variables (including the 'population

structure' variable) were considered for SAMβADA multivariate analysis (see Table S1, Supporting information).

2.5.5 Protocol of analysis

Four approaches were applied to detect selection signatures among the 40 019 SNPs from 804 samples. As SAMβADA processes each genotype independently, while BAYENV, LFMM and ARLEQUIN treat each locus as a whole, we defined a locus as 'detected' by SAMβADA if at least one of its three genotypes showed a significant association with an environmental variable. For BAYENV, LFMM and ARLEQUIN, the selection signatures are analysed per locus.

Data preparation: Since Ugandan cattle globally comprises two admixing populations (Fig. S1, Supporting information), the 33 samples from the two smaller populations were excluded from the analyses with SAMβADA and LFMM, leading to a set of 771 samples for these methods. To estimate whether the population structure could be efficiently summarized by the Ankole and zebu clusters, a PCA was run on the coefficients of ancestry for the subset of 771 samples taken from the results of ADMIXTURE for $K = 4$. The first principal axis of this PCA accounted for 95% of the variance among all molecular markers, so that a single coefficient is sufficient to provide an overall view of an individual's ancestry. Given this configuration, SAMβADA's multivariate analysis needed a single variable, that is the first axis of the PCA, to summarize the population structure. As the cattle population is essentially constituted of two clusters, the number of latent factors tested with LFMM covered a range of values of K that included the estimated K as described by Frichot & François (2015). This range consisted of values of K from $K = 1$ to $K = 4$. For BAYENV and ARLEQUIN, as these approaches require the samples to be clearly assigned to a population, the 804 samples were classified into populations based on their coefficient of ancestry and using a threshold of 0.85, below which samples were excluded from the analysis. This led to, respectively, three clusters of 162 Ankole cattle, 8 zebus and 10 cattle from the third population; samples from the fourth population were highly admixed and none satisfied the condition. This method was preferred over a classification based on sampling locations or phenotypic traits because Ugandan cattle are generally admixed (see Fig. S1, Supporting information). The univariate correlative approaches – SAMβADA, BAYENV and LFMM – used a selected set of 23 environmental variables, while SAMβADA multivariate analysis used a set of 15 environmental variables (see 'Environmental data' for details).

Computational set-up for correlative Bayesian approaches: BAYENV (v. 2.0, Coop *et al*. 2010; Günther & Coop 2013) first estimated the interpopulation covariance matrix with a run of 100 000 iterations over a set of 1000 loci selected at random among the loci identified as neutral by SAMβADA's univariate analysis. Then, the full data set was analysed for another 100 000 iterations to detect the signatures of selection. LFMM models were run with the package LEA (v. 1.4.0; Frichot & François 2015) in R (v. 3.3.0; R Core Team 2016) using the following parameters: 10 000 iterations with a burn-in of 5000 iterations, and five replicate runs for each value of the number of latent factors.

Models selection: The statistical significance threshold for SAMβADA, LFMM and ARLEQUIN was set to $\alpha = 0.01$ before applying the Bonferroni correction. The analysis of SAMβADA's multivariate models followed the same protocol as its counter-part on the simulation data: the univariate models involving

the 'population structure' variable were used as 'null models' for assessing the significance of bivariate models involving the 'population structure' variable and one environmental variable; all other models were discarded. For LFMM, the median *z* -score and *P* -value were chosen from each set of five runs. The number of latent factors was set to $K = 2$ based on the quantile – quantile (QQ) plots (see Fig. S2, Supporting information). For BAYENV, model selection was based on the Jeffreys' scale of evidence (Jeffreys 1961) and on the distribution of Bayes Factors (BF) for neutral loci (Coop *et al*. 2010). This distribution was estimated by selecting a random subset from the loci identified as neutral by SAMβADA. BAYENV's results were analysed separately for each environmental variable and models showing a BF higher than 10 (strong evidence) or higher than the 1st percentile of the neutral distribution (if higher than 10) were used to build the set of candidate loci.

3 Results

3.1 Results for the simulated data

3.1.1 Detection of selection signatures

Univariate models in SAMβADA show that on average both the TPR and the genome–environment association index (GEA index) increase with the strength of selection (see Table 1a and Table S3, Supporting information, for detailed results). TPR ranges from 60% for the weak (1%) selection, to 90% for intermediate (5%), and to 100% for strong selection (10%), while the GEA index takes the values of 0.7, 1.6 and 2.1 for the corresponding selection pressures. The FPR is high (43–45%) but consistent among the different scenarios. When population structure is taken into account using multivariate models, the TPR index and the GEA index decrease for the weak and intermediate levels of selection compared to the univariate models, but their values remain unchanged for the stronger level of selection, whereas the FPR decreases for all levels of selection (2–4%, see Table 1b and Table S4, Supporting information, for detailed results). Overall, LFMM behaved very similar to the SAMβADA univariate approach showing the same TPR and FPR and marginally better GEA values (Table 1c and Table S5, Supporting information, for detailed results).

Table 1. Average true-positive rate (TPR), false-positive rate (FPR) and genotype–environment association index (GEA index) across the 10 replicates for each simulation scenario. All simulations use a dispersal level of 5% and a discrete landscape with an aggregation index *H* of 0.1. TPR scales from 0% (worst performance, locus under selection not detected) to 100% (best performance, locus under selection detected); FPR scales from 0% (best performance, no false detection) to 100% (worst performance, 99 neutral loci detected as significant); GEA index scales from 0 (worst performance, no detection) to 3 (best performance, correct detection). Results for (a) SAMβADA univariate models, (b) SAMβADA multivariate models taking into account the population structure (c) LFMM

Selection (%)	TPR (%)	FPR (%)	GEA index
SAMβADA univariate			
1	60	45	0.7
5	90	43	1.6
10	100	45	2.1

SAMβADA multivariate			
1	10	4	0.1
5	50	2	0.5
10	100	2	2.1
LFMM			
1	50	43	0.6
5	90	43	2.0
10	100	43	2.8

3.1.2 Spatial autocorrelation

Spatial statistics were computed for one genotype per locus for each replicate of the three selection scenarios. The choice of the genotypes was based on SAMβADA's univariate models: for each locus, the genotype in the model with the highest *G* score was chosen to represent the locus in the subsequent analyses. Spatial autocorrelation was measured using Moran's *I*, and the spatial ponderation was based on the number of nearest neighbours. The weighting schemes included 5, 15, 30, 45 and 60 neighbours. The threshold of pseudo-*P* -values was set to 0.01 (99 permutations) for assessing the significance of global and local values of Moran's *I*. Figure 1 presents an overview of the correlograms obtained for each simulation scenario. For each scenario, the loci were ordered in three groups: loci under selection (L0), neutral loci detected by SAMβADA (i.e. false-positive detections) and neutral loci not detected by SAMβADA (i.e. true-negative detections). On average, the group of false positives shows a higher value of Moran's *I* than the group of true negatives. The loci under selection show values of Moran's *I* similar to the group of true negatives for the weak selection scenario, while their values of Moran's *I* tend to be higher than both groups of neutral loci for the intermediate and strong selection scenarios (see Table 1). The individual correlograms for each replicate of the three selection scenarios are found in Figs S4–S6, Supporting information.

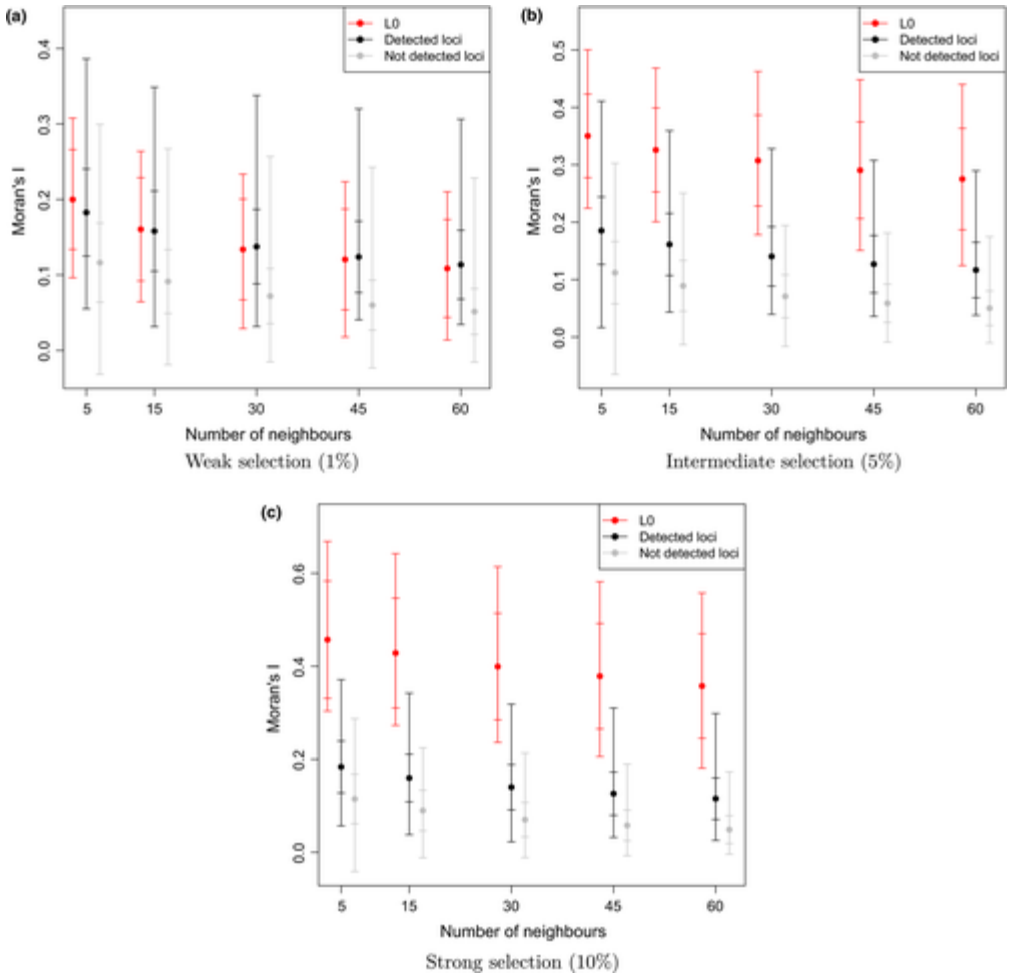


Figure 1: **Summary of correlograms computed for the simulation data. Spatial autocorrelation was measured using Moran's *I* , and the spatial ponderation was based on the number of nearest neighbours.** The weighting schemes included 5, 15, 30, 45 and 60 neighbours. Each locus was represented by its genotype involved in the model with the highest *G* score. Each graph summarizes the correlograms of one of the selection scenario s: a) weak, b) intermediate, and c) strong selection. The loci were sorted in three groups: the loci under selection (L0 – red bars), the neutral loci detected by SAMβADA (black bars) and the neutral loci not detected by SAMβADA (grey bar). For each group, the averaged Moran's *I* is represented by the dot on the bar, the two marks above and below indicate the standard deviation and the outer bounds show the minimal and maximal values of Moran's *I* for this group.

Local indicators of spatial association were summarized for each locus by counting the number of sampling points showing a significant value. The amount of significant LISA points is generally higher for the locus under selection than the averaged values of each of the two groups of neutral loci (see central part of Fig. S6, Supporting information). For the replicates where the locus L0 was detected by SAMβADA's univariate models, all detected loci were ordered according to the decreasing number of significant LISA points. For the intermediate and strong selection scenarios, the locus L0 is often found among the first loci. For instance, L0 is found between positions 1 and 5 for the LISA computed with 15 neighbours in the intermediate selection scenario (see right part of Fig. S6, Supporting information).

3.2 Results for the Ugandan cattle

3.2.1 Detection of selection signatures

Using univariate models, SAMβADA identified 2354 SNPs (5.9%) potentially subject to selection, BAYENV 1169 (2.9%), LFMM 970 (2.4%) and ARLEQUIN did not identify any locus as significant. Among the 2354 loci detected by SAMβADA, 967 were <100 000 base pairs apart from another detected locus, suggesting that some loci may be detected simply due to physical linkage to selected regions. Figure **2** counts the number of common detections between landscape genomic approaches. SAMβADA's results partially match those of BAYENV with 214 common loci (i.e. 9% of SAMβADA' and 18% of BAYENV's detections). Concerning the third correlative approach, LFMM is more conservative than SAMβADA and the overlap is smaller because 79 loci (i.e. 3% of SAMβADA' and 8% of LFMM's detections) are detected by both SAMβADA and LFMM, while 24 loci (i.e. 2% of BAYENV's and 2% of LFMM's detections) are detected by both BAYENV and LFMM. However, 110 SNPs detected only by LFMM are <100 000 base pairs apart from loci detected by SAMβADA, potentially identifying the same selection signature. Lastly, ARLEQUIN's best results involved 17 SNPs with *P* -values lower than 10^{-4}. Although these results are not significant – the threshold corrected for multiple comparisons was $\alpha' = 2.5 \times 10^{-7}$ – it is interesting to compare them with the other methods. Among these 17 SNPs, one was common with SAMβADA, 16 were common with BAYENV and none with LFMM, suggesting that population-based methods, whether using outliers or environmental correlations, tend to overlap substantially in detecting selection signatures. Quantile – quantile (QQ) plots of SAMβADA and LFMM results are presented on Fig. S2 (Supporting information).

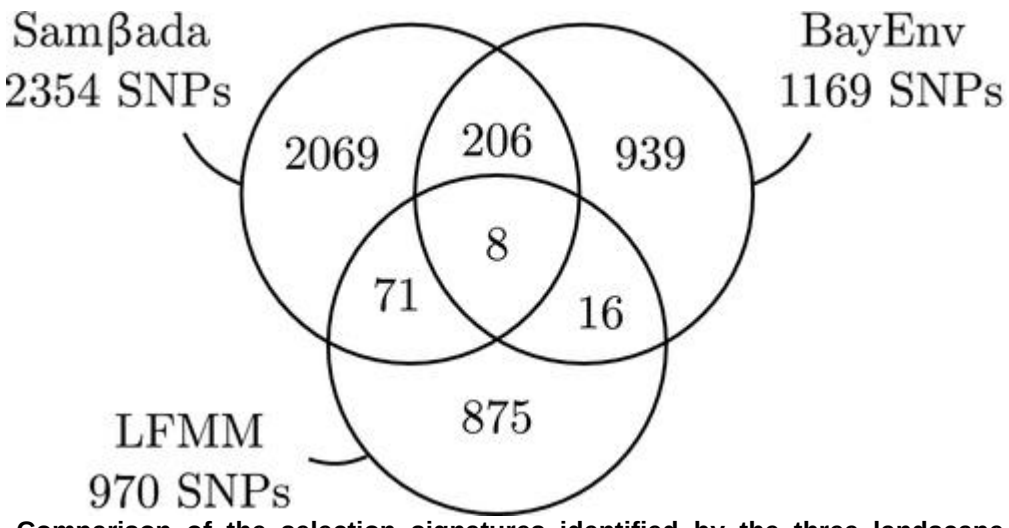


Figure 2: **Comparison of the selection signatures identified by the three landscape genomic approaches.** The total number of SNP s detected by each method is indicated below the name. The diagram shows how these sets of SNP s overlap between methods.

The loci detected by SAMβADA's univariate analysis with the highest *G* scores were compared among methods. Table 2 shows that BAYENV generally agreed with SAMβADA's detections, while LFMM's results differed. Some of the most significant loci detected by SAMβADA were ignored by LFMM. A total of eight

loci were identified by the three correlative methods and four of them were among the most significant models detected by SAMβADA (see Table 2). Three of these SNPs occur close to each other on chromosome five.

Table 2: List of SNP s detected by SAMβADA corresponding to the univariate models with the highest *G* scores. Loci are identified by their name, their chromosome and their position in million base pairs (Mbp). The following columns show whether SAMβADA (univariate), BAYENV and LFMM detected them with the corresponding environmental variables and *P*-values (SAMβADA, LFMM) or Bayes Factor (BAYENV). Loci in bold type are the common discoveries of SAMβADA univariate and bivariate, LFMM and BAYENV. Local indicators of spatial autocorrelation were analysed for SNPs on lines 4 and 7

Loci	Chr.	Pos (Mbp)	SAMβADA		BAYENV		LFMM	
			Env	*P* -value	Env	BF	Env	*P* -value
1. Hapmap41074-BTA-73520	5	48.35	prec7	48.35×10^{-47}	tmin10	136		
			latitude	1.41×10^{-43}	bio9	89.7		
			bio7	6.07×10^{-43}	prec6	74.2		
2. ARS-BFGL-NGS-113888	5	48.32	prec7	4.86×10^{-47}	tmin10	39.3		
			latitude	1.06×10^{-43}	bio9	27.6		
			bio7	1.26×10^{-42}	prec6	24.9		
3. Hapmap41762-BTA-117570	5	18.94	prec7	2.74×10^{-44}	bio9	15.3		
			latitude	3.95×10^{-41}	prec6	13.3		
			prec6	4.95×10^{-37}	prec5	12.6		
4. ARS-BFGL-NGS-46098	20	2.95	prec7	2.94×10^{-44}				
			latitude	2.58×10^{-39}				
			prec6	4.35×10^{-39}				
5. BTA-73516-no-rs	5	48.75	prec7	2.51×10^{-39}	bio9	12.8		
			latitude	4.57×10^{-36}	prec6	11.8		
			prec6	7.61×10^{-33}	prec5	11.5		
6. Hapmap41813-BTA-27442	5	49.04	prec7	6.06×10^{-39}	bio9	16.7		
			latitude	7.37×10^{-36}	prec6	15.3		
			prec6	2.26×10^{-32}	prec5	14.9		
7. Hapmap28985-BTA-73836	5	70.34	bio3	6.98×10^{-36}	bio9	12.5	bio3	4.01×10^{-19}
			prec6	1.18×10^{-35}	prec6	11.5	bio7	3.94×10^{-14}
			bio7	1.61×10^{-33}	prec5	11.1	latitude	6.63×10^{-10}

Loci	Chr.	Pos (Mbp)	SAMβADA		BAYENV		LFMM	
			Env	*P* -value	Env	BF	Env	*P* -value
8. ARS-BFGL-NGS-106520	5	70.2	bio3	6.26×10^{-35}	tmin10	79.5	bio3	3.61×10^{-17}
			bio7	3.55×10^{-33}	bio9	23.3	bio7	1.18×10^{-12}
			latitude	1.13×10^{-31}	prec6	18.7	prec6	2.03×10^{-10}
9. BTA-73842-no-rs	5	70.18	bio3	8.95×10^{-34}	bio9	13.4	longitude	3.19×10^{-15}
			bio7	2.64×10^{-30}	prec6	11.3	prec6	1.35×10^{-9}
			latitude	4.13×10^{-30}	prec5	10.7	bio15	2.55×10^{-9}
10. Hapmap31863-BTA-27454	5	48.99	prec7	1.08×10^{-33}				
			latitude	3.00×10^{-30}				
			prec6	3.26×10^{-27}				
11. Hapmap50523-BTA-98407	5	46.74	prec7	6.36×10^{-32}	bio9	14.4		
			prec6	7.61×10^{-28}	prec6	12.8		
			latitude	9.69×10^{-28}	prec5	12.3		
12. BTB-01400776	20	2.7	prec7	4.71×10^{-31}				
			latitude	5.23×10^{-30}				
			prec6	1.65×10^{-25}				
13. ARS-BFGL-NGS-10586	2	128.64	latitude	9.47×10^{-29}	bio9	11.5		
			bio7	1.73×10^{-25}	prec6	10.1		
			prec7	1.81×10^{-25}				
14. Hapmap23956-BTA-36867	15	47.2	latitude	1.59×10^{-28}	bio9	23.1		
			prec7	2.17×10^{-26}	prec6	20		
			prec6	8.85×10^{-25}	prec5	19		
15. ARS-BFGL-NGS-94862	11	103.53	longitude	1.23×10^{-27}	bio9	45.6	longitude	9.52×10^{-10}
			prec7	1.26×10^{-22}	prec6	42.1		
			latitude	4.26×10^{-20}	prec5	40.8		
16. BTA-122374-no-rs	14	16.44	latitude	1.97×10^{-27}				
			prec7	1.05×10^{-23}				
			prec11	1.26×10^{-23}				

Loci	Chr.	Pos (Mbp)	SAMβADA		BAYENV		LFMM	
			Env	*P* -value	Env	BF	Env	*P* -value
17. ARS-BFGL-NGS-43694	5	49.65	prec7	8.16×10^{-27}				
			latitude	3.41×10^{-25}				
			prec6	5.93×10^{-24}				
18. BTB-01356178	20	2.49	latitude	1.49×10^{-26}	tmin10	62.7		
			prec7	6.28×10^{-26}	bio9	33		
			prec6	6.69×10^{-23}	prec6	27.9		
19. BTA-108359-no-rs	14	16.31	longitude	2.35×10^{-26}				
			prec7	3.87×10^{-26}				
			prec11	6.28×10^{-25}				
20. ARS-BFGL-NGS-15960	5	28.02	prec7	3.20×10^{-26}	bio9	76.8		
			prec6	7.57×10^{-24}	prec6	74.1		
			longitude	1.78×10^{-23}	prec5	72.9		
21. ARS-BFGL-NGS-116294	2	128.58	latitude	6.05×10^{-26}	tmin10	43		
			prec7	3.34×10^{-23}	bio9	18		
			bio7	6.44×10^{-23}	prec6	15.2		
22. Hapmap52789-rs29018750	5	70.26	bio7	1.05×10^{-25}				
			bio3	1.32×10^{-24}				
			latitude	1.08×10^{-23}				
23. ARS-BFGL-NGS-86183	8	43.5	prec7	4.73×10^{-25}				
			prec6	1.27×10^{-21}				
			latitude	3.35×10^{-21}				
24. ARS-BFGL-NGS-16554	20	1.44	bio7	1.18×10^{-24}	tmin10	55.4		
			prec7	1.27×10^{-24}	bio9	15.2		
			latitude	4.91×10^{-23}	prec6	12.7		
25. ARS-BFGL-NGS-30091	22	47.94	longitude	1.25×10^{-24}				
			prec7	3.08×10^{-14}				
			tmax10	3.63×10^{-14}				

SAMβADA's multivariate analysis identified 12 significant bivariate models, corresponding to 8 loci (see Table S2, Supporting information). In SAMβADA's framework, this means that these models involving one environmental variable and the variable 'population structure' provided a significantly more accurate estimation of the genotype's frequency than their univariate parent involving the variable 'population structure' only. Therefore, although population structure might partly explain the distribution of these genotypes, adding an environmental variable provided a significantly more accurate estimation of their distribution ($\alpha' = 5.9 \times 10^{-9}$). The loci detected by SAMβADA's multivariate analysis include three loci that were detected by all correlative approaches (Hapmap28985-BTA-73836, ARS-BFGL-NGS-106520 and BTA-73842-no-rs, see lines 7, 8 and 9 in Table 2).

