EUROPEAN SCIENCE FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

Research Networking Programme
NEWSLETTER No. 5 March 2014

ESF GENOMIC-RESOURCES

Life, Earth and Environmental Sciences (LEE)

Introduction

In 2013, the ESF GENOMIC RESOURCES Research Networking Programme (RNP) decided to fund a special science event named "SANDPIT meeting". In this issue, we present the concept and the main outcomes of this particular way of exchanging knowledge and experience among distinct scientific communities (page 2).

Last year, GENOMIC RESOURCES also funded 7 exchange visits. The results obtained in the context of these projects are presented in this issue (page 5).

Eight exchange visit grants will be funded in 2014 and the selected projects will be described in the following GENOMIC-RESOURCES Newsletter.

What's new?

Call for papers: journal "Frontiers in Livestock Genomics"

Research Topic: Advances in farm animal genomic resources

Deadlines: April 30, 2014, for abstract submission and September 30, 2014, for full paper submission

This Research Topic is proposed by GENOMIC-RESOURCES to constitute a special issue dedicated to the project. Therefore publishing fees related to papers accepted for publication should be covered by the GENOMIC-RESOURCES project. In case the budget allocated is not sufficient, we will apply ESF priority rules (see ESF guidelines). For more information, see:

http://www.frontiersin.org/Livestock_Genomics/researchtopics/Advances_in_Farm_Animal_Genomi/2123

High-throughput SNP genotyping is about to be available for all major farm animal species. The recent development of sequencing techniques calls for new methods of data management and analysis and for new ideas for the extraction of information. However, to make sense of this information in practical conditions, integration of geo-environmental and socio-economic data are key elements. The study and management of farm animal genomic resources (FAnGR) is indeed a major multidisciplinary issue.

The goal of the proposed Research Topic is to collect contributions of high scientific quality relevant to biodiversity management, and applying new methods to new genomic and bioinformatics approaches for characterization of FAnGR, to the development of FAnGR conservation methods applied ex-situ and in-situ, possibly including socio-economic aspects of FAnGR conservation, transfer of lessons between wildlife and livestock biodiversity conservation, and considering

the contribution of FAnGR to a transition in agriculture (FAnGR and agro-ecology).

We invite the submission of abstracts of maximum 1 A4 format page of work related to the focus of the research topic. Authors will be notified by the host editor whether their abstract has been accepted.

Deadlines: April 30, 2014, or sooner for abstract submission - **September 30, 2014,** for full paper submission

Type of contribution, number of words and costs: Book Review - 1,000 words - free | General Commentary - 1,000 words - free | Opinion Article - 2,000 words - free | Mini Review Article - 3,000 words - €575 | Perspective Article - 3,000 words - €575 | Original research - 12,000 words, €770 for associate editors - €960 for other cases | Review article - 12,000 words - €770 for associate editors - €960 for all other cases | Method article - 12,000 words - €770 for associate editors - €960 for all other cases

Sandpit meeting

Social science meets animal science

How to favor scientific integration and capacity building for sustainable conservation of farm animal genetic resources (FAnGR)?
September 9-11, 2013, Louvain-la-Neuve (BE)

Summary
Animal breeders and social scientists are both working in a rapidly changing environment as new technologies - often linked to massive amounts of data - emerge, consumer demands change and global topics (climate change, global trade of animals and their products) arise. Both researcher communities are confronted with complex adaptive systems, but each group has a different focus on these systems. While animal breeders tend to concentrate on technical solutions, social scientists put human beings and their attitudes, actions and behavior in the centre of their research.
A three days Sandpit meeting was organized from 9 to 11th of September in Louvain-la-Neuve (Belgium). Animation was achieved by a duo made of a professional facilitator and a scientific facilitator. This animation and original activities helped to meet the three objectives: (1) a better mutual knowledge between social scientists and animal geneticists, (2) exchange on successful case studies, (3) identification of good rules of interaction.
The audience was made of 12 social scientists and 16 animal scientists. This balance of competences was one of the key elements of success of the meeting. During the meeting both sides got a better understanding about concepts, approaches and methods of the other discipline. It was confirmed that each group of scientists in itself is very heterogeneous. Events such as speed dating were essential to bridge the gap on Day 1. Activities such as beer tasting and farm visit were reinforcing the process.
Based on common understanding, participants could develop in a very free setting new ideas for the identification of key stakeholders and issues during the second day.
The third day was dedicated to further discussion of the modes of convergence between fields, to a second session of

speed dating and to a dynamic and reflexive evaluation of the meeting.