Computation time was measured for the three correlative approaches using a desktop computer with 8-core CPUs at 4.0 GHz and 16 Gb of RAM, except for BAYENV, which used a slightly less powerful computer (8-core CPU at 3.1 GHz and 8 Gb of RAM). SAMβADA analysed the univariate models within 1.5 h using a single processing thread and both univariate and bivariate models in 2.6 h using four threads. LFMM analysed the data set in 26.9 h for each value of *K* using five threads (one per run) and BAYENV in 41.3 h with a single thread, for one run. Ratios between computation times tend to increase with larger data sets (see Table S7, Supporting information).

3.2.2 Spatial autocorrelation

Global and local indicators of spatial autocorrelation were computed for two genotypes with a weighting scheme based on the 20 nearest neighbours and a pseudo *P* -value threshold of 1%: (i) ARS-BFGL-NGS-46098 (genotype GG) (hereafter ARS-46 (GG)), which was detected by SAMβADA only with one of the highest *G* scores (Table 2, line 4), and (ii) Hapmap28985-BTA-73836 (genotype GG) (hereon HM-28 (GG)), which was detected by SAMβADA while the corresponding locus HM-28 was detected by BAYENV and LFMM (Table 2, line 7). SAMβADA identified isothermality, the stability of temperature across the year, as strongly associated with both genotypes. Figure 3 shows local indices of spatial autocorrelation for these two genotypes. On the one hand, ARS-46 (GG) was positively autocorrelated for the majority of points and the index was significant for half of them. Although the distribution of this genotype shows spatial dependence, nonsignificant associations were found at the edge of Lake Victoria and in a corridor in the North of the Lake with some occurrences in the West of Uganda. On the other hand, the local indices of spatial association of HM-28 (GG) showed lower values in general and were only significant in the northwest of Uganda. This particular region also showed the lowest values of isothermality in Uganda, that is a high variability of temperatures. This correlation between HM-28 (GG) and isothermality also appeared with bivariate LISAs, where the presence of the genotype was compared with the mean value of isothermality among neighbouring points (not shown).

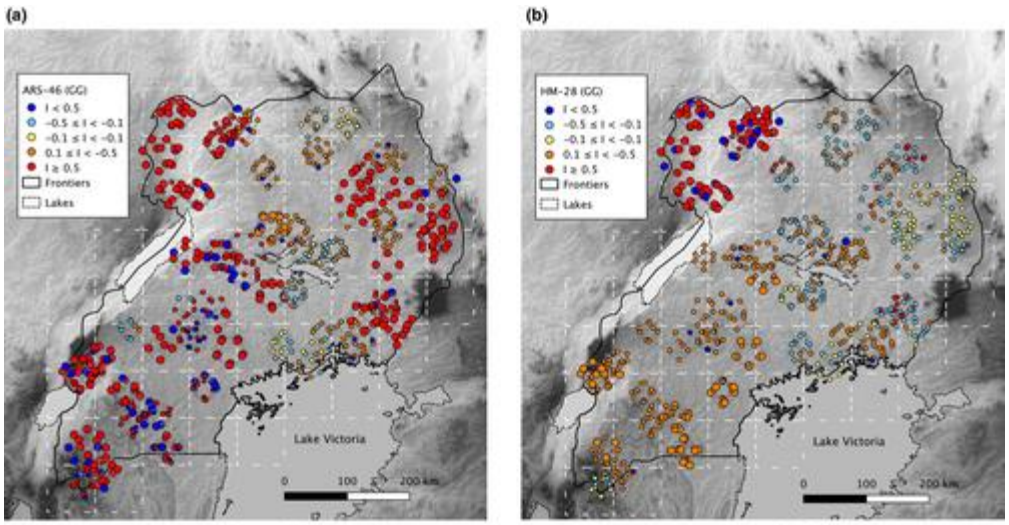


Figure 3: Local indicators of spatial association of markers ARS-46 (genotype GG) and HM-28 (genotype GG). The weighting scheme is based on the 20 nearest neighbours. Red points tend to be similar to their neighbours, while blue points differ from them. Yellow points are independent from their neighbourhood. Small points indicate nonsignificant values ($P > 0.001$). The map in the background represents the relief, the darker the shade, the higher the altitude. Samples coming from the same farm have been spread on a circle around their actual location.

4 Discussion

The main features of SAMβADA are the processing speed, the multivariate modelling and the measurement of spatial autocorrelation. Processing speed is key when dealing with high-throughput data, while multivariate modelling and spatial autocorrelation measurements improve the interpretation of results, particularly when the data set includes population structure. Models may indeed include the global ancestry coefficients provided by a preliminary analysis (e.g. ADMIXTURE). This facilitates the detection of genotypes correlated with the environment while taking population structure into account. Additionally, introducing measurements of spatial autocorrelation into these analyses takes into account the valuable contribution of spatial statistics in landscape genomics. Unlike most current and nonspatial approaches (e.g. Coop *et al* . 2010; Frichot *et al* . 2013; Frichot & François 2015), SAMβADA allows the determination of whether the observed data reflects independent samples, a requirement of the underlying statistical model. Spatial autocorrelation measurements help assess whether the occurrence of a genotype is related to its frequency in the surrounding locations. More specifically, local indices of spatial autocorrelation allow the mapping of areas prone to spatial dependence. The results of the present analysis show that using spatial statistics in conjunction with correlative models may lower the risk of false positives in landscape genomics. This is important when the individuals under study share demographic history (e.g. individuals within breeds of a livestock species – Orozco-terWengel *et al* . 2015 – or absence of gene flow in a divergence-after-speciation model configuration – Cruickshank & Hahn 2014), in the presence of isolation by distance (Meirmans 2012) or cryptic relatedness (Corbett-Detig *et al* . 2015), and when genetic background are ignored (François *et al* . 2016). However, while some population structures do not show significant spatial autocorrelation, one has to keep in mind that particular demographic

structures may totally mimic selection signatures (Holderegger *et al* . 2008) and that in this case, correlative approaches are not able to recognize the cause of the spatial pattern observed. SAMβADA can analyse such cases with the multivariate models including the global ancestry coefficients.

4.1 Simulation study

The simulation study shows that SAMβADA univariate models and LFMM are able to detect the locus under selection in discrete, low-agglomerated landscapes, provided that the strength of selection is high enough. In the weak selection scenario, the mortality at birth is compensated by the dispersal of individuals in approximately half the replicates, so that the locus under selection is not detected. On the contrary, it is only missed once for the intermediate selection strength and is always detected for the strong selection scenario. However, this power of detection comes at the cost of high FPRs. The relatively low dispersal capacity of individuals leads to isolation by distance, so that frequencies of neutral alleles vary across space (Forester *et al*. 2016). This induces some spurious correlations with the '*x* ' and '*y* ' coordinates, used as proxies for continuous gradient-like environmental variables. These false detections affect both the SAMβADA univariate models, which do not correct for population structure, and LFMM, which tries to model it as unobserved variables. Besides their comparable TPR and FPR, LFMM seems to recognize the variable 'habitat' as the driver of selection in more replicates than SAMβADA which tends to assign better scores to models involving '*x* ' or '*y* '. The GEA index of both methods increases with the selection strength, showing that higher selection strengths increase the power of detection and the ability to distinguish the environmental variable driving local adaptation.

SAMβADA's multivariate analysis leads to a considerably lower FPR than the previous methods (2–4% vs. 39–45%). Therefore, including population structure as a set of covariates improves the ability of SAMβADA to distinguish between signals of selection and differences in allelic frequencies due to isolation by distance. In the strong selection scenario, the multivariate models have the same power of detecting the locus under selection as the univariate models. However, the TPR is lower for the intermediate level of selection and very low for the weak selection scenario. Thus, controlling for population structure in multivariate models with a conservative significance threshold (e.g. Bonferroni correction) may decrease the power of detecting loci under weak to moderate selection strengths. These results illustrate the trade-off which exists between the power of detection of correlation-based approaches and the specificity of the said detections obtained by taking the population structure into account.

The analysis of spatial autocorrelation enables the comparison of the locus under selection (L0) to neutral loci detected by SAMβADA (false positives) and neutral loci not detected by SAMβADA (true negatives). False-positive loci tend to have higher values of Moran's *I* than the group of true negative for all selection scenarios (see Fig. 1 and Figs S4–S6, Supporting information, for details). This illustrates the fact that spatial dependency in neutral loci increases their probability of being detected as potentially subject to selection. The spatial autocorrelation of both groups of neutral loci (false-positive and true-negative) stays stable with increasing selection pressure, while the spatial autocorrelation of true positive clearly increases with the selection pressure. The latter effect may be emphasized by the fact that several genotypes are positively selected in distinct habitats and

negatively selected in the other habitats. Therefore, loci with high values of spatial autocorrelation can also be subject to selection and should not be discarded from the analysis on this sole criterion. Local indicators of spatial autocorrelation draw the same picture as the global Moran's *I* : when counting the number of sampling points showing a significant LISA value, the locus under selection is often among the loci showing the most significant LISA points, and this trend also increases with selection pressure (Table S6, Supporting information).

4.2 Ugandan cattle

In the study of Ugandan cattle, SAMβADA detected the highest number of SNPs as potentially subject to selection among the four approaches. However, SAMβADA's detection rate may reflect false positives probably due to population structure. This interpretation is supported by the shape of the quantile–quantile plots, where SAMβADA univariate analysis shows an excess of models with small *P* -values (see Fig. S2, Supporting information, part a). Indeed, the distribution of cattle populations follows roughly a north–south axis which corresponds to the gradient shown by some environmental variables. This overlay may result in some spurious associations. Regardless, environmental conditions can underlie the intensity of some health threats, such as the trypanosomiasis. The two cattle species bore some specific traits before they met in Uganda (e.g. drought tolerance and disease resistance). These specificities have contributed to shape their respective distribution in the country. In this case, the observed genome–environment associations can reflect the local adaptation of cattle in Uganda. Moreover, the discrepancy between the results may indicate that the more conservative approaches induce some false negatives. The zebus are indeed highly admixed with Ankole cattle and only eight of them were retained in the reference population used by BAYENV and ARLEQUIN (compared with 162 Ankole cattle). This difference in sample size may have affected ARLEQUIN's analysis and prevented the detection of selection signatures. Another potential source of discrepancy between approaches is the use of a pre-existing SNP chip to analyse local adaption. Some ascertainment bias could result from the choice of the set of loci as neither Shorthorn zebus nor Ankole cattle were included in the SNP chip development. However, using the observed heterozygosity of both populations as a proxy of the effect of ascertainment bias, we can see that the average observed heterozygosity of Ankole is ~0.27 and that of the one of zebu is ~0.25, largely reflecting that if there is a bias it probably affects both groups similarly. Additional data from the BovineHD Genotyping BeadChip (Illumina Inc., San Diego, CA, USA) suggest that both Ankole and zebu here have similar observed heterozygosity (L. Colli, personal communication).

Comparing these results in the light of spatial dependence gives information about the differences between SAMβADA's, BAYENV's and LFMM's detections. The locus ARS-46 was detected by SAMβADA only, and its genotype GG showed a widespread pattern of spatial autocorrelation (Fig. 3a). This pattern could originate from the underlying population structure, as Ankole cattle are more common in the southwest, while zebus are more common in the northeast of the country. This spatial dependence in the occurrence of this genotype is in contradiction with the assumptions of SAMβADA's statistical model. Thus, the correlation detected by logistic regressions between ARS-46 (GG) and environmental variables could be spuriously driven by demographic factors, as described above. Patterns of spatial dependence for HM-28 presented a different situation (Fig. 3b). The low value of spatial autocorrelation for HM-28 (GG) implies that the distribution of this genotype was mostly independent

of location, thus the logistic models are reliable for this genotype. HM-28 was also detected by the three landscape genomic approaches and by SAMβADA multivariate analysis, and this supports a possible adaptive origin of the observed correlation with isothermality. Maps of local spatial autocorrelation for the genotypes ARS-46 (GG) and HM-28 (GG) illustrated a general trend: BAYENV and LFMM discarded SNPs showing significant local spatial autocorrelation for a large proportion of the sampling locations, while SAMβADA detected them. Thus, in this case, measuring the local autocorrelation of candidate genotypes may help distinguishing between the effects of local adaptation and those of population structure among SAMβADA's detections.

Regarding common detections, three of the SNPs identified by SAMβADA when population structure was included as a covariate were among the common detections of the three correlative approaches. SAMβADA bivariate analysis is rather conservative with only eight detected loci; however the distribution of *P* -values is close to the expected distribution, suggesting that population structure was taken correctly into account (see Fig. S2, part b, Supporting information). Thus, pre-existing knowledge on demography may be built on to refine correlation-based detections of selection signatures. One possible approach consists of assessing population structure and then including one or a few variables summarizing this structure in the constant model used by SAMβADA. In this way, only genotypes showing a significant correlation with the environment while taking the population structure into account are detected. In case there are more than two main populations, hence requiring several variables to summarize the samples' ancestry, these summary variables could for instance be derived from a PCA of the samples' coefficients of ancestry. In the present study, the coefficients of ancestry for the Ankole and zebu populations are essentially complementary for most samples, thus using the first principal axis of the PCA is similar to using one of these coefficients of ancestry as the summary variable.

Concerning the biological function of frequently detected loci, these three loci are located on chromosome 5, near the gene POLR3B whose mouse counterpart is involved in limiting infection by intracellular bacteria and DNA viruses (UniProt, www.uniprot.org). Moreover, genotype HM-28 (GG) shows spatial autocorrelation in the northwestern part of Uganda and this area overlaps with one of those where the highest load of tsetse fly (*Glossina* spp.) occurs in the country (Abila *et al*. 2008; MAAIF *et al*. 2010). Hence, the risk of cattle trypanosomiasis is high in this region and the detected mutations may be involved in parasite resistance.

4.3 Comparison between simulated and empirical data

The analyses of the simulation and cattle data lead to some common observations. SAMβADA's univariate modelling detects some spurious associations in scenarios with population structure. As a countermeasure, multivariate analysis, which includes predictors variables accounting for this population structure, lowers the rate of false positives. However, the assumption that the main axis of molecular variation represents only the population structure may induce some false negatives, especially when the selection pressure is low (simulated data) or when the full data set was used to assess the said population structure (cattle data). The comparison of the two types of data also reveal some differences: the environmental variable 'habitat' which drives selection in the simulation data is discrete with a complex spatial distribution (low-agglomeration), while there are many continuous

environmental variables describing the habitat in Uganda and most of these present a north – south gradient. Another difference is the spatial distribution of individuals: each sample came from a distinct location in the simulation data, while several individuals were sampled at each location in Uganda. These differences may be reflected in the observed patterns of spatial autocorrelation. The simulated data show that molecular markers displaying a high spatial dependence can actually be subject to selection. In fact, as many environmental variables are auto-correlated in nature, it can be expected that the distribution of a molecular marker selected by one of these variables will also present some spatial correlation. Therefore, it is currently not possible to distinguish between true and false positives solely on the basis of their spatial dependence. The most efficient approach involves comparing the results of several methods taking the population structure into account, and to observe the patterns of spatial autocorrelation to analyse how the detected GEAs are linked to the spatial distributions of markers and environmental variables.

4.4 Perspectives

The increasing availability of large molecular data sets raises challenges regarding their analysis. Correlative approaches in landscape genomics enable fast detection of candidate loci to local adaptation. However, these methods must take into account the effect of population structure (De Mita *et al*. 2013; Frichot *et al* . 2013; Joost *et al*. 2013; Frichot & François 2015). Limited dispersal of individuals leads to spatial autocorrelation of marker frequencies, which may cause spurious correlations with the environment. SAMβADA addresses these issues by rapidly detecting selection signatures with the possibility of including prior knowledge of the population structure in the analysis and by measuring the level of spatial autocorrelation for candidate loci. The next methodological step involves developing spatially explicit models that directly include autocorrelation. SGLMM (Guillot *et al*. 2014) provides such a model; however, the current R-based implementation does not enable whole-genome analysis.

The recent availability of whole-genome sequence (WGS) data also raises issues regarding the statistical assessment of multiple comparisons. Indeed, while many individuals and few genetic markers were available 10 years ago, the current high costs of WGS limit the number of sequenced samples. Therefore, standard procedures for multiple comparisons, such as the Bonferroni correction, are over-conservative and may lead to discard some adaptive loci. In this context, alternatives procedures focus on controlling the ratio of false positives in a set of significant results. Among them, Storey and Tibshirani's false discovery rate (2003) was especially designed for large molecular data sets and suits any detection method relying on significance tests. This method is available as an R package (*q* value, Storey *et al.* 2015) and its implementation in SAMβADA is ongoing.

Computation time is critical when processing large data sets. In this context, SAMβADA is able to swiftly analyse high-density SNP-chips and variants from WGS. When taking population structure into account, SAMβADA's multivariate analysis is approximately 10 times quicker than LFMM and 16 times than BAYENV for a data set comparable to this study, and these ratios increase with larger data sets (see Table S7, Supporting information). SAMβADA's simple underlying model has the advantage that the computation time grows linearly with the size of the genetic data under study. Therefore, SAMβADA's module for parallelized processing enables the analysis of WGS data sets on

desktop computers (see Table S9, Supporting information). SAMβADA's processing speed, combined with its ability to analyse the spatial autocorrelation in molecular data and to incorporate prior knowledge on population structure, suits a wide range of applications, especially those involving whole-genome sequence data.

4.5 Acknowledgements

We thank Sergio Rey for his advice on assessing the significance of LISA, Stephan Morgenthaler for fruitful discussions on assessing the significance of multivariate logistic models, Olivier François and Eric Frichot for their explanations on LFMM and Gilles Guillot for providing us with SGLMM for testing purposes. We thank Kevin Leempoel for his help in analysing the spatial autocorrelation and Estelle Rochat for her careful reading and useful comments on the manuscript.

4.6 Funding

This research was funded by EU FP7 project NextGen (Grant KBBE-2009-1-1-03).

4.7 Resources

4.7.1 Software availability

sam βada is an open source software written in C++ available at http://lasig.epfl.ch/sambada (under the license GNU GPL 3). Compiled versions are provided for Windows, Linux and MacOS X.

4.7.2 Data availability

NextGen data are described at http://projects.ensembl.org/nextgen/. Ugandan cattle SNP data are available at ftp://ftp.ebi.ac.uk/pub/databases/nextgen/bos/variants/chip_array/ in PLINK format (files UGBT.bovineSNP50.UMD3_1.20140307.[ped/map].gz) with the following data policy ftp://ftp.ebi.ac.uk/pub/databases/nextgen/documentation/README_data_use_policy.
Simulation data, landscape surfaces and individual sample files are available at Dryad doi:10.5061/dryad.v0c77.

5 Authors contribution

P.T., S.J., M.B., L.C. and R.N. designed research. S.S., P.O.T.W., L.C., S.J., B.F., C.M., R.N. and S.D. performed research. S.S., S.J. and P.O.T.W. contributed to new analytical tools. S.S., S.J., P.O.T.W., B.F., S.D., M.J. and E.L. wrote and reviewed the manuscript. All the authors undertook revisions, contributed intellectually to the development of this manuscript and approved the final manuscript.

6 References

Abila PP, Slotman MA, Parmakelis A *et al* . (2008) High levels of genetic differentiation between Ugandan *Glossina fuscipes fuscipes* populations separated by Lake Kyoga. *PLOS Neglected Tropical Diseases*, **2**, e242.

Ajmone Marsan P, Garcia JF, Lenstra JA, the Globaldiv Consortium (2010) On the origin of cattle: how aurochs became cattle and colonized the world. *Evolutionary Anthropology*, **19**, 148– 157.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655– 1664.

Anselin L (1995) Local Indicators of Spatial Association – LISA. *Geographical Analysis*, **27**, 93– 115. GISDATA (Geographic Information Systems Data) Specialist Meeting on GIS (Geographic Information Systems) and Spatial Analysis, Amsterdam, Netherlands, Dec 01–05, 1993.

Anselin L, Syabri I, Kho Y (2006) Geoda: an introduction to spatial data analysis. *Geographical Analysis*, **38**, 5– 22.

Arguez A, Vose RS (2010) The definition of the standard WMO climate normal: the key to deriving alternative climate normals. *Bulletin of the American Meteorological Society*, **92**, 699– 704.

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301– 302.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969– 980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B: Biological Sciences*, **263**, 1619– 1626.

Bivand R, Piras G (2015) Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, **63**, 1– 36.

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3– 62.

Colli L, Joost S, Negrini R *et al* . (2014) Assessing the spatial dependence of adaptive loci in 43 European and Western Asian goat breeds using AFLP markers. *PLoS One*, **9**, e86668.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411– 1423.

Corbett-Detig RB, Hartl DL, Sackton TB (2015) Natural selection constrains neutral diversity across a wide range of species. *PLoS Biology*, **13**, 1– 25.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133– 3157.

De Mita S, Thuillet A-C, Gay L *et al* . (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383– 1399.

Dobson AJ, Barnett AG (2008) *An Introduction to Generalized Linear Models*, 3rd edn. Chapman & Hall, Boca Raton, FL.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564– 567.

Farr TG, Rosen PA, Caro E *et al* . (2007) The shuttle radar topography mission. *Reviews of Geophysics*, **45**, RG2004.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977– 993.

Forester BR, Jones MR, Joost S, Landguth EL, Lasky JR (2016) Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Molecular Ecology*, **25**, 104– 120.

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454– 469.

Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, **6**, 925– 929.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687– 1699.

Gardner RH (1999) *RULE: Map Generation and a Spatial Analysis Program*, pp. 280– 303. Springer, New York, NY.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555– 1579.

GDAL Development Team (2013) *GDAL – Geospatial Data Abstraction Library, Version 1.10*. Open Source Geospatial Foundation, Beaverton, Oregon.

correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spatial Statistics*, **8**, 145– 155.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205– 220.

Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics*, **7**, 1– 32.

Henry P, Russello MA (2013) Adaptive divergence along environmental gradients in a climate-change-sensitive mammal. *Ecology and Evolution*, **3**, 3906– 3917.

Hijmans R, Cameron S, Parra J, Jones P, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965– 1978.

Holderegger R, Herrmann D, Poncet B *et al* . (2008) Land ahead: using genome scans to identify molecular markers of adaptive relevance. *Plant Ecology & Diversity*, **1**, 273– 283.

Jeffreys H (1961) *The Theory of Probability*, 3rd edn. Oxford University Press, Oxford, UK.

Jones MR, Forester BR, Teufel AI *et al* . (2013) Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinical populations. *Evolution*, **67**, 3455– 3468.

Joost S, Bonin A, Bruford MW *et al* . (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955– 3969.

Joost S, Kalbermatten M, Bonin A (2008) Spatial Analysis Method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources*, **8**, 957– 960.

Joost S, Vuilleumier S, Jensen JD *et al* . (2013) Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Molecular Ecology*, **22**, 3659– 3665.

Landguth EL, Cushman SA (2010) CDPOP: a spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, **10**, 156– 161.

Legendre P (1993) Spatial autocorrelation – trouble or new paradigm? *Ecology*, **74**, 1659– 1673.

Lv F-H, Agha S, Kantanen J *et al* . (2014) Adaptations to climate-mediated selective pressures in sheep. *Molecular Biology and Evolution*, **31**, 3324– 3343.

Meirmans PG (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**, 2839– 2846.

Ministry of Agriculture, Animal Industry and Fisheries, Uganda, Uganda Bureau of Statistics, Food and Agriculture Organization of the United Nations, International Livestock Research Institute, and World Resources Institute (2010) *Mapping a Better Future: Spatial Analysis and Pro-Poor Livestock Strategies in Uganda*, pp. 30– 37. World Resources Institute, Washington, DC and Kampala.

Mitton JB, Linhart YB, Hamrick JL, Beckman JS (1977) Observations on genetic structure and mating system of ponderosa pine in Colorado front range. *Theoretical and Applied Genetics*, **51**, 5– 13.

Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17– 23.

Ndumu DB, Baumung R, Hanotte O *et al* . (2008) Genetic and morphological characterisation of the Ankole Longhorn cattle in the African Great Lakes region. *Genetics Selection Evolution*, **40**, 467– 490.

Orozco-terWengel P, Barbato M, Nicolazzi EL, Biscarini F, Milanesi M (2015) Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Frontiers in Genetics*, **6**, 191.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, 2074– 2093.

Pemstein D, Quinn KM, Martin AD (2011) The Scythe statistical library: an open source C++ library for statistical computation. *Journal of Statistical Software*, **42**, 1– 26.

Purcell S, Neale B, Todd-Brown K *et al* . (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**, 559– 575.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rey SJ, Anselin L (2010) PySAL: A python library of spatial analytical methods. In: *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications* (eds M Fischer, A Getis), pp. 175– 193. Springer, Berlin.

Shaffer JP (1995) Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561– 584.

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, **10**, 199– 228.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440– 9445.

Storey JD with contributions from Bass AJ, Dabney A, Robinson D (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.4.2. http://github.com/jdstorey/qvalue

de Villemereuil P, Gaggiotti O (2015) A new F_{ST} method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, **6**, 1248– 1258.

Vitalis R, Dawson K, Boursot P, Belkhir K (2003) DetSel 1.0: a computer program to detect markers responding to selection. *Journal of Heredity*, **94**, 429– 431.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97– 120.

Appendix C
Supporting information for article in chapter 3

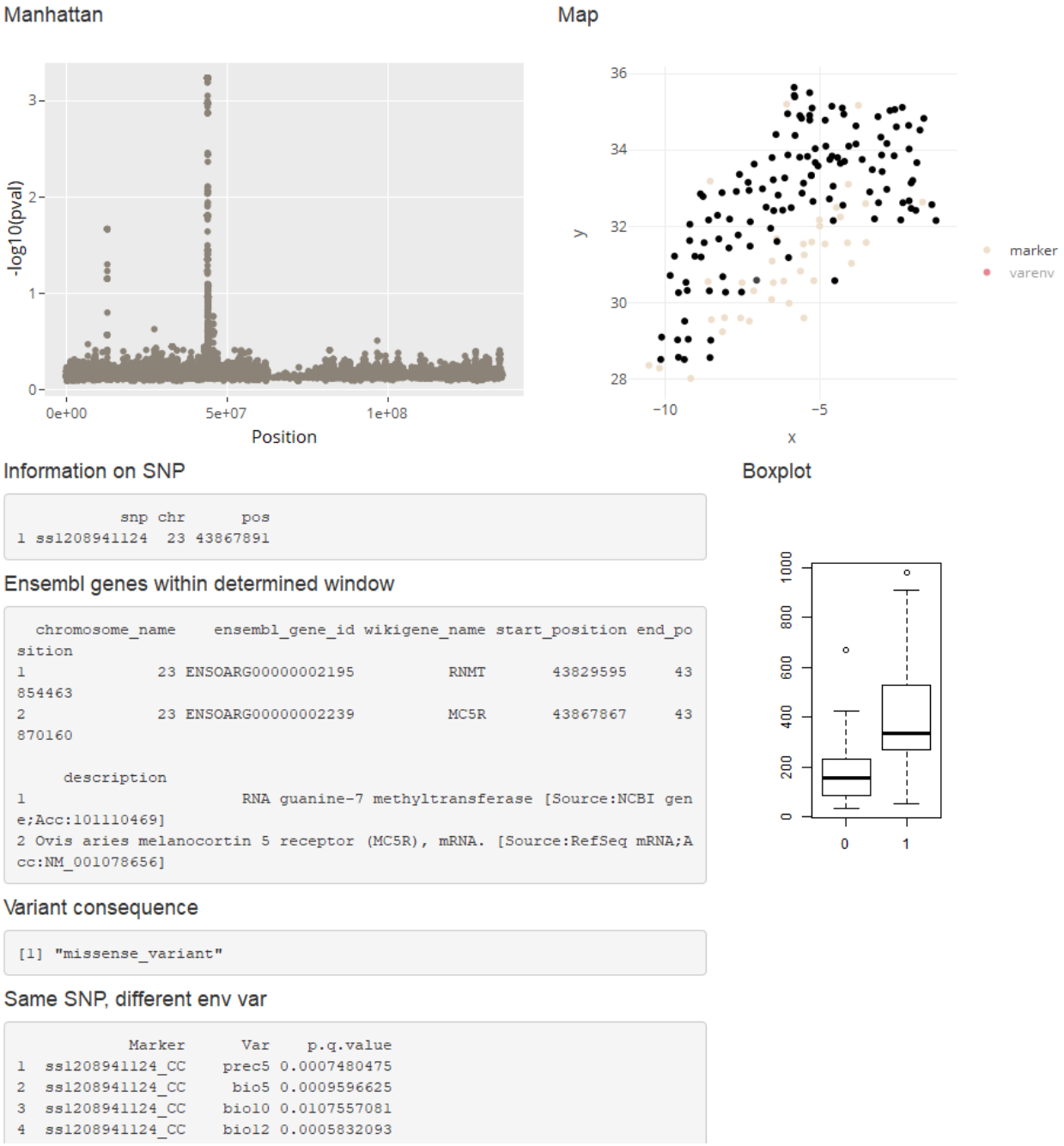


Figure S1: Interactive plot as a result of the function plotResultInteractive. The manhattan plots the q-value of SNPs included in the chromosome 23 of the sheep dataset. In this graph, the genotype ss1208941124_CC was selected, so as to show information on the SNP (position, nearby genes, variant consequence). A map of the genotype (or environmental variable if selected) is available, as well as a boxplot showing the distribution of the environmental variable for the two classes of individuals (absence versus presence of genotype CC). Finally, the window also displays the results of the same SNP included in other models.

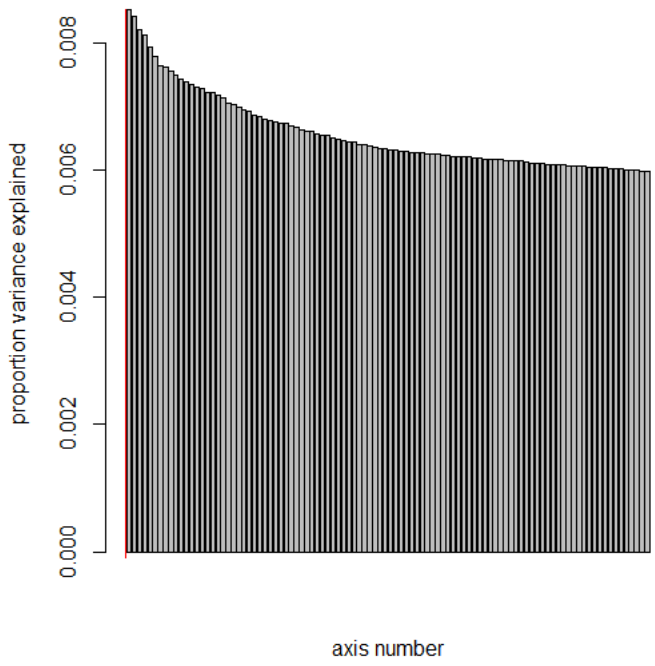


Figure S2: Proportion of variance explained for the first 100 axes of the PCA on molecular markers of Moroccan sheep. The red line indicates how much axes were retained as population variables (in this case none).

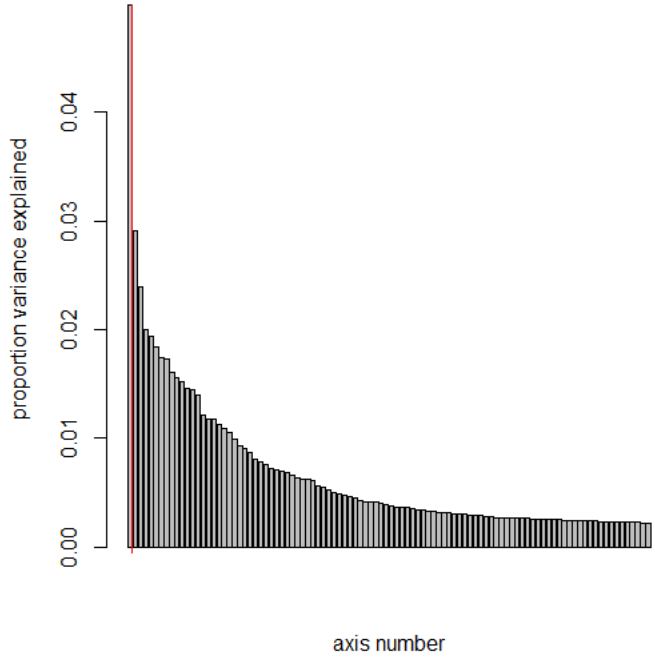


Figure S3: Proportion of variance explained for the first 100 axes of the PCA on molecular markers of the Spanish cattle. The red line indicates how much axes were retained as population variables (in this case only one axis).

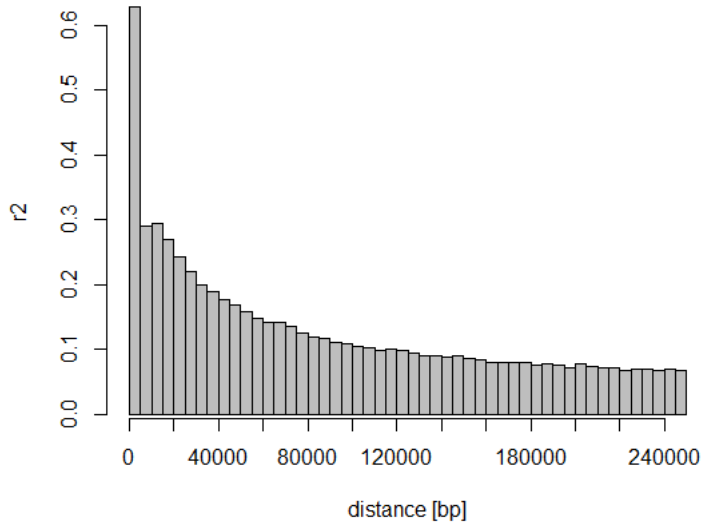


Figure S4: Average Linkage Disequilibrium (LD) per distance class in the Spanish cattle molecular data. This plot is useful to determine if the initial SNP density is sufficient to capture most of the selection signatures. Since the *Bos taurus* genome entails around 3 billion bp and since our SNP chip contains 50k genetic variations, any SNP in the genome will be within an average distance of 30000 bp to a genotyped SNP. The corresponding average LD in this graph for such a distance is 0.2, which can be considered as a sufficiently high value to reach a good coverage of the genome.

Table S1: List of genes and corresponding description within a window of 25000 bp around the two major peaks of the manhattant plot on Moroccan sheep (Figure III.2, results from models involving SNPs on the 23rd chromosome and bio12 – annual precipitation -).

Gene Ensembl ID	Wiki gene	Description
ENSOARG00000005159	LOC101104705	plasminogen activator inhibitor 2-like
ENSOARG00000021375		5S ribosomal RNA
ENSOARG00000002195	LOC114118402	RNA guanine-7 methyltransferase
ENSOARG00000002195	RNMT	RNA guanine-7 methyltransferase
ENSOARG00000002239	MC5R	Ovis aries melanocortin 5 receptor (MC5R), mRNA
ENSOARG00000003950	MC2R	melanocortin 2 receptor

Table S2: List of genes and corresponding description within a window of 50000 bp around SNPs belonging to major peaks of the manhattan plot on Spanish cattle (Figure III.5, investigating associations between SNPs on all chromosomes and bio1 - mean annual temperature -). Note that some peaks are visible on the manhattan plot but have no corresponding entry in this table since SNPs are not necessarily located next to a gene.

Chr #	Gene ID	Wiki gene	Description
2	ENSBTAG00000051630	-	-
6	ENSBTAG00000015649	TMEM156	transmembrane protein 156
6	ENSBTAG00000006044	KLHL5	Kelch like family member 5
6	ENSBTAG00000004988	CCKAR	Cholecystokinin A receptor
8	ENSBTAG00000033396	C8H9orf135	chromosome 8 C9orf135 homolog
10	ENSBTAG00000025634	FMN1	Formin 1
11	ENSBTAG00000053165	-	-
11	ENSBTAG00000013290	DYSF	dysferlin
11	ENSBTAG00000009242	ZNF638	zinc finger protein 638
11	ENSBTAG00000007689	LPIN1	lipin
15	ENSBTAG00000011578	CD44	CD44 molecule
15	ENSBTAG00000004282	AMBRA1	autophagy and beclin 1 regulator 1
15	ENSBTAG00000004639	HARBI1	harbinger transposase derived 1
15	ENSBTAG00000017325	ATG13	autophagy related 13
15	ENSBTAG00000014513		
15	ENSBTAG00000046527	LOC100139830	olfactory receptor 1052
16	ENSBTAG00000002266	NPL	N-acetylneuraminate pyruvate lyase
16	ENSBTAG00000019821	DHX9	DExH-box helicase 9
16	ENSBTAG00000000793	LAMC2	laminin subunit gamma 2
16	ENSBTAG00000049987	-	-
17	ENSBTAG00000007988	STX2	syntaxin 2
17	ENSBTAG00000009797	RIMBP2	RIMS binding protein 2
17	ENSBTAG00000055033	RAN	RAN, member RAS oncogene family
17	ENSBTAG00000001303	HSPB8	heat shock protein family B (small) member 8
23	ENSBTAG00000025718	HMGCLL1	3-hydroxymethyl-3-methylglutaryl-CoA lyase like 1
23	ENSBTAG00000005888	MDGA1	MAM domain containing glycosylphosphatidylinositol anchor 1
26	ENSBTAG00000019759	IDE	insulin degrading enzyme
26	ENSBTAG00000009383	KIF11	kinesin family member 11
27	ENSBTAG00000003509	PLEKHA2	pleckstrin homology domain containing A2
27	ENSBTAG00000001463	TNKS	tankyrase
28	ENSBTAG00000009587	RNG3	Neuregulin 3
28	ENSBTAG00000054349		
28	ENSBTAG00000047155	C28H10orf71	chromosome 28 C10orf71 homolog
28	ENSBTAG00000011991	DRGX	dorsal root ganglia homeobox

Appendix D
Summary statistics of alped Braunvieh cows

Data: This report is the first step in the analysis of a rich dataset collected on the Braunvieh cattle breed describing several aspects of cows. It is an extract of the whole database described in Datenschnittstelle. Only lactating alped cows are included in this extract, with measures posterior to 2000, from breed Braunvieh (BV) or Original Braunvieh (OBV). The available tables are the following:

- B01: Information on the farm location
- K01: General information on the cow, its pedigree
- K04: Global information on lactations (estimated milk quantity and parameters, duration)
- K07: Traits from linear description (length, height, udder)
- K33: Individual milk records (milk quantity and parameters, date, location)

Goals: This small report has three goals:

1) Look at the distribution of different variables to determine reasonable thresholds to apply for quality control. In particular, the following criteria will be considered
   - Is the lactation complete?
   - Is the total milk yield of the lactation within expectable range?
   - Is the interval between two lactations long enough?
   - Are phenotypes known for those cows?
   - Are all cows alped?
   - Is the interval between calving and first milk record within acceptable range?
   - Is the interval between the birth of the cow and its first calf within expectable range?
2) Once determined, these parameters will be applied to discard erroneous data and biased outliers. Some descriptive statistics about the dataset will then be provided.
   - Number of cows per year
   - Distribution of milk record parameter (milk yield, protein content...)
   - Comparison between first, second, third and more lactation

1 Quality control

General numbers

Table 1 specifies the total number of cows, lactations and milk records present in the database before applying any filters.