Description of the scientific content and discussion

Day 1 The workshop started with an introduction to the concept of "complex adaptive systems" by the facilitator. The different elements and their linkages were highlighted and some practical examples from the field of animal breeding were given. The idea was to find a common understanding between different research disciplines and identify areas of a common research agenda. Four participants gave a short presentation on their current research projects. This activity opened discussions on different research approaches and methodologies:

- Emerging markets, emerging strategies under the genomic era, by Julie Labatut, INRA, France
- Why do we like cattle? by J.A. Lenstra, molecular geneticist, Faculty of Veterinary Medicine, Utrecht University
- Economics and genetic resources, by Dominic Moran
- Opportunities provided by Gene Editing technology, by Bruce Whitelaw, Roslin Institute

A speed dating session was held, where participants got five minutes to talk to another person. The task given to the participants was to make the other person interested in their research and to identify possible common interest. This exercise was highly appreciated by all participants as it allows familiarizing oneself within a short period of time with a number of different people.

Therefore the organizers agreed with the participants to repeat this exercise during the workshop once more.

The next step was to look at perception of different research groups and what they think of each other. Animal scientists were asked to list characteristics of the work of social scientists. They also had to list what social scientists would think about animal science. The group of social scientists did the same group work in parallel. Outputs of this exercise are available on the meeting website:http://www.genresandpit.eu/wpsp/?page_id=49. At the end the results from both groups were presented to the plenary.

The purpose of this exercise was to stimulate thinking about research cultures of different disciplines. Perceptions and prejudices can be discovered by this technique. This helps to appreciate that there are different ways of doing research. This is seen as a first step that allows in the long run developing joint research proposals. The aim of day 1 was also to "break the ice". Therefore some social activities as a joint dinner and a beer tasting session were organized.

Day 1 concluded with an interesting keynote presentation entitled "Sustainability Science for Strong Sustainability and held by Tom Dedeurwaerede (Université de Louvain). He clearly indicated that real world problems are often of high complexity and therefore inter-and transdisciplinary research is needed to address these questions. He also raised the point that this approach asks for a shift in the structure and administration of universities and research organisations.

On *day 2* the whole team moved to Louvain-la-Neuve University. The morning session was dedicated to identify research areas. This was first done in smaller groups, and then presented to the audience. In a joint effort the ideas were grouped in broad themes. Finally, each participant could choose his/her topic of interest. Based on common interests smaller groups were formed and a discussion about possible research topics and projects were discussed.

After the exercise a visit to the genomic laboratory was organized. This was planned to make it more explicit how laboratory work is done. People had the

opportunity to interact with staff members and get an inside in the routine of a lab.
After lunch a visit to a commercial farm was organized. A goat farmer, who is processing all milk into cheese and sells it on the nearby market of Brussels was visited. This again gave participants the opportunity to interact in a very informal way and discuss different points of views.

On *day 3*, further discussions were organized about the modes of convergence between fields. A second session of speed dating took place on a voluntary basis and the last hour was the occasion of a dynamic and reflexive evaluation of the meeting.
Video footages were filmed in order to be put on the website : www.genresandpit.eu

Assessment of the results and impact of the Sandpit meeting on the future directions of the field

The program allowed certain flexibility in adjusting to the needs and interests of the participants. The wide range of diverse disciplines opened the way for lively discussions and gave insights in the research culture of other disciplines. All participants were open-minded and interested to learn from colleagues and this positive and enabling environment facilitated the learning process. Participants learned what other approaches are currently used and where possible collaborations might be established.

Quantitative evaluation of the Sandpit meeting (17 surveys)
General organisation: Very fine (10) to Fine (7) - *Content*: Very fine (3), Fine (12), Fair (2) - *Animation* : Very fine (8), Fine (7), Fair (1), no answer (1)

Qualitative evaluation
What did you learn ?