Table 1: Data size before applying filters

Category	Number
Cows	245’313
Lactations	616’081
Records	5’681’498

Incomplete years

The dataset was extracted from a general database in the beginning of the year 2017 (and a complement was sent in 2018 for the breed "Original Braunvieh"). As a result, the year 2016 is not complete because some data were not already entered in the database. There are 41’177 lactations for which calving happened after March 2015 (this cut-off leaves 270 days before the end of the year, so that we can already take these lactation into account in the analysis).

Furthermore, the datasets contains data from 2000 onwards, but the calvings of 1999 are not in the dataset. As a result, the year 2000 is incomplete. Lactations for which calving took place before August 2000 should be discarded. This represents 13’779 lactations.

Decisions to make:

1) Should we remove year 2000. **Yes, <08.2000**
2) Should we remove year >2015. **Yes, >04.2015**

Days in milk

In order to have a global understanding of lactations, it is important to consider only lactations with the right length. Commonly-agreed standard lactations have a length between 207 and 305 days.

In our data, the number of days in milk ranges from 0 to 2’695. It is useful to notice that, for lactations longer than 305 days, table K04 entails two records: one with the full lactation, and one with a standardised lactation stopped at 305 days.

The total number of lactations available is 616’081. From this number, 13’078 are still run- ning lactations (calf born end of year 2015), but 4’587 of those have already exceeded 305 days. Furthermore, 94’195 lactations have a duration shorter than 270 days. In summary, we have 508’809

finished and longer than 270 days, plus 4’587 non-ended but reaching 305 days, summing up to 83% of the total number of lactations.

While it is important to discard too short lactations, we should also cut-off long ones. Indeed, a cow might lactate much longer if it is not pregnant. If we decided to cut lactations after 500 days, we would discard 35’640 records. Figure 1 shows the days in milk for all finished lactations except those with more than 750 days.

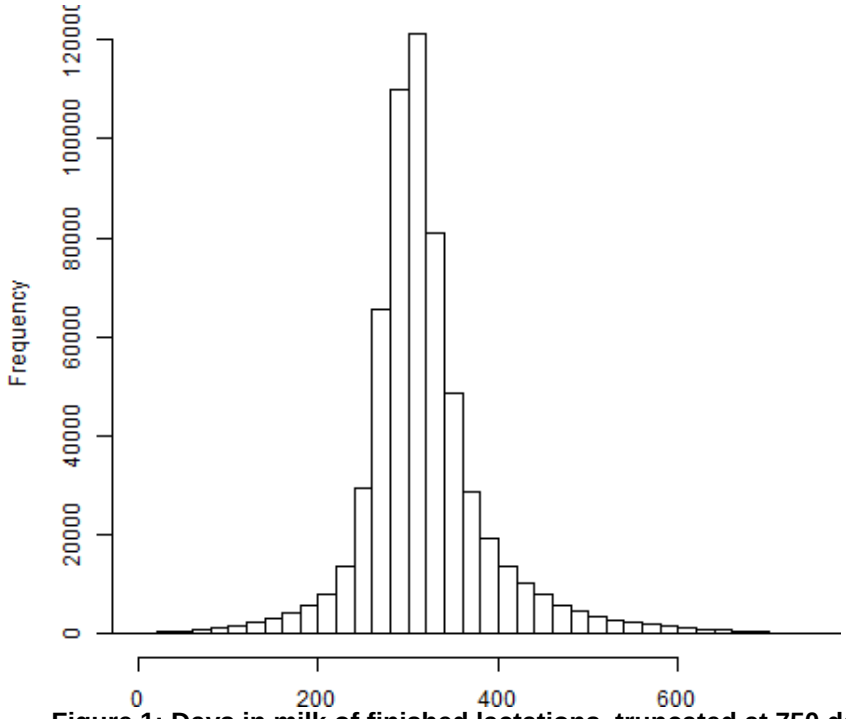


Figure 1: Days in milk of finished lactations, truncated at 750 days

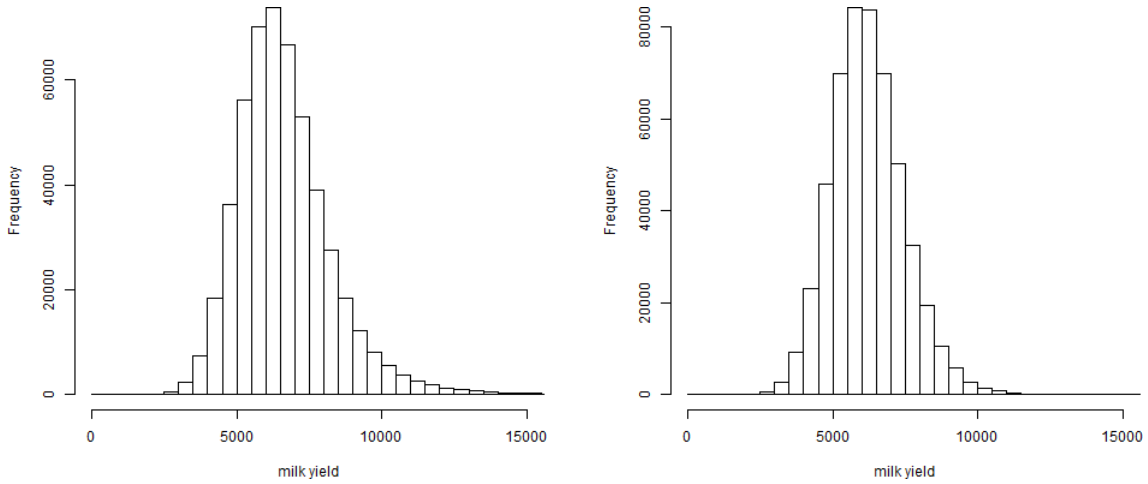
Decisions to make:

1) Should we remove <270 days lactations. **Yes**
2) Should we take full or 305-day standardised lactations. **Depends on the analysis**
3) If we take full lactations should we cut or remove too long ones. Threshold? **Yes. >450- 500?**
4) If we take standardised lactations should we keep running lactations already reaching 305 days. **No. Delete year 2016 - incomplete**

Milk yield

In order to avoid outlier milk yields that might bias our analysis, we should look at the distribution of the milk yield to apply a cut-off.

Figure 2 shows the milk yield obtained during the whole lactations. It is interesting to compare both full and standardised lactations to show the impact of choosing one or the other. As expected, the right tail of the full lactations is a more stretched out, but the rest of the shape is relatively unchanged.



(a) Full lactations (b) Standardised lactations

Figure 2: Milk yield during full lactations (left) and 305-day standardised lactations (right). The full lactations only contain finished ones while the standardised also contains still running lactations. Both graphs only show lactations >270 days

Decisions to make:

1) Should we remove lactations with too low milk yield. Which threshold should we apply. **No. Filter on duration of lactation more relevant**
2) Should we remove or cut lactations with too high milk yield. Which threshold should we apply. **No. Filter on duration of lactation more relevant**

Interval between two calvings

The interval between two calvings should not be too short, because it could imply possible mistakes in the data. However, there are 94 lactations for which the interval between two calvings <270 days, 255 with <290 days and 630 with <305 days.

Figure 3 shows the distribution of this interval. It has also been inspected whether the graph substantially changes between 1, 2 and 3 and higher lactation (graph not shown here). While it is true that the average interval between calving is slightly higher for older cow, the shape of the histogram stays substantially similar.

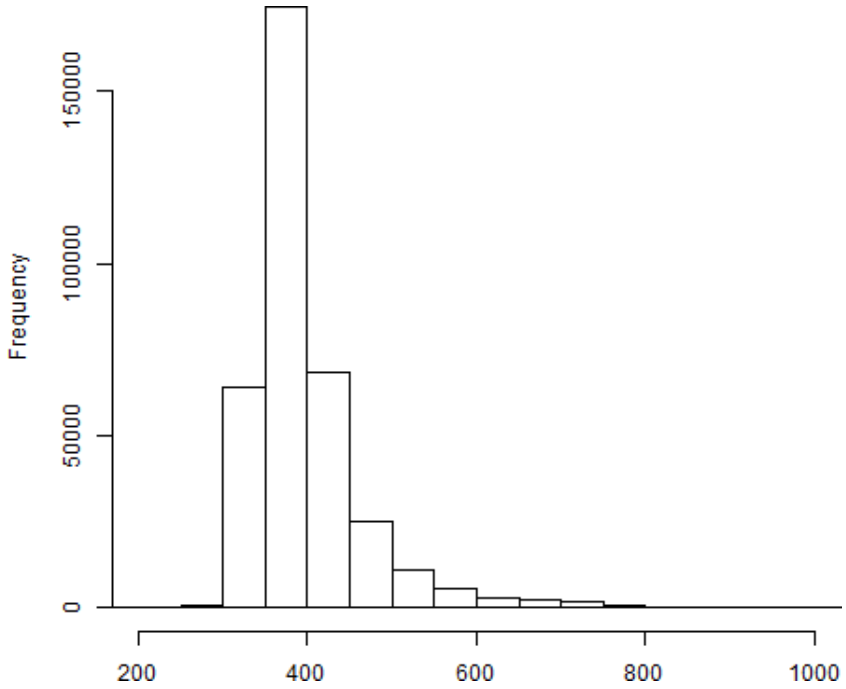


Figure 3: Interval in days between two calvings

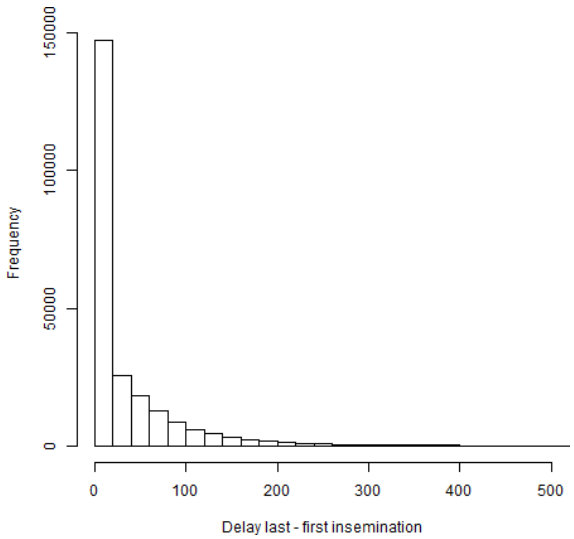
Decisions to make:

1) Should we remove lactations with interval between two calvings too small. Which cut-off should we apply. **Yes, 290 days**
2) Should we remove cows with interval between two calvings too long. Which cut-off should we apply. **No, see fertility**

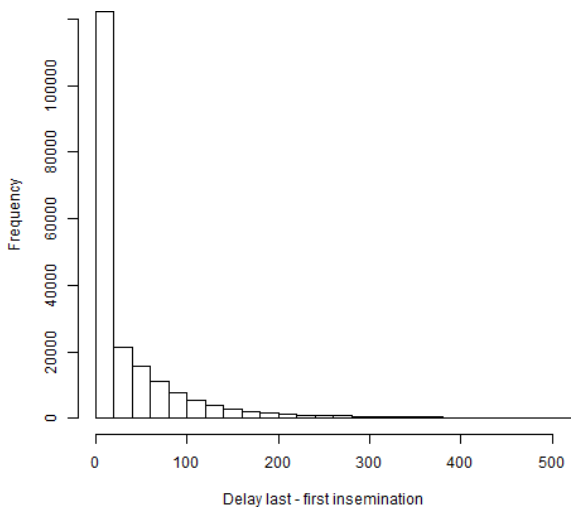
Fertility

Linked to the question of interval between calving and lactation duration is the problem of fertility. Indeed, longer lactations tend to arrise from the fact that a cow could not be successfully inseminated and it took several trials to get it pregnant.

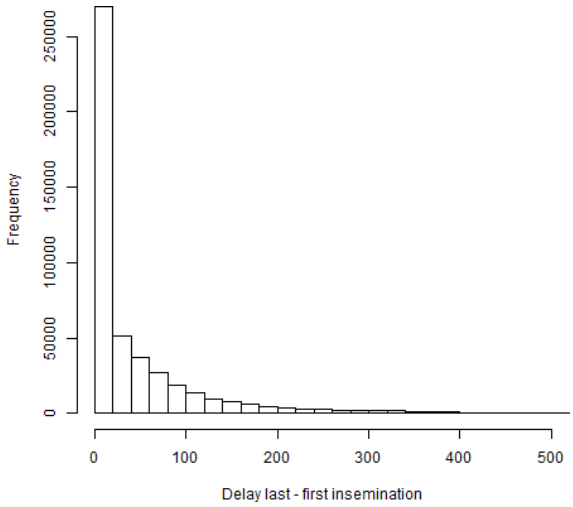
Figure 4 shows the interval between first and last insemination for the 3 groups of lactation.



(b) 1-2 lactation



(b) 2-3 lactation



(c) 3-4 lactation

Figure 4: Interval between first and last insemination, for all 3 lactation group

Decisions to make:

1) Should we remove cows that calved less than 2 years after their birth **Yes**
2) Should we remove cows that had their first calf while being quite old. What cut-off. **Yes, 4 years (1460 days)**

If we accept an average delay of 100 days for the 3 first lactation, we would discard 16’158 cows.

Decisions to make:

1) Should we remove cows whose interval between first and last insemination is too long. Which cut-off? **Yes, 100 (average 3 first lactations)**

Interval between birth and first calf

A cow rarely calves when it is younger than 2 years. Again, if this interval is not satisfied the database might contain an error.

There are 1’087 cows that calved before two years of age, and 365 before 700 days after their own birth.

Additionally, there are 406 cows that had their first calf while being older than 4 years (and 63’106 older than 3 years). In fact this does not necessarily mean that they had their first calf while being that old, since there might be an unregistered calving. However, in the data analysis, we will consider all first lactations together. Older cows might bias the analysis when comparing their lactation parameters with much younger cows.

Figure 5 shows the distribution of the age of the cow at first calving.

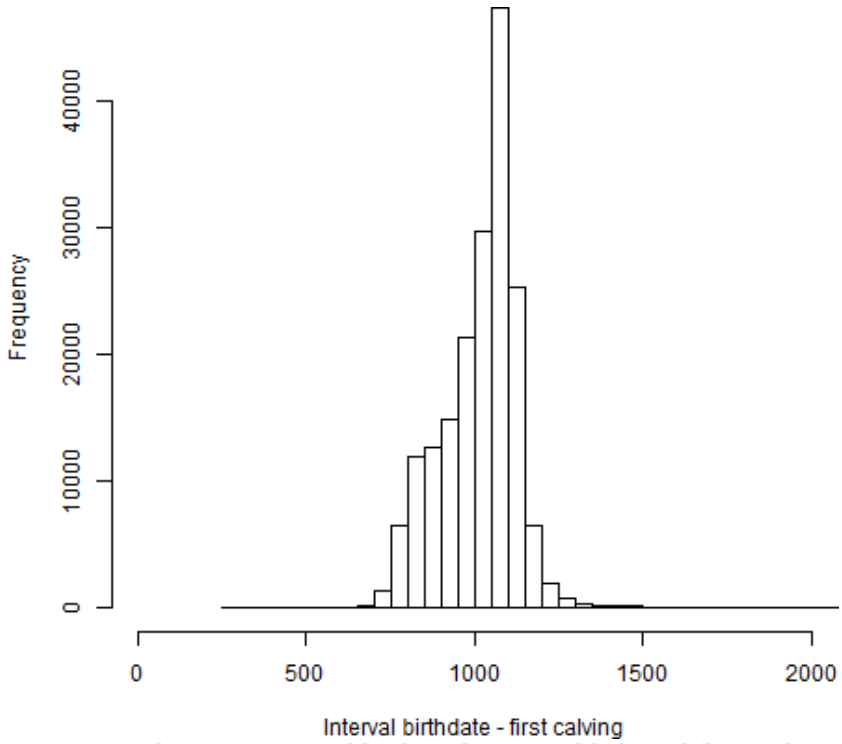


Figure 5: Interval in days between birth and first calf

Known morphological traits

Given that we want to relate lactation characteristics with body phenotypes, we are interested to consider only cows with known morphological traits.

Out of the 243'632 morphological description, 212'058 are taken during the first lactation. Only those values should be accounted for, given that descriptions in subsequent lactations are not comparable with the one in the first lactation. We have 33'277 (16%) cows that do not have any description in their first lactation. In terms of lactation, 73'764 lactations out of the 616'081 lactations (i.e. 12%) would be discarded.

Actually the number of available morphological traits depends on the age of the cow. For the oldest cows it is unknown. For cows born after 1995, the percentage of cows whose morphological description is entirely absent is around 5%. Table 2 shows this evolution.

Table 2: Number of cows with known and unknown morphological traits per year (birth year of the cow)

Year	Known pheno	Unknown pheno	Percent unknown
1979	0	1	100
1981	0	4	100
1983	0	8	100
1984	0	50	100
1985	7	76	91.6
1986	14	170	92.4
1987	39	359	90.2
1988	87	615	87.6
1989	499	1055	67.9
1990	676	2210	76.6
1991	1610	3659	69.4
1992	4852	5226	51.9
1993	10909	3524	24.4
1994	22289	1804	7.5
1995	37606	1968	5
1996	47659	1996	4
1997	63606	3074	4.6
1998	72086	4245	5.6
1999	66905	3807	5.4
2000	69573	3796	5.2
2001	64848	3737	5.4
2002	60004	3663	5.8
2003	58728	3638	5.8
2004	59325	3916	6.2
2005	61094	3929	6
2006	57988	4155	6.7
2007	54864	4110	7
2008	53026	4875	8.4
2009	46419	4796	9.4
2010	38628	6104	13.6
2011	26802	5615	17.3
2012	17255	3978	18.7
2013	6403	2097	24.7
2014	648	347	34.9
2015	12	8	40

Decisions to make:

1) Should we discard lactations from cows with missing phenotypic information from every analyses. **Depends on the analysis**

Alped cows

In the first round of analysis, we only want to consider alped cows. This condition should have been a filter already applied by Qualitas when the data was extracted. Nevertheless this filter should be checked.

In table K04, the parameter "alpung" should tell us if the cow is alped. It has the value 1 for all animals, as expected. Furthermore, we can identify which records were taken in the alp, because if this is the case, the altitude of the alp is given (on the contrary, if the record is taken on the lowland farm, the altitude is not documented). There are 62 lactations which do not have any records taken in the alp. This is probably due to the fact that the cow was alped but no records were taken during its stay in the alp.

However, it is important to notice that there are only 349'958 out of the 616'081 lactations (58.6%) that have at least two different "BetriebsID" indicating that the record was taken in two different farms. This comes from the fact that some alps are actually an extension of the main farm, so that they have the same farm ID. It can be annoying because, except from the altitude, we will not know the geographic and environmental characteristics of the alp as well.

Decisions to make:

1) Should we remove cows that have no records taken during the alp. **Yes**
2) Should we only keep lactations for which we have at least two milk recording location. **Depends on the analysis**

Alp altitude

The altitude of the alps in the dataset ranges from 700 to 5500m. Since Switzerland does not reach 5500m, it seems clear that there are some mistakes in the given altitude. If we go to a bit lower altitudes, we see that there are very few lactations (172) for which the altitude of the Alp is above 2700m. As for the low altitudes, according to Braunvieh-CH, they consider that a cow is alped if the altitude of the alp is at least 1100m and the difference in altitude is at least 100m. Therefore, altitudes below 1'100m should be discarded.

Table 3: Number of lactation per altitude of alp

Altitude	Num lactations
700	1
1000	58
1100	3025
1200	45107
1300	48210
1400	60535
1500	42169
1600	55407
1700	50440
1800	85871
1900	63862
2000	104740
2100	23906
2200	21169
2300	6977
2400	3583
2500	293
2600	485
2700	50
2800	21
2900	73
3000	1
3100	22
5500	5

Decisions to make:

1) Should we remove too high and low alps? Which cut-off? **Yes, 1100 and 2600?**

Interrupted alp stays

It sometimes happen that a cow is alped, and that for some reasons it goes down to the lowland farm (disease, weather condition, food shortage, ...). This behaviour will bias our analysis. In fact, if the cow stayed long enough in the alp before this interruption, we can already take it into account. Removing only the end of the lactation when a second alping occurs discards 75'688 out of the 5'681'498 records.

Decisions to make:

1) Should we discard the end of the lactation when a second stay in the alp occurs **Yes**

Interval between calving date and first milk record

The guidelines describing the way milk records should be taken specify that the first milk record should be taken between 5 and 42 days after calving. It also adds that in some cases, a record can be taken before the 5th day (aceton-related problems) but that this record should not be taken into account for data analysis.

There are 185 records (out of 5'681'498) from 97 lactations which were taken before the calf was born, often several months before. This could potentially indicate an error in the calving date. Then 15'556 records were taken less than 5 days after calving. There are also 8'384 lactations for which the first record was taken after the 42th day after calving.

Decisions to make:

1) It seems pretty clear that we should discard records taken less than 5 days after calving **Yes**
2) Should we remove the whole lactation when the first record was taken before the calf was born **Yes**
3) Should we remove lactations for which the first record was taken after the 42th days after calving **Yes**

Breed

In this analysis, we want to focus on Braunvieh (BV) and Original Braunvieh (OB) cows. In the data sent, we have 132 cows that have 'BS' as breed code representing 506 lactations. Furthermore, 65'521 (out of 245'313) cows have a sire or dam from a different breed, though most of them (64'202) come from a 'BS' parent.

Decisions to make:

1) Should we remove 'BS' cows ? **Yes**
2) Should we remove cows with parents from different breeds? If yes, do we accept BS as a potential parent ? **Yes, but keep BS as potential parent**

Lactation cycles

The maximum lactation number we observe is 18 (though there is only one occurrence). The number of cows decreases from the first lactation onwards, as expected. When studying lactations, a standard way of proceeding is to treat separately 1st, 2nd and 3rd and higher lactation. However, it seems clear that a cow being in its 18th lactation will perform differently from one that is in its 3rd lactation. Therefore, we should probably apply a condition to filter out too old cows.

To give actual numbers, out of the 616'081 lactations, there are 23'076 lactations for which the cow is in its 8th lactation cycle or more, 2'866 lactations if we consider only lactation cycle greater than 10.

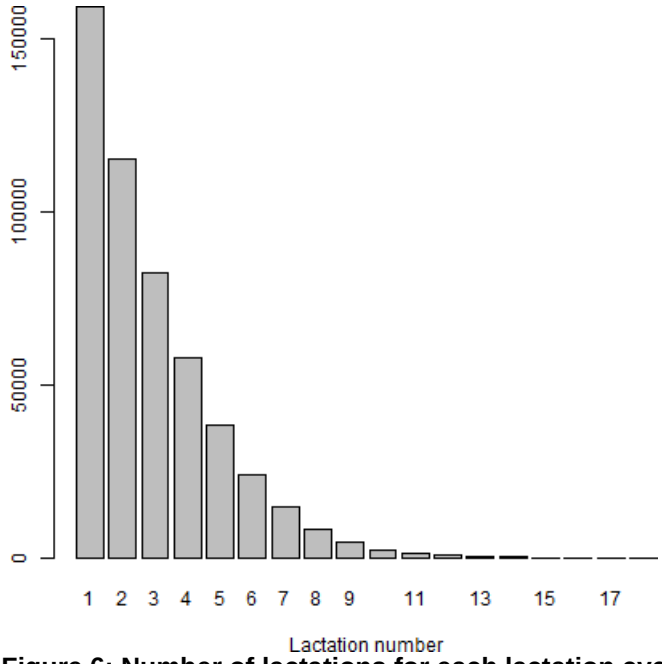


Figure 6: Number of lactations for each lactation cycle

Decisions to make:

1) Should we remove too old cows? Which cut-off. **Yes, >10th lactation**

Calving months

It is also interesting to study the calving season. Table 4 shows the number of calves born per month. We clearly see that the number of births decreases during summer, which is expected when the cow is alped, as farmers generally avoid calving during the alping season.

Furthermore, we will see in the descriptive statistics that these cows behave quite differently in terms of lactation curve. Finally, the fact that the first record is taken in the alp will disturb our analysis.

Table 4: Number of calving per month

Month	Num calves
1	43176
2	31508
3	27486
4	18748
5	10278
6	4882
7	5281
8	15342
9	78889
10	129734
11	87919
12	55566

Decisions to make:

1) Should we remove cows that calved during summer? Which month? **Yes. From March to August**

Filter summary

Table 5 summarises the number of cows (resp. number of cows with known phenotype), lactations and records before applying filters (discribed in 6). Table 6 lists all conditions applied to filter the data and the number of rows it impacts. It should be noted that when cows are deleted, it will also influence the number of lactations and records, while when lactations are deleted, it also impacts the number of records.

For information, the corresponding number of alps, as well as the number of alps whose location is known is also given. Finally the number of alps for which at least the PLZ is known is reported. Though the location of the alp might be missing, the altitude is always reported.

Table 5: Data size before and after applying filters

Category	unfiltered	filtered
Cows	245’313	187’327
Cows (known morpho)	212’037	168’172
Lactations	616’081	421’900
Records	5’681’498	4’003’001
Alps	8’266	6’443
Alps (known location)	3’264	3’022
Alps (known PLZ)	7’619	6’093

Table 6: Conditions used to filter out data and the number of cows/lactations/records it impacts

Condition	impact
Calving date after 03.2015	41’177 lact.
Calving date before 08.2000	13’779 lact.
Lact duration < 270 days	94’195 lact.
Records after 500 days after calving	35’640 rec.
Interval bet. calvings < 290 days	255 lact.
Interval 1st-last insemination (avg 3 lact)>100 days	16’158 cows
First calf < 2 years	1’087 cows
First calf > 4 years	406 cows
No record in the alp	62 lact.
Alp alt <1’100m or >2’600	231 lact.
Interrupted alp stay	75’688 rec.
Lact with records before calving	97 lact.
First record before 5 days after calving	15’556 rec.
First record after 42 days after calving	8’384 lact.
Breed = BS	132 cows
Parent breed <> BV, OB or BS	1’319 cows
11th and higer lactations	2’866 lact.
Calving month = May-July	31’351 lact.

2 Descriptive statistics

Unless otherwise specified, filters described in the previous section have been applied for the rest of the report.

Total number of lactations

Table 7 shows the number of cows (unfiltered) for which we have lactation information. It displays the total number of lactations as well as the number of standard lactations. The total number of lactations is a proxy for the number of alped cows per year. Indeed, even though some cows are not reported in the database, the relative number from year to year gives us relevant insights: generally, we observe a relatively constant decreasing trend between the year 2000 and 2014, summing up to a loss of 25% over the last 15 years.

Table 7: Number of with known lactation characteristics per year (year of birth of the cow - year 2015 not complete)

Year	Total number of lactations	Number of standard lactations
2000	42474	38347
2001	41868	39474
2002	40210	38762
2003	38268	37186
2004	37455	36458
2005	37146	36132
2006	37304	36248
2007	36276	35404
2008	36776	35891
2009	36796	35687
2010	36546	35437
2011	35015	34034
2012	34149	33178
2013	34722	33653
2014	34003	32761
2015	33377	30312
2016	13432	10682
2017	1445	1170

Number of daughters per sire

The distribution of the number of daughters per sire is, as expected very unequal. The most represented sire of the database conceived 6’998, representing 32’201 lactations (given that many cows have several lactations). At the other end of the picture, 3’832 bulls sired only 1 cow of the database.

Figure 7 shows on the x-axis how many bulls we must have to account for a given percentage (y-axis) of lactations. As an example, we see that with 500 bulls, we can account for a bit more than 60% of the lactations (unfiltered data).

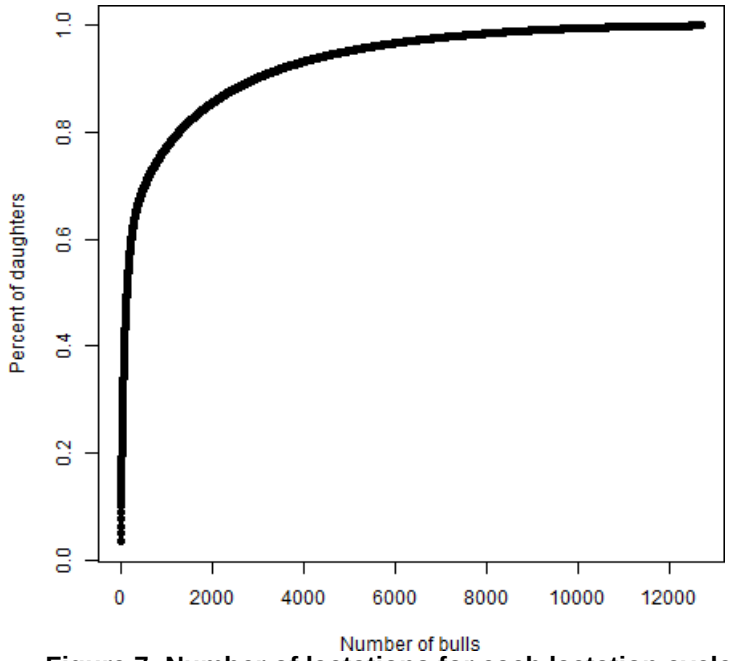
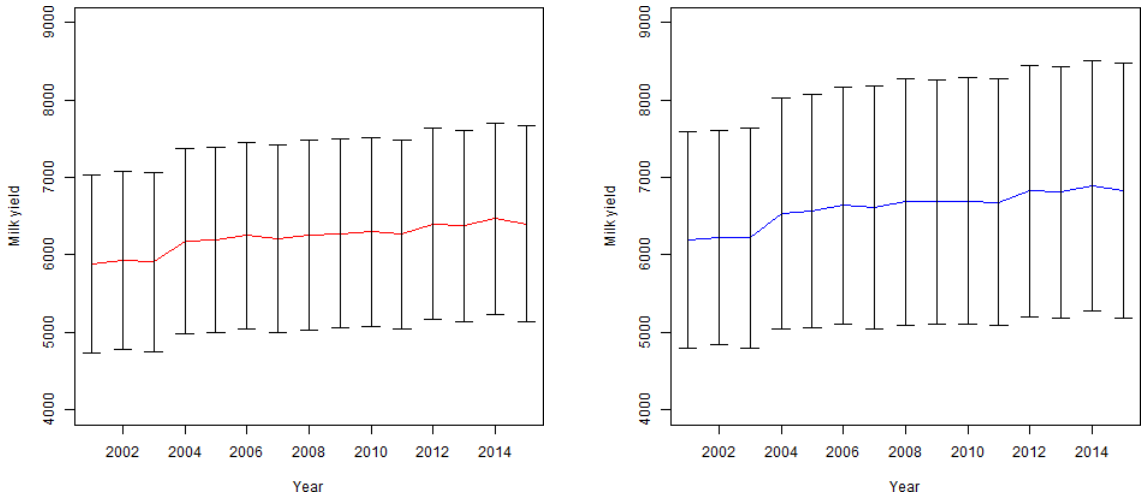


Figure 7: Number of lactations for each lactation cycle

Milk yield and lactation duration

Figure 8 shows the average milk yield as well as its standard deviation for standardised and full lactations per year. We clearly see a relatively steady increase in the milk yield, suggesting more efficient cows due to increased selection. The mean milk yield over standardised lactations goes from 5'888 kg in the year 2000 to 6'401 kg today, i.e. an increase of 9%. However, it is also worth noting that the milk production is extremely variable given that the standard deviation is around 20% of the mean for standardised lactation (around 24% for full lactation). We can also see some slight drop in the milk production for some years. One year worth mentioning is the year 2003, known to have had an extremely hot and dry summer. Thus we see that the environmental and climatic conditions a priori have an influence on the milk production.

The analysis of the difference between standardised and full lactation, not surprisingly reveals a higher mean but also a larger standard deviation.



(c) Standardised lactations (b) Full lactations

Figure 8: Mean and standard deviation of the milk yield for standardised lactation (a) and full lactation (b), per year

As mentioned earlier, milk production is extremely influenced by the lactation number. Figure shows the mean milk yield per year and lactation number as well as its standard deviation for standardised lactation. Globally, the average milk yield for 1st lactation is 500kg smaller than for 2nd lactation, and 1000kg smaller than for 3rd and higher lactation. The ratio "mean over standard deviation" is slightly lower than when considering all lactations together, which is expected since the lactation number should explain part of the variation. This ratio is of 20% for 1st and 2nd lactation 22% for 3rd and higher lactation. As already explained, this slightly higher last number probably comes from the grouping of many lactation numbers together.

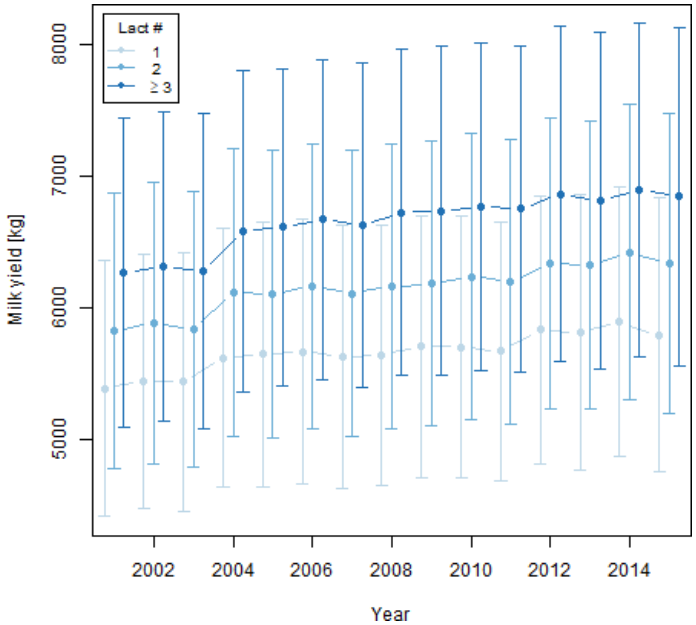


Figure 9: Mean milk yield +/-1 *σ* for standardised lactations grouped by lactation number (1, 2, 3 and more) per year

Milk yield according to calving month

The next figure summarises the average total milk production over the whole lactation (only for cows in their first lactation, over unfiltered data to also show milk production for cows calving in summer). Cows having calved during spring and summer produce, on average, significantly less milk than those that calved during fall and winter (4611 for May versus 6439 for September).

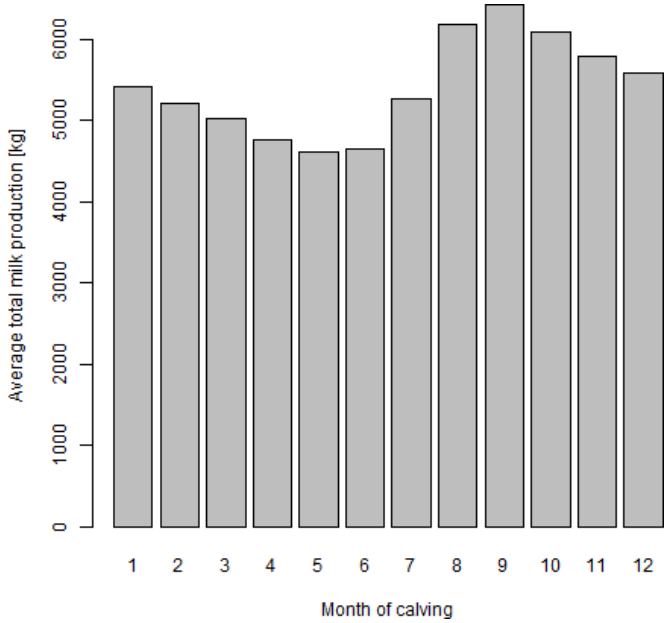


Figure 10: Average milk production of cows in their first lactation depending on the month of calving

Milk properties

Fat

Figure 11 shows the evolution of the mean fat content in % over the years for all 3 lactations group.

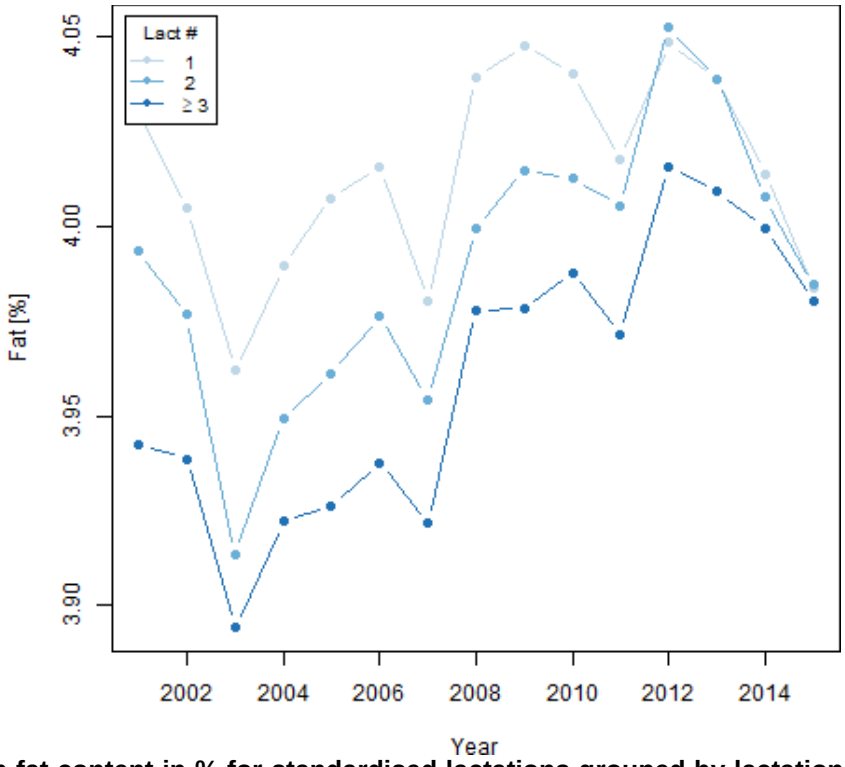


Figure 11: Mean fat content in % for standardised lactations grouped by lactation number (1, 2, 3 and more) per year

Although the fat content sensibly varies across years, there is no clear general increasing trend, but one seem to notice a slight increase over the years. It is interesting to notice that the two years highlighted as those with a smaller milk production (i.e. 2003 and 2015) also are the ones with the poorest fat content. A peculiar distinction between these two years mentioned above is the difference between the three lactation groups: while 2003 shows a wide range between 1st and 3rd lactation, the distinction between the three groups is quasi null for the year 2015.

The amplitude of the variation is of the order of 0.15%

First lactation cows produce slightly fatter milk, followed by second lactation cows.

Protein

Figure 12 shows the evolution of the mean protein content in % over the years for all 3 lactation groups.