- That there is a good community of scientists interested in collaboration across disciplines.
- How to improve interdisciplinary exchanges and communication with colleagues;
- There are similarities and differences between Natural Sciences (NS) and Social Sciences (SS) on how they do research, goals, communication, methodology, etc.
- Much about the social scientific approach to agriculture and diversity.
- That SS have same overall goal; but the approach is reflective and more inclusive than in NS. Diversity within groups is as large as between groups – this adds value to the discussion.
- Organisational models are changing; Animal scientists can benefit from SS to generate hypotheses in an evolving social context; we need to work together;
- Meeting geneticists that I knew by name;
- Exchange of news in my old field;
- Sociologists are working with genetic/natural resources;
- Move towards a holistic approach to develop sustainable models;
- NS and SS have a good mutual understanding and of themselves;
- I have to take the first step to work with social scientists;
- Importance of communication processes;
- Confirmation of some 'ideas'. I already need cooperation with other fields.
- That NS and SS know more about each other than I thought;
- I have learned about complexity of livestock conservation as well as capability of social scientists to properly communicate with the society;
- Yes we can find new ways of working ; new ways to create positive interactions among disciplines.
- Communication needs time & attitude to be able to listen;

- Combining the contributions of natural and social scientists in the same project is a real challenge;
- I learned that communication can be favored by a good animation.;
- Importance of collaboration between NS and SS to highlight the importance of FAnGR management among policy makers and citizens.

What are the points to improve in case of another Sandpit event?

- Propose more challenging exercises to develop research proposals. It was almost perfect!
- A better balanced number of participants among the two groups;
- Have reward = identified funding opportunity / POT;
- Making sure having more balanced groups of SS (we need more) and NS
- Less workload for workshop leaders ('we are tired');
- Avoid obvious areas (content of project proposal) ;
- Avoid trips longer than 1/2 hour;
- No computers in conference room!
- I would like a little more stringent schedule;
- Great to have man focused on "getting to know people better". But having an additional focus on animation;
- To put information to funders would have been helpful;
- In the different activities, force more the encounter between NS and SS;
- Narrow the focus;
- Clarifying the aims;
- Stronger theoritical contextualization in the beginning of the meeting.

Perspectives ?

- Circulation of information on interdisciplinary proposals that have been successful;
- Information about funding sources and how they deal with interdisciplinary proposals;
- Putting in proposal ideas to DG Research for future calls;
- Writing up of a manifesto to strengthen the research community;
- Use of the web site to place useful documents, recommended lectures e.g. why genetic resources are important.

Exchange grants 2012
Funded projects and final reports

Project N°1

Shedding light on local adaptation in livestock by means of landscape genomics and whole-genome sequencing (LightStock)

Sylvie Stucki, Laboratory of Geographic information systems (LASIG) School of Architecture, Civil and Environmental Engineering (ENAC) École Polytechnique Fédérale de Lausanne (EPFL)

Purpose of the visit

The visit aimed at assessing the genetic diversity of sheep and goats in Morocco with whole genome data. The animals were sampled and sequenced during the project NextGen in which both host and guest are involved. However the delivery was delayed and whole genome data was not available for the visit. Thus the host and the guest agreed to rather analyse SNP datasets from Ugandan cattle that were already provided by NextGen partners. The 917 individuals sampled in Uganda had been genotyped in two batches: 813 individuals using the 54k BovineSNP50 BeadChip assay and 102 distinct individuals using the 800k BovineHD

BeadChip assay (Illumina Inc., San Diego, USA).
The main topics addressed during the visit were the population structure, the detection of molecular signatures of selection in relation with the environment (landscape genomics) and the fine-tuning of Samβada, a software focusing on spatial analysis of genomic data.

Description of the work carried out during the visit

Both datasets were filtered with a call rate of 95% for SNPs and individuals, the minimum allele frequencies (M. A. F.) were set to 1% for the first group and 5% for the second, resulting in 804 samples genotyped for 41,215 SNPs and 102 samples genotyped for 634,849 SNPs respectively (Purcell et al., 2007).
The environment was characterised with the WorldClim dataset, which consists of minimum, maximum and mean monthly temperatures, monthly amount of precipitation and 19 derived variables at a resolution of 1 kilometre (Hijmans et al., 2005). The topography was described with the digital elevation model SRTM which mesh is 90 meters (Farr et al., 2007). The slope and curvature were derived from the altitude. Environmental data was prepared with SAGA GIS (www.saga-gis.org) and the values corresponding to the sampling locations were extracted in Quantum GIS (www.qgis.org). A total of 72 environmental variables were included in the analysis.

Population structure

The first analysis of population structure was processed with BAPS on the 54k dataset (Corander and Marttinen, 2006). However this software did not manage to solve the population structure for the 800k SNPs dataset in a reasonable amount of time. Thus both analysis were repeated with Admixture (Alexander et al., 2009), which uses a different algorithm to cluster individuals into populations.