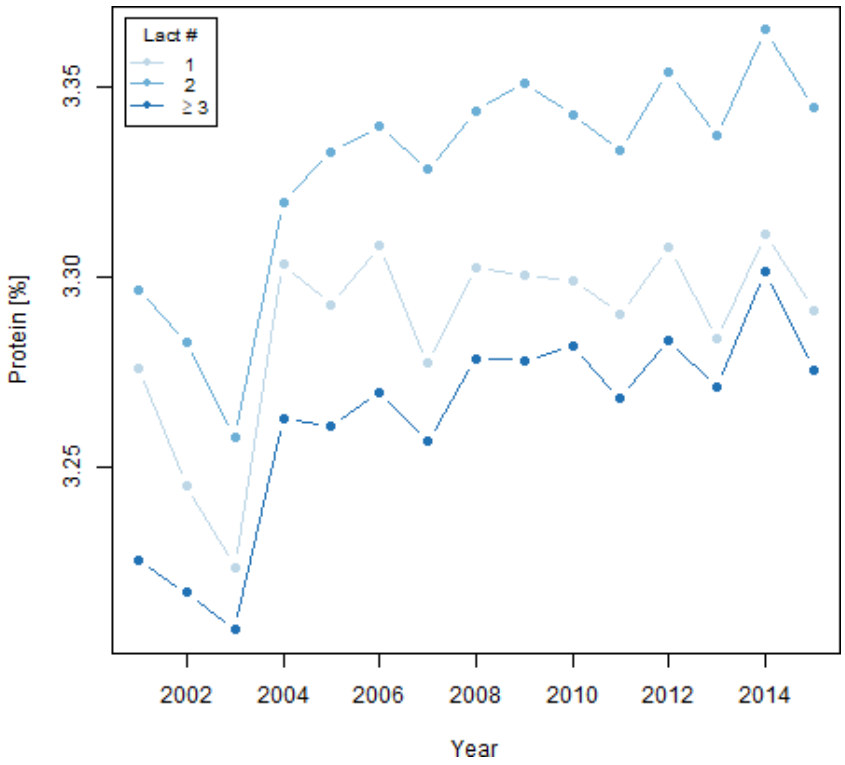


Figure 12: Mean protein content in % for standardised lactations grouped by lactation number (1, 2, 3 and more) per year

Here again, we see no significant increase or decrease over the years. In fact, the evolution is rather flat, except for second lactation cows that tend to have a milk richer in protein over the years (0.1% increase between 2005 and 2015). The year 2003 is again clearly differentiated from its neighbours with a clear drop in protein content.

However this time, 1st lactation cows have a milk poorer in protein than 2nd lactation ones but richer than 3rd and higher. The graph is not shown here but if consider 4 lactation groups (with 1,2,3,4 and higher lactation), 3rd lactation cows still have a smaller protein content than 1st lactation cows.

Lactose

Figure 13 shows the evolution of the mean lactose content in % over the years for all 3 lactations group.

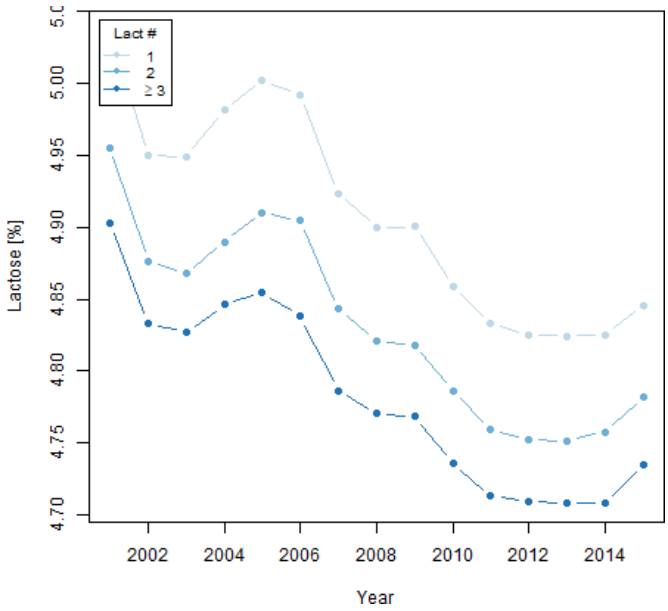


Figure 13: Lactose content in % for standardised lactations grouped by lactation number (1, 2, 3 and more) per year

Unlike fat and protein content, the lactose content shows a clear trend over the years: the mean lactose content has decreased of 0.2% over the last 15 years, though the situation seems to be stabilised between 2010 and 2015.

First lactation cows have a higher lactose content than second, which have themselves a higher content than third and higher lactation.

Somatic cell count

Figure 14 shows the evolution of the mean somatic cell count (SCC) over the years for all 3 lactations group.

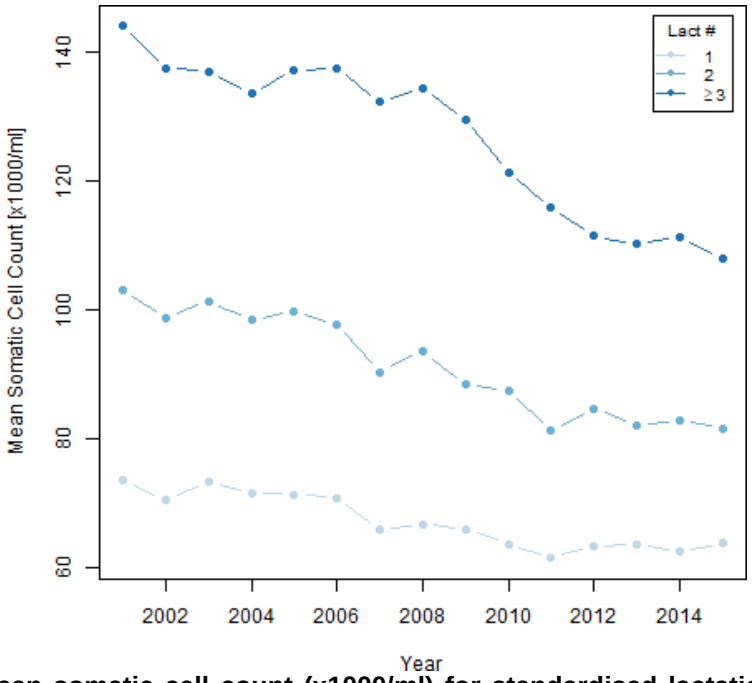


Figure 14: Mean somatic cell count (x1000/ml) for standardised lactations grouped by lactation number (1, 2, 3 and more) per year

Here it is clearly visible that 3rd lactation cows have a substantially higher count of somatic cell than 2nd and that 2nd lactation cows have in return a higher count than 1st lactation cows. In fact this trend is also visible for higher lactations (4th>3rd and so on - not shown here) and this holds for up to the 6th lactation. After the image is not so clear.

Beside this observation, we also see a decline among the years of SCC. This is especially true for 3rd and higher lactation cows.

It must also be noted that the range of values for SCC is very wide (between 5 and 9999). Therefore, the mean value can be very much affected by extreme values. It is thus relevant to have a look at the same graph as before representing the median instead of the mean (Figure 14).

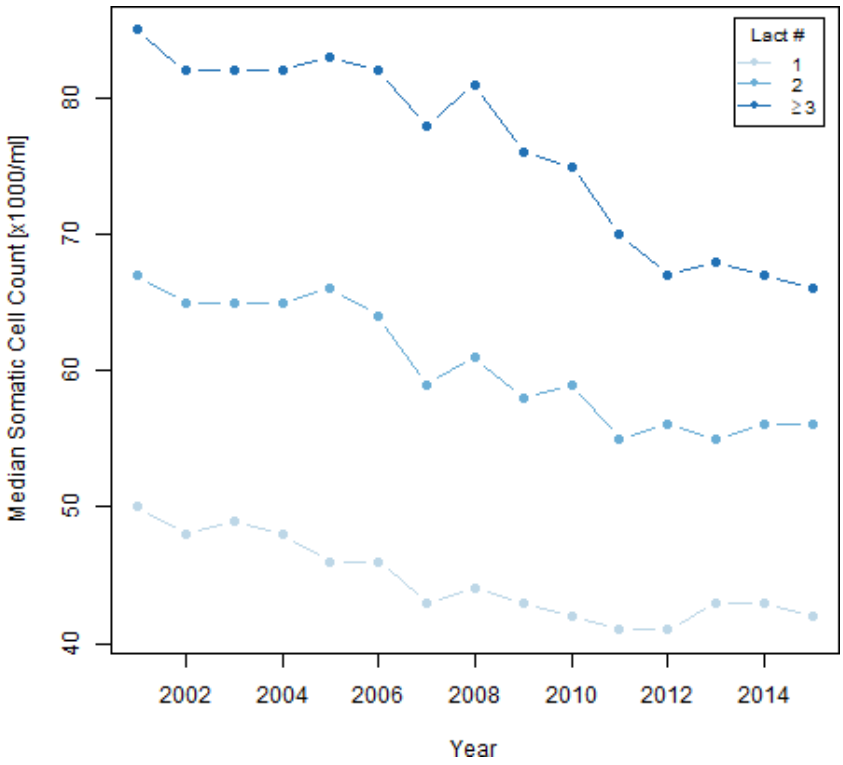


Figure 15: Median somatic cell count (x1000/ml) for standardised lactations grouped by lacta- tion number (1, 2, 3 and more) per year

The shape of the graph is relatively similar, though the values are much lower. Both previously mentioned observations still holds for the median value.

Influence of altitude

Figure 16 shows the influence of altitude of the alp on milk quantity.

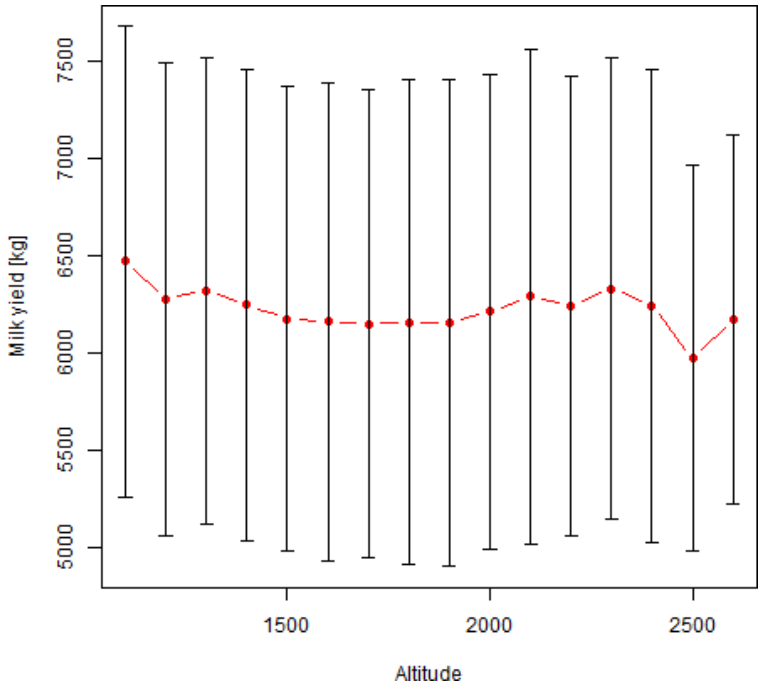
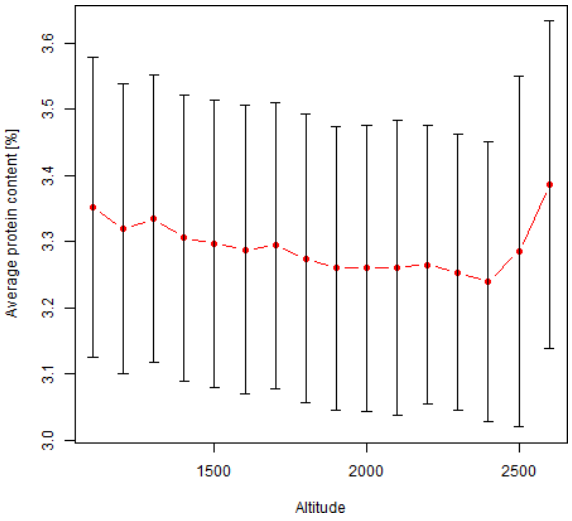


Figure 16: Milk yield +/- 1σ for standardised lactations grouped by lactation (1, 2, 3 and more) and altitude windows (100m)

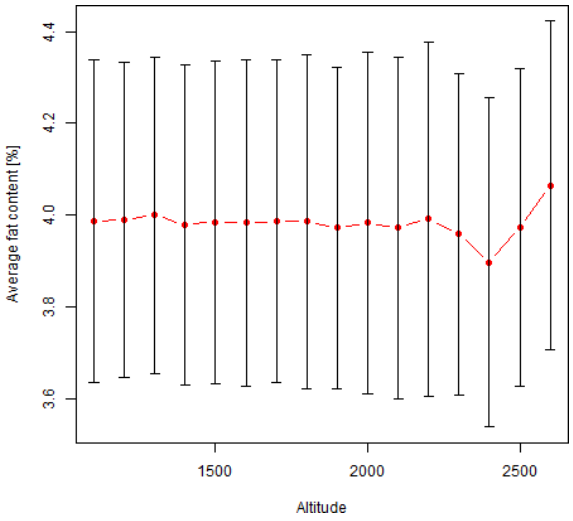
Here we see no clear relationship between the milk yield and altitude. The slope of the regression decreases a bit at first but then increases again.

However, this figure is not very informative, because we consider the milk yield over the whole lactation period and not the one during alping. The problem is that we do not know the milk yield during the alping season since we only have a few milk records and we do not know when the season exactly starts and ends. Furthermore, single milk records are influenced mainly by the duration since calving but also on the climatic conditions of that day so that it is difficult to compare them. We need a more complex model to account for all these variables.

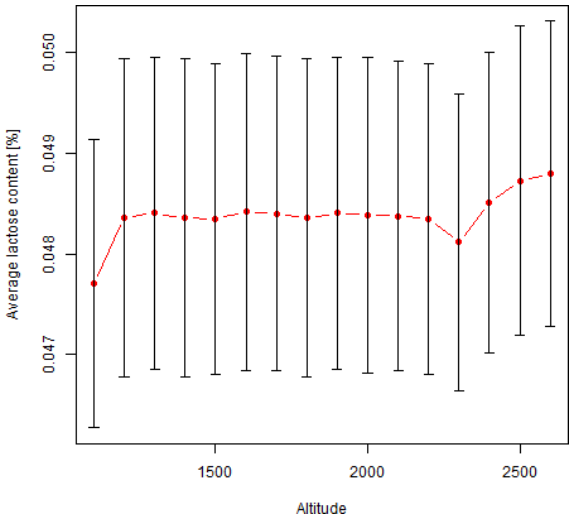
We can also look at the influence of altitude on milk properties (Figure 17). If we look within the same altitude range between 1000 and 2500m, there is no significant influence of altitude on milk properties. Only the protein content slightly decreases (less than 0.1%) but considering the high standard variation we observe, this decrease is hardly interpretable.



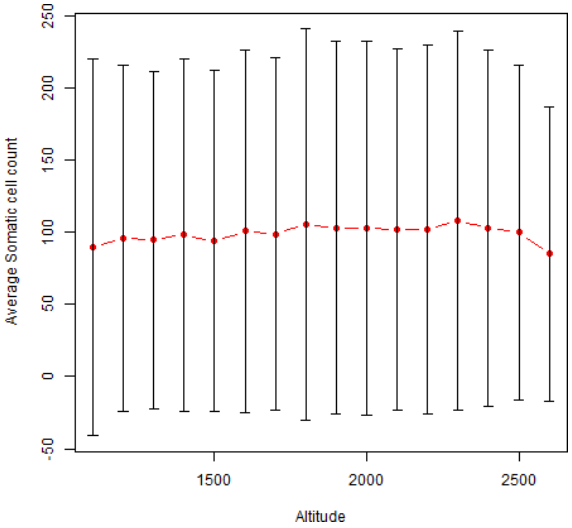
(d) Protein



(b) Fat



(c) Lactose



(d) SCC

Figure 17: Average milk component as a function of altitude

Persitency

Figure 18 shows the evolution of persistency over the years for the three lactations

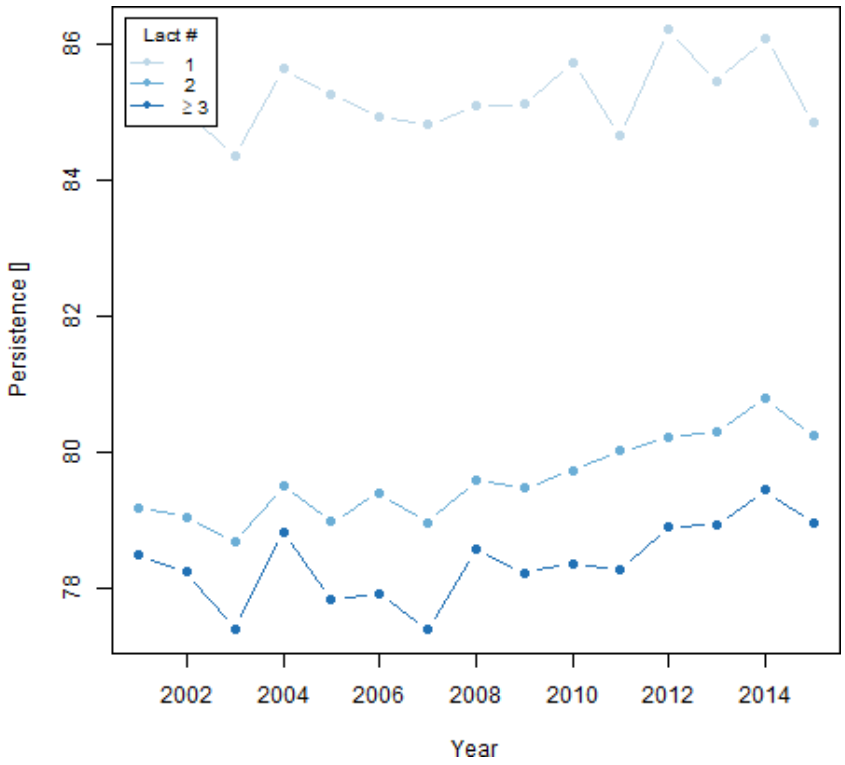


Figure 18: Persistency over the years of standardised lactations grouped by lactation (1, 2, 3 and more)

There is no clear difference between the years. However first lactation cows clearly show increased persistency, while second lactation have only slightly higher persistency than third and higher lactation cows. This seems consistent with the literature. Putting this observation together with the one on milk yield, we thus see that first lactation cows have a lower overall milk yield but a higher persistency throughout the milking season.

Shape of lactation curves

Figure 19 shows the average shape of lactation curves. It gives some specifics about the section on persistency. Indeed, the milk yield peak is lower for first lactation cows and the slope after the peak is more gentle. Third and higher lactation cows have the most pronounced peak. The end of the lactation period (i.e. after 200 days) is sensibly similar for the three lactation numbers. The same kind of figure is given to show the shape of the lactation curve in terms of milk content, such as protein, fat and lactose content (Figure 20)

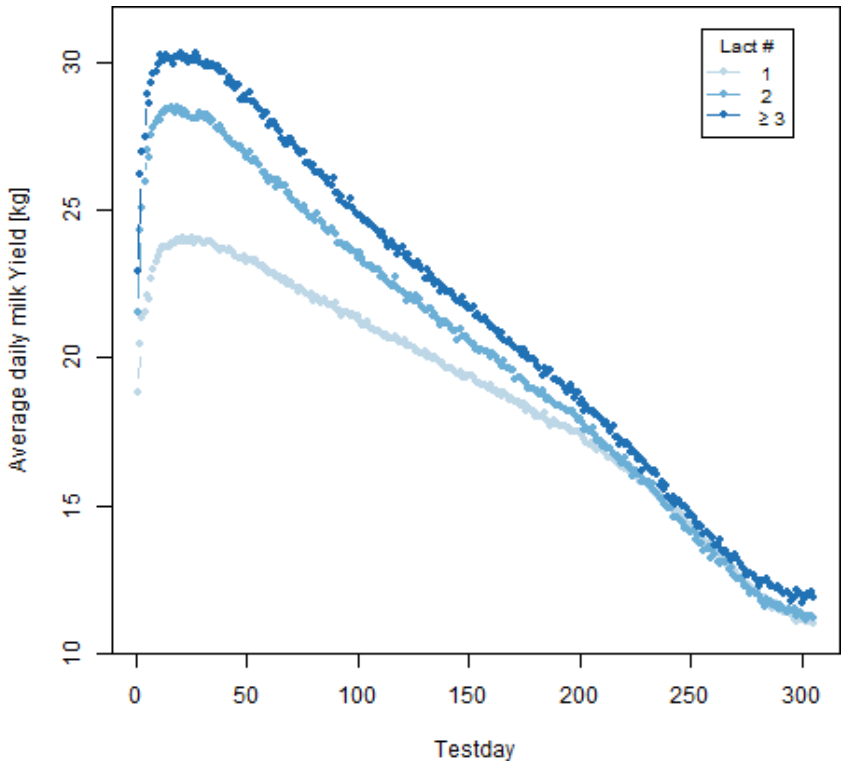
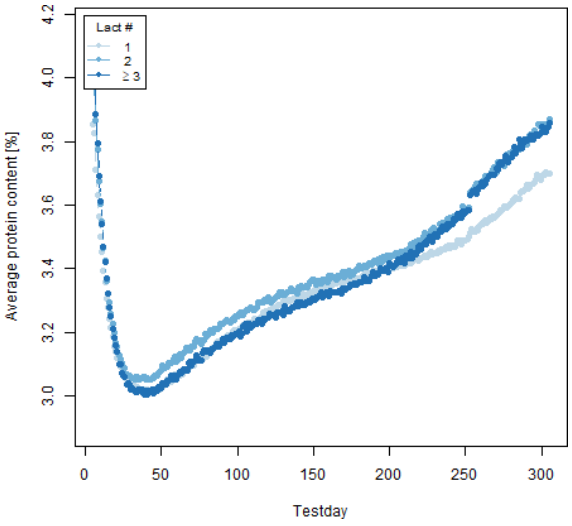
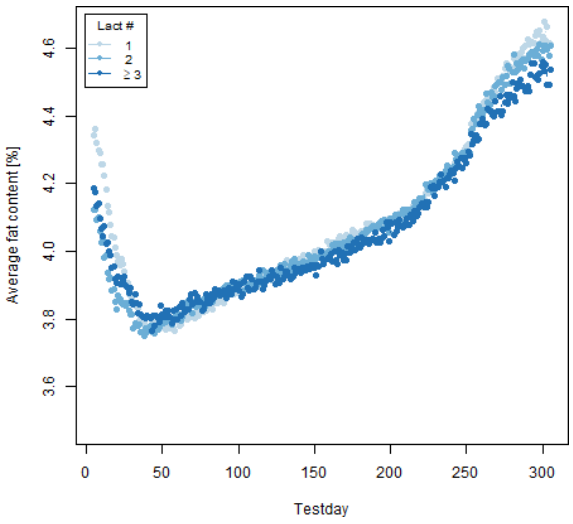


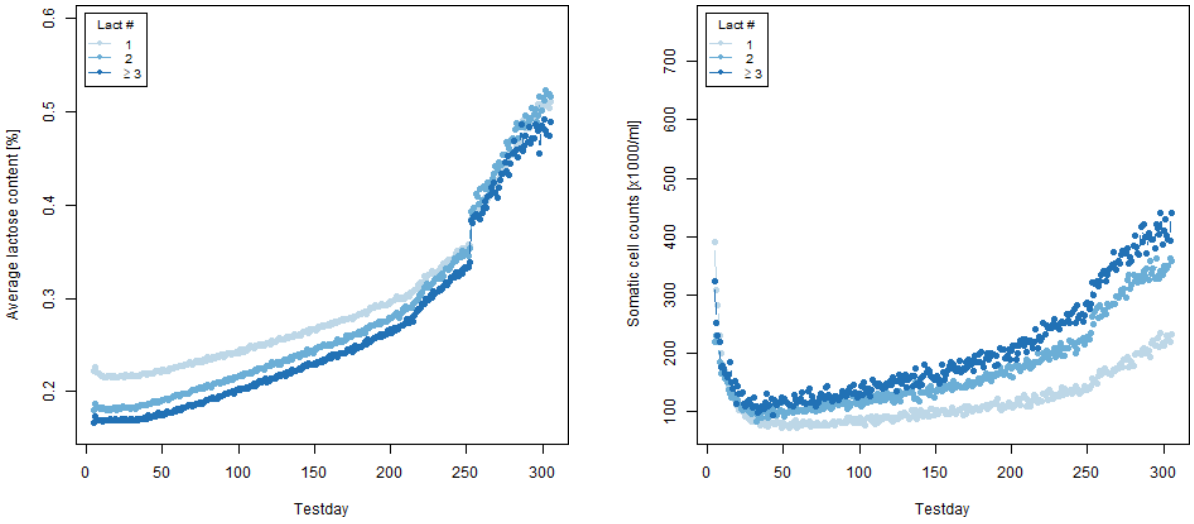
Figure 19: Shape of lactation curves for all three lactations (1, 2, 3 and more). Each dot corresponds to the average milk yield for a given day after calving



(e) Protein



(b) Fat

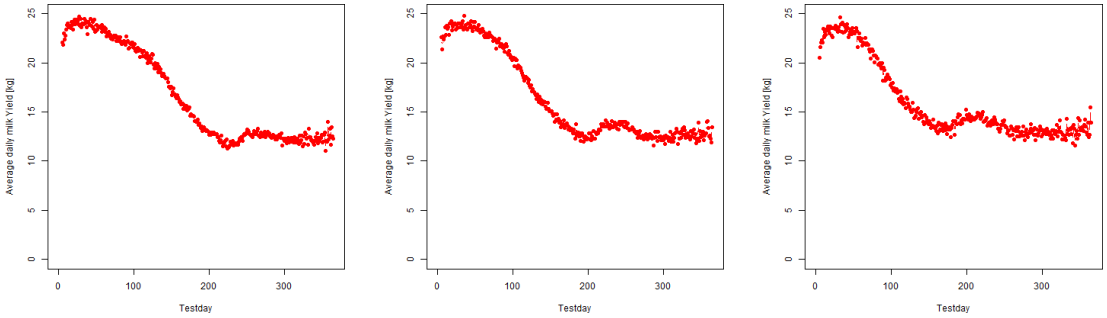


(c) Lactose (d) SCC

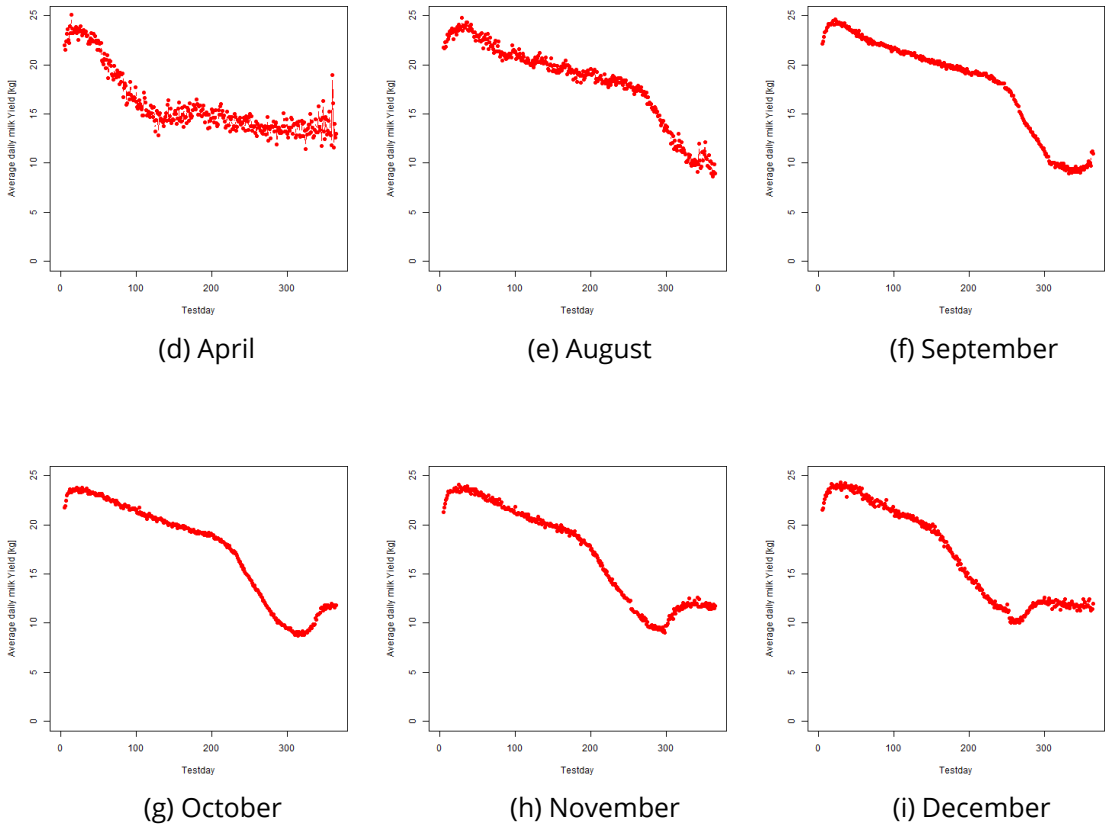
Figure 20: Average milk component curves: protein [%], fat [%] and lactose [%] content and Somatic cell counts [x1000/ml]

Shape of lactation curves according to the month of calving

Figure 21 shows the lactation curve depending on the calving month. To simplify the analysis, we only consider cows in their first lactation in these plots.



(f) January (b) February (c) March



(d) April (e) August (f) September

(g) October (h) November (i) December

Figure 21: Average milk production per testday depending on the calving month for cows in their first lactation

The first thing that catches our attention is that lactation curve are indeed different depending on the month in which the cow calved. This is probably the consequence of the influence of climate on milk production. Something interesting to notice is that milk production reaches a minimum towards the end of summer and then comes slightly up again for cows that calved during winter. This phenomenon is not seen in the total average lactation curve since it is averaged and thus disappears.

The fuzzier plots obtained for calving in spring and summer months simply comes from the fact that we have less cows from which the average is calculated therefore leading to a less stable result.

The period in which milk yield raises again at the end of the lactation period seen in figure 21seems to correspond to the end of the alping season. Thus, cows coming down from alpine pastures apparently have again an increased milk production. This is proven in Table 8 where we see that for cows having 2 or more records taken during the alping period, the milk yield slightly increases right after. Cows whose alping period is shorter are not affected similarly; in fact, on average, cows whose alping period is long experience a more significant increase after the alping period. It is necessary to remind that those increase are small (0.5 kg at most) and we could speak of it staying stable, but the mere fact that the milk yield does not decrease is relevant enough to be mentioned.

Table 8: Average milk yield per record right before, during and after the alp. Rows are grouped according to the number of records taken in the alp, which is a proxi of the duration of the alping season

nb records	num cows	before	begin	1st	2nd	end	after
1	23422	21.33	16.87	-	-	-	14.74
2	80726	22.81	17.91	-	-	13.98	14.02
3	140562	23.59	19.05	15.42	-	12.79	13.91
4	19519	25.21	21.98	18.3	15.81	14.05	14.5

Correlations between milk properties and phenotypes

Figure 22 depicts the correlation between phenotypes and milk yield, properties and persistency of lactation. The first thing to notice is that while milk yield and protein content is correlated to several phenotypes, both fat content and persistency have no clear correlations with whatever trait. Milk yield and protein are most affected by the format (i.e. the size and shape) and the udder description of the cow. Generally, the bigger the cow, the more milk it will produce with higher protein content. Regarding the udders, the bigger (large and long) its udders, the more milk and protein the cow will produce.

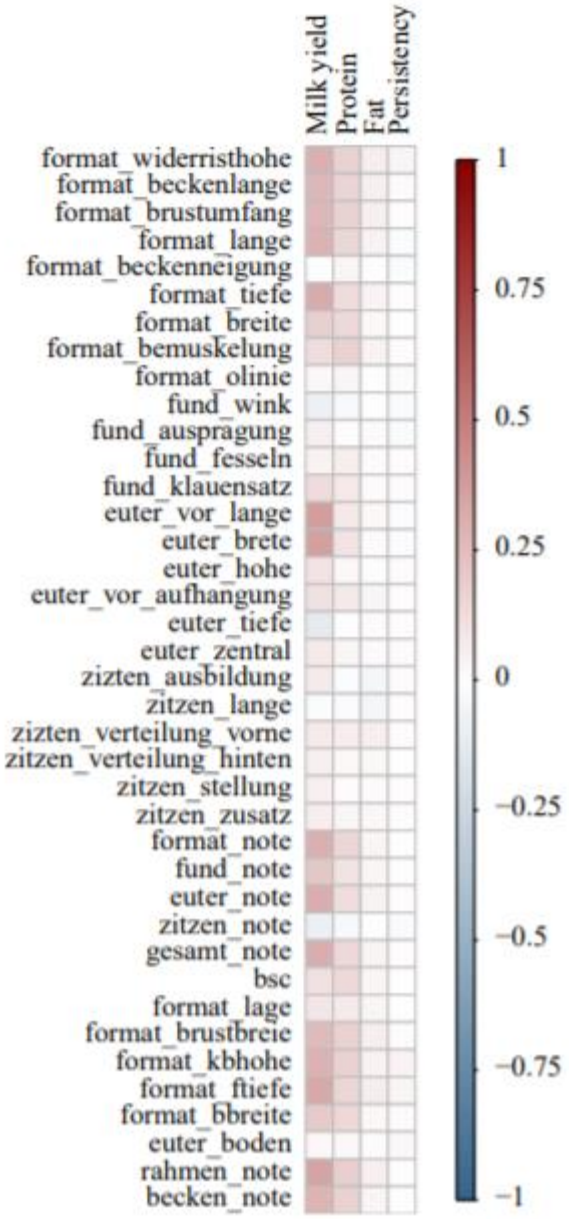


Figure 22: Correlation between milk yield, properties or persistency and phenotypes

Geographic distribution

Figure 23 shows the average persistency of lactation for 2nd lactation cows grouped by PLZ. The grouping has been done in order to allow a better visualisation. Interestingly, we see that persistency is higher in the Alp region (Grison, Nord of Tessin, Valais) than in central and North-Eastern Switzerland.

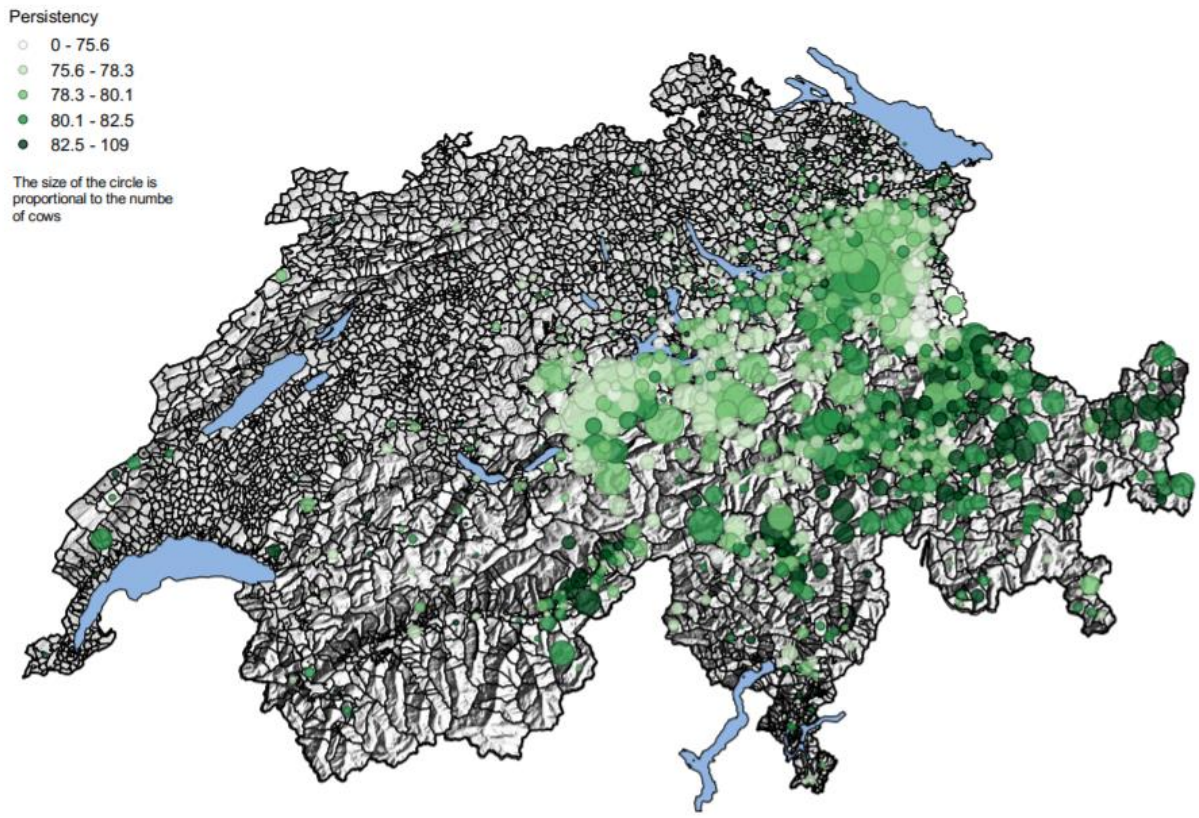


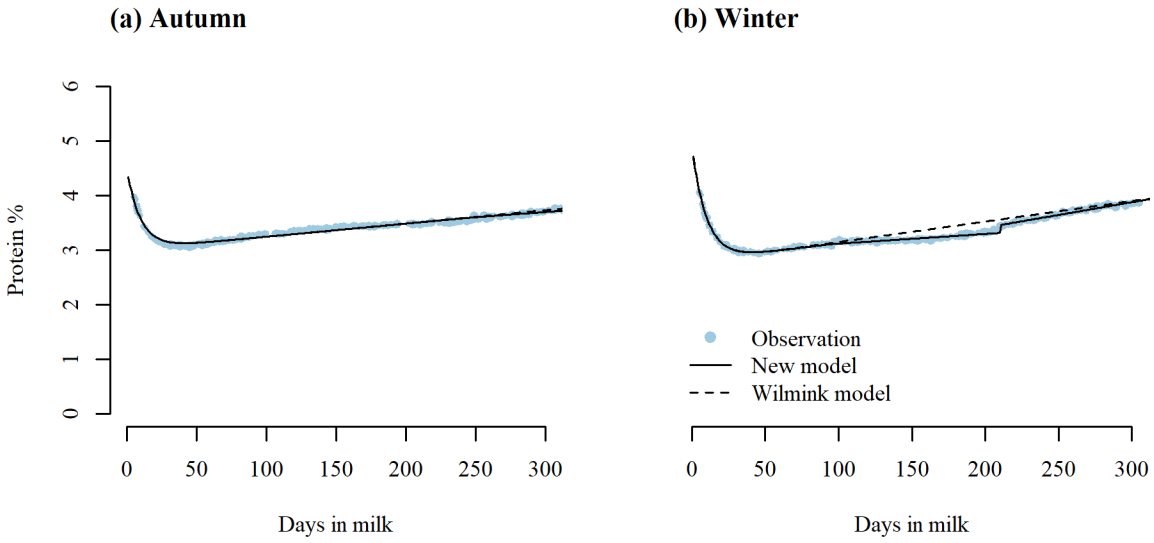
Figure 23: Geographic distribution of persistency of 2nd lactation cows grouped by PLZ

Appendix E
Supporting information for article in chapter 4

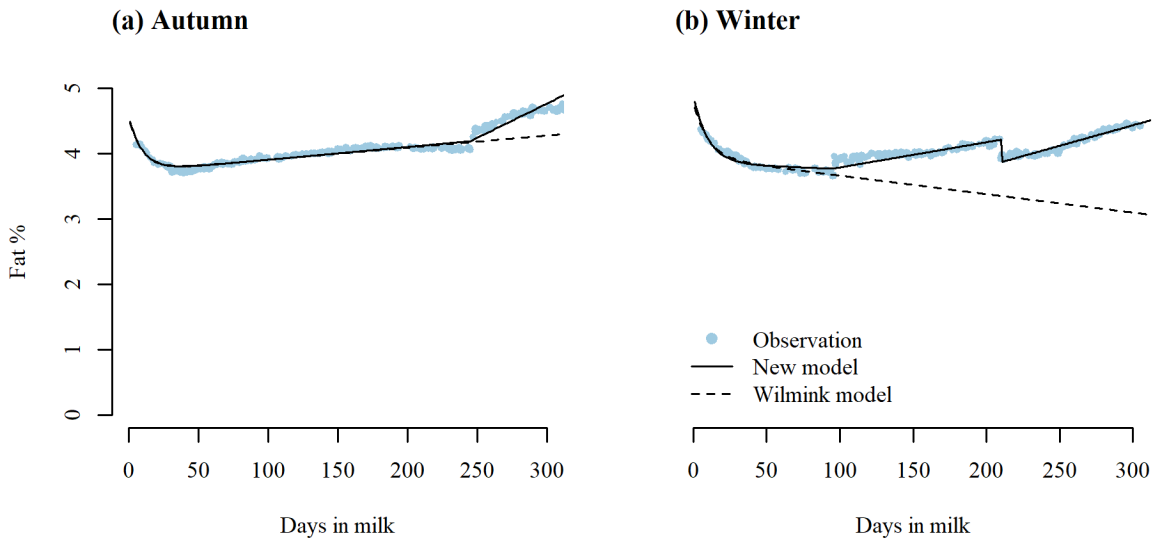
Sup. Mat. S1: The between-group difference of all tested factors influencing milk production. For each criterion and calving month, the between-group difference in milk production in the alp and over the total lactation cycle is reported, together with the Δd (reflecting how differently the two groups are impacted by alping) and its associated significance.

Crit	Calving month	Δmilk alp	Δmilk total	Δd	p-value
Lact #	9	5.3	16.9	0.0070	1.16E-01
	10	7.0	17.5	-0.0036	5.74E-01
	11	11.3	18.8	-0.0071	4.55E-04
	12	16.2	19.8	-0.0077	1.69E-03
	1	18.9	19.4	-0.0066	4.61E-01
	2	21.4	19.8	-0.0080	1.00E+00
Pregnancy stage	9	23.2	2.5	0.0333	2.16E-15
	10	18.4	4.3	0.0278	1.98E-20
	11	12.1	4.1	0.0250	1.40E-27
	12	6.4	3.6	0.0203	6.84E-18
	1	2.5	2.9	0.0161	1.63E-04
	2	1.0	2.8	0.0202	1.93E-01
THI-3d	9	3.2	0.2	0.0065	1.78E-01
	10	1.8	0.3	0.0039	1.17E-01
	11	1.3	0.4	0.0030	1.28E-01
	12	0.1	0.1	0.0004	1.00E+00
	1	0.8	0.6	0.0023	5.38E-01
	2	1.5	1.7	0.0046	4.96E-02
THI-30d	9	0.1	0.0	0.0002	1.00E+00
	10	2.3	0.3	0.0049	3.57E-02
	11	1.9	0.5	0.0044	5.24E-03
	12	0.8	0.4	0.0022	3.10E-01
	1	0.6	0.5	0.0018	1.00E+00
	2	1.0	1.1	0.0030	5.30E-01
CSI-3d	9	-1.3	-0.1	-0.0026	1.00E+00
	10	-0.1	0.0	-0.0002	1.00E+00
	11	-0.7	-0.2	-0.0017	1.00E+00
	12	-0.4	-0.2	-0.0011	1.00E+00
	1	-0.5	-0.4	-0.0015	1.00E+00
	2	-0.8	-0.9	-0.0023	9.23E-01

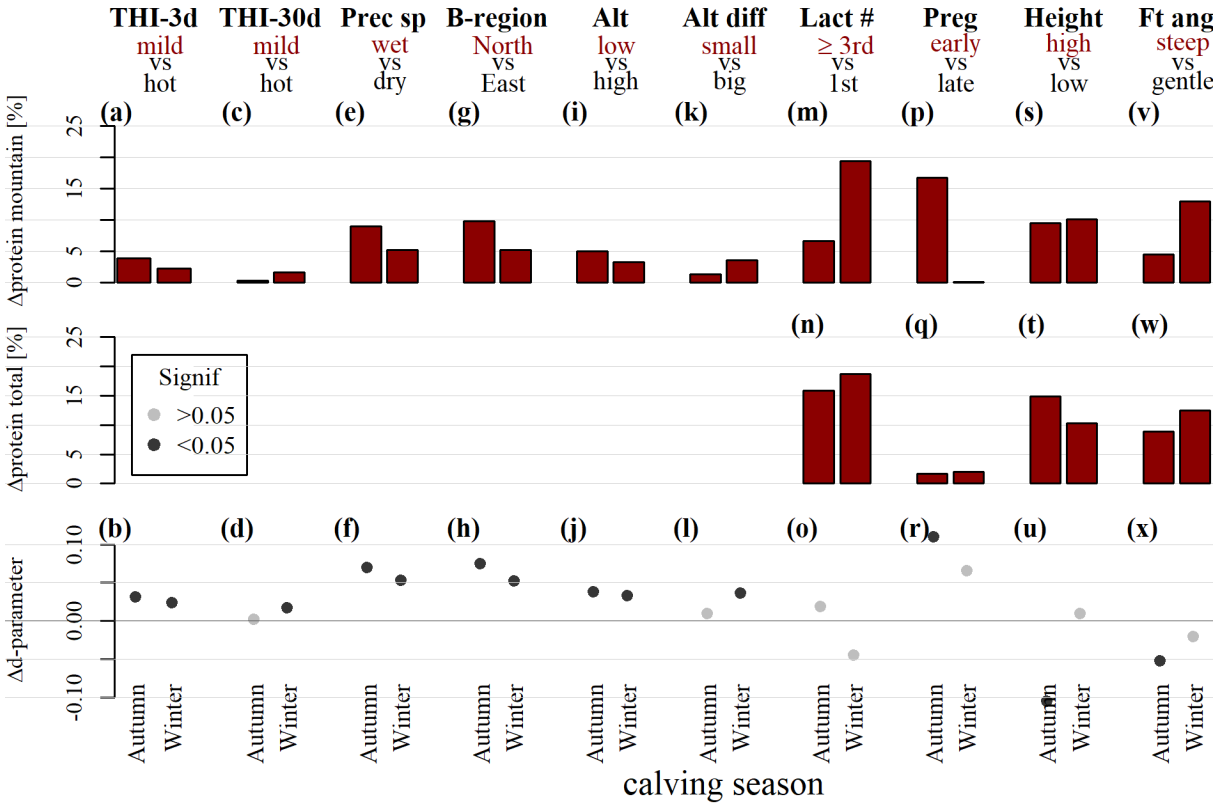
Crit	Calving month	Δmilk alp	Δmilk total	Δd	p-value
CSI-30d	9	-0.7	0.0	-0.0014	1.00E+00
	10	-0.2	0.0	-0.0004	1.00E+00
	11	0.2	0.1	0.0005	1.00E+00
	12	0.2	0.1	0.0004	1.00E+00
	1	-0.6	-0.5	-0.0018	9.44E-01
	2	-1.4	-1.6	-0.0043	7.45E-02
Biogeographical region	9	10.4	0.6	0.0209	3.82E-20
	10	9.2	1.3	0.0198	9.38E-47
	11	8.2	2.3	0.0196	1.69E-73
	12	6.9	3.3	0.0183	3.84E-59
	1	5.3	4.0	0.0152	4.47E-19
	2	5.1	5.9	0.0159	2.37E-11
Altitude	9	5.3	0.3	0.0107	6.72E-07
	10	5.3	0.7	0.0115	1.18E-18
	11	5.6	1.6	0.0133	1.35E-44
	12	4.6	2.2	0.0121	1.10E-37
	1	3.6	2.8	0.0104	2.57E-14
	2	3.7	4.2	0.0115	5.20E-12
Difference in altitude	9	1.2	0.1	0.0025	1.00E+00
	10	1.1	0.2	0.0023	3.30E-01
	11	2.2	0.6	0.0052	2.27E-07
	12	2.6	1.3	0.0070	2.14E-13
	1	3.5	2.7	0.0101	3.84E-15
	2	3.8	4.3	0.0118	2.59E-13
Aspect (100m)	9	0.4	0.0	0.0008	1.00E+00
	10	0.8	0.1	0.0018	6.76E-01
	11	0.3	0.1	0.0006	1.00E+00
	12	0.3	0.1	0.0007	1.00E+00
	1	0.3	0.2	0.0008	1.00E+00
	2	-0.2	-0.3	-0.0008	1.00E+00
Aspect (1km)	9	1.6	0.1	0.0032	5.04E-01
	10	1.1	0.2	0.0025	1.78E-01
	11	0.8	0.2	0.0020	7.99E-02
	12	0.7	0.4	0.0020	8.61E-02
	1	0.5	0.4	0.0015	1.00E+00
	2	0.4	0.4	0.0012	1.00E+00
Height	9	7.7	14.9	-0.0137	7.31E-04
	10	8.5	14.5	-0.0125	4.68E-06
	11	9.9	14.7	-0.0091	3.89E-04
	12	11.5	14.7	-0.0089	9.68E-03
	1	11.2	12.4	-0.0022	1.00E+00
	2	11.9	10.9	0.0105	9.12E-01
Foot angle	9	2.8	6.6	-0.0095	1.32E-02
	10	3.3	6.2	-0.0079	2.60E-03
	11	4.8	7.2	-0.0082	3.21E-03
	12	7.4	8.4	-0.0065	4.39E-01
	1	9.6	9.3	-0.0025	1.00E+00
	2	11.1	9.4	-0.0005	1.00E+00



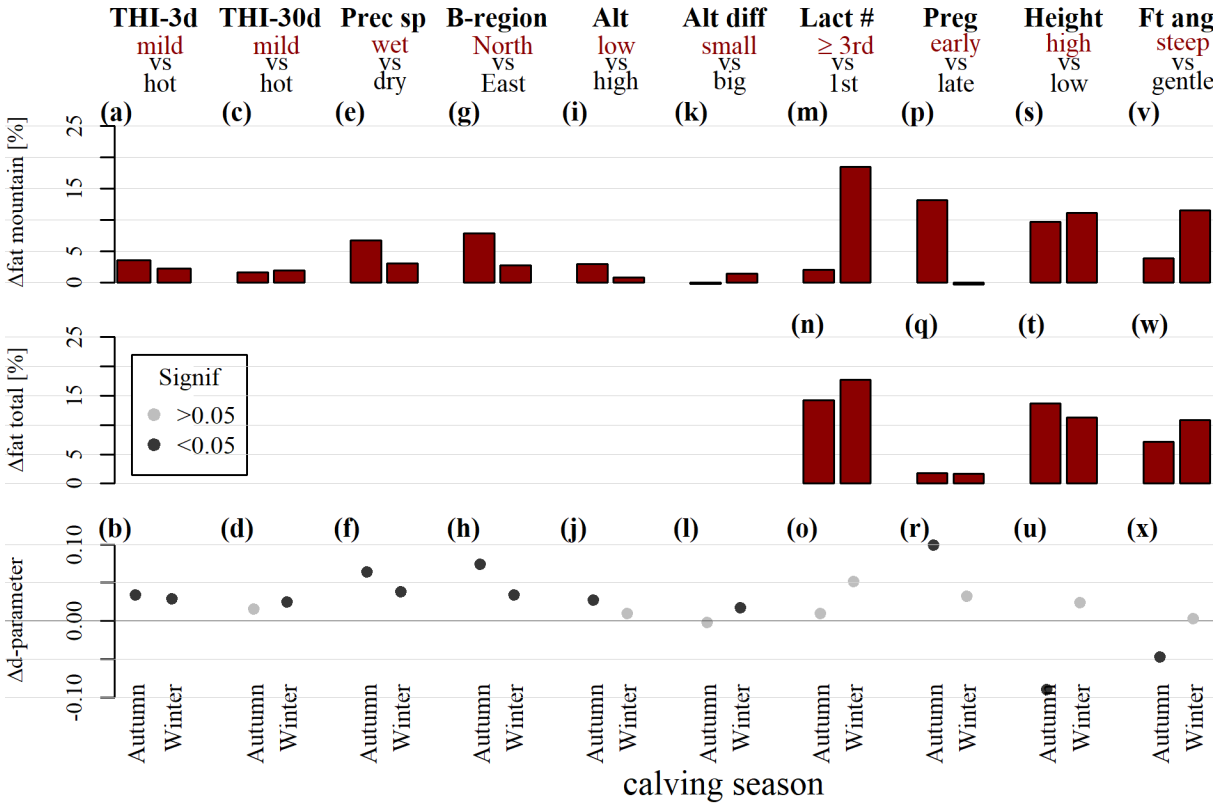
Sup. Mat. S2: Evolution of protein percentage over a lactation cycle as derived from the proposed model (full line) and the Wilmink model (dashed line) for cows that calved in September (a) and February (b). The Wilmink model was fitted using points from the beginning of the curve only, i.e. before alping. Each dot represents the average of milk records per day. When t>245 (a) and between 95 and 210 (b), records from the alp only are used to calculate the average, whilst records from the lowland farm only are included for the remaining time frame. The impact of alping on protein content is considerably smaller compared its influence on milk yield (Fig. 2).



Sup. Mat. S3: Evolution of fat percentage over a lactation cycle as derived from the proposed model (full line) and the Wilmink model (dashed line) for cows that calved in September (a) and February (b). The Wilmink model was fitted using points from the beginning of the curve only, i.e. before alping. Each dot represents the average of milk records per day. When t>245 (a) and between 95 and 210 (b), records from the alp only are used to calculate the average, whilst records from the lowland farm only are included for the remaining time frame. It should be noted that the shape of the Wilmink curve in plot b seems inaccurate, since we typically expect fat content to increase again after the minimum is reached. This is due to the fact that alping starts shortly after the minimum and too few observations are present to model correctly the end of the lactation curve.



Sup. Mat. S4: Effect of influencing factors on protein yield. The effect of influencing factors is tested by investigating the difference in protein yield (protein content * milk production) between two groups of animals coming from contrasted conditions (first and third tertiles, except for THI where second and third tertile are chosen). Each factor is here reported in a separate column. At the top of each column, the factor name as well as the contrasted groups are reported; the group with highest protein yield during alping is chosen as the reference group, highlighted in red. In each barplot, the first bar shows the result for autumn calving, and the second for winter calving. The between-group difference in protein yield during alping is displayed in the top panel, the between-group difference in protein yield during the whole lactation in the intermediate panel, the change in the d-parameter at the bottom. The Δd-parameter indicates how the reference group is impacted by alping compared to the other group, with positive values meaning lower negative impact (see Eq. 4). Significant Δd values are plotted in black, while grey indicates non-significance. Environmental factors affects production during alping only, making a comparison of the whole protein production redundant (which is why no graph is present in the intermediate panel of the concerned variables).



Sup. Mat. S5: Effect of influencing factors on fat yield. The effect of influencing factors is tested by investigating the difference in fat yield (fat content * milk production) between two groups of animals coming from contrasted conditions (first and third tertiles, except for THI where second and third tertile are chosen). Each factor is here reported in a separate column. At the top of each column, the factor name as well as the contrasted groups are reported; the group with highest fat yield during alping is chosen as the reference group, highlighted in red. In each barplot, the first bar shows the result for autumn calving, and the second for winter calving. The between-group difference in fat yield during alping is displayed in the top panel, the between-group difference in fat yield during the whole lactation in the intermediate panel, the change in the d-parameter at the bottom. The Δd-parameter indicates how the reference group is impacted by alping compared to the other group, with positive values meaning lower negative impact (see Eq. 4). Significant Δd values are plotted in black, while grey indicates non-significance. Environmental factors affects production during alping only, making a comparison of the whole fat production redundant (which is why no graph is present in the intermediate panel of the concerned variables).

Solange Duruz
Grandchamp 14
1018 Lausanne
solange.duruz@citycable.ch

Strengths

- EPFL engineer with PhD in biogeoinformatics
- Programming (R, Web-, SQL)
- Languages (French, English, German)

Curriculum Vitae

Education

2015-2020	PhD at EPFL: "Biogeoinformatics for the management of Farm Animal Genetic Resources (FAnGR)"
2009-2014	Master degree at EPFL in Environmental Sciences and Engineering, specialisation "monitoring and modelling of the environment"
2011-2012	Exchange year at the University of Waterloo, Canada
2006-2009	High school at Gymnase de Beaulieu (Lausanne), bilingual French-German
2007-2008	Exchange year in Vienna (Austria), as part of the bilingual high school program

Technical skills

GIS	GIS, Spatial analysis Web-GIS, Database management, SQL, Decision aid
Informatics	Programming (*R, PHP/HTML,Matlab*), Latex, *Microsoft/OpenOffice* suite
Biology	Bioinformatics, treatment of molecular data

Professional experiences

2015-2020	**EPFL** PhD Research: Biogeoinformatics for the management of farm animal Teaching: teaching assistant in GIS and analysis of geodata
2014	**ICRC** (4-month internship) Contributed to the creation of an on-line training platform in hydraulics

Languages

French	Native language
English	Very good skills (Cambridge Certificate of Proficiency), C1-C2 level
German	Very good skills (one year in Austria), C1 level
Italian	Very good skills (option in high school), C1 level

Extra-professional activities

Music (flute for 15 years, 5 years in an orchestra), Hiking, Skiing

Personal situation

29 years old, Married with two children, Swiss nationality

Publications

Papers

Duruz, S., Vajana, E., Flury, C., Burren, A., Joost, S. (2020) Big dairy data to disentangle the effect of geo-environmental, physiological and morphological factors on milk production of mountain-pastured Braunvieh cows. bioRxiv.

Duruz, S., Sevane, N., Selmoni, O., Vajana, E., Leempoel, K., Stucki, S., ... & Joost, S. (2019). Rapid identification and interpretation of gene–environment associations using the new R. SamBada landscape genomics pipeline. Molecular ecology resources, 19(5), 1355-1365.

Vajana, E., Widmer, I., Rochat, E., **Duruz, S.**, Selmoni, O., Vuilleumier, S., ... & Joost, S. (2019). Indication of spatially random occurrence of Chlamydia-like organisms in Bufo bufo tadpoles from ponds located in the Geneva metropolitan area. New microbes and new infections, 27, 54-63.

Duruz, S., Flury, C., Matasci, G., Joerin, F., Widmer, I., & Joost, S. (2017). A WebGIS platform for the monitoring of Farm Animal Genetic Resources (GENMON). PloS one, 12(4).

Stucki, S., Orozco-terWengel, P., Forester, B. R., **Duruz, S.**, Colli, L., Masembe, C., Negrini, R., Landguth, E., Jones, M. R., The NEXTGEN Consortium, Bruford, M. W., Taberlet, P. & Joost, S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. Molecular Ecology Resources, 17(5), 1072-1089.

Leempoel, K., **Duruz, S.**, Rochat, E., Widmer, I., Orozco-terWengel, P., & Joost, S. (2017). Simple rules for an efficient use of geographic information systems in molecular ecology. Frontiers in Ecology and Evolution, 5, 33.

Joost, S., **Duruz, S.**, Marques-Vidal, P., Bochud, M., Stringhini, S., Paccaud, F., ... & Vollenweider, P. (2016). Persistent spatial clusters of high body mass index in a Swiss urban population as revealed by the 5-year GeoCoLaus longitudinal study. BMJ open, 6(1).

Joost, S., **Duruz, S.**, Rochat, E., & Widmer, I. (2016). Open computational landscape genetics (No. e1721v3). PeerJ Preprints.

Bruford, M. W., Ginja, C., Hoffmann, I., Joost, S., Orozco-terWengel, P., Alberto, F. J., Amaral A. J., Barbato, M., Biscarini, P., Colli, L., Costa, M., Curik, I., **Duruz, S.**, ... (2015). Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. Frontiers in genetics, 6, 314.

Conferences

Rochat, E., **Duruz, S.**, Widmer, I., Clémence, A., Desrichard, O., Rappo, D., ... & Joost, S. (2015, March). Relationship between land cover type and Body Mass Index in Geneva. In 2015 Joint Urban Remote Sensing Event (JURSE) (pp. 1-4). IEEE.

Duruz, S. (2014). A WebGIS application for the monitoring of Farm Animal Genetic Resources. In Meetings del Comitato Italo-Svizzero per la Geoinformatica (No. POST_TALK).