Landscape genomics

The detection of selection signatures was carried out with Samβada which models the frequency of each genetic marker with logistic regressions on the environmental variables (Joost et al., 2008). The significance of the association is assessed by both log likelihood-ratio (G) and Wald tests. The analysis reveals which genomic regions are subject to selection.
The study also allowed for testing and improving three features of Samβada:

- Multivariate models are assessed against the simpler nested models, so the gain in prediction is worth the added complexity.
- The module for spatial autocorrelation was completed with the computation of significance levels for local Moran's I, a Local Indicator of Spatial Association (Anselin, 1995).
- The interface between Samβada and other bioinformatic softwares was improved by providing a module to recode PED and MAP files, a popular format for SNP data (Purcell, 2009), into input files for Samβada.

Description of the main results obtained

Population structure

The original filtering of 54k data kept 786 individuals and 38,597 SNPs. The best classification found by BAPS consists of four populations shown on Fig 1. The classification provided by Admixture on 804 individuals and 41,215 SNPs is similar: 771 out of 785 common samples were classified the same way by both algorithms. The available pictures of the animals were sorted by cluster based on BAPS results. As shown on fig. 2, the two large clusters correspond to the Zebu (no 2) and Ankole (no 3) cattle populations. No pictures were taken for the small clusters 1 and 4. However their small sizes and the location of cluster 4 around Kampala might indicate a recent introgression while the cluster 1 might also correspond to an hybrid of Zebu and Ankole. These hypotheses need further investigation.

Landscape genomics

Detection of loci under selection

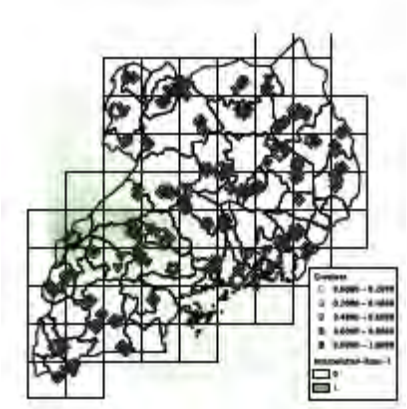
When recoded as binary variables, the 54k dataset lead to 120,869 polymorphic markers. Samβada processed 8 millions univariate models with these markers and 72 environmental variables. Out of them, 46,862 models were significant at p=0.01 (score threshold=37 after Bonferroni correction). Table 1 shows the most significant models.

Marker	Env_1	Gscore	WaldScore	Type of correlation
Hapmap39368-BTA-104532_A A	tmax11	289.29	173.06	+
Hapmap39368-BTA-104532_A A	prec7	289.16	182.93	+
Hapmap39368-BTA-104532_A A	tmax12	284.49	172.37	+
Hapmap39368-BTA-104532_A A	latitude	277.88	185.05	+
Hapmap39368-BTA-104532_A A	bio4	262.50	173.98	+
ARS-BFGL-NGS-36736_AA	prec7	268.09	176.51	+
Hapmap44320-BTA-95767_AA	prec7	267.20	176.10	+
Hapmap44320-BTA-95767_AA	latitude	265.07	180.58	+
ARS-BFGL-NGS-107270_AA	prec7	266.59	175.99	+
ARS-BFGL-NGS-114888_GG	prec7	253.73	171.19	+
ARS-BFGL-NGS-114888_GG	latitude	244.50	171.94	+
ARS-BFGL-NGS-31523_GG	prec7	253.65	171.38	+

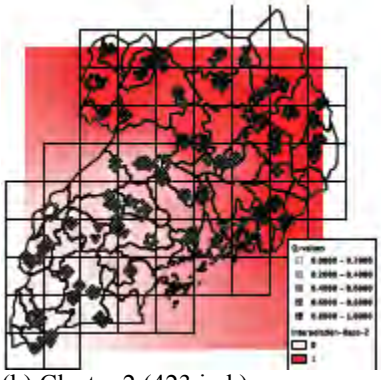
Table 1: Most significant models with 804 individuals and 41,215 SNPs (=120,869 binary markers).

The first column is the marker name, formed by the SNP name and the allele, then come the name of the environmental variable, the log likelihood-ratio (G) score, and the Wald score. The last column shows the sign of the correlation between the marker frequency and the value of the environmental variable. These loci are located on the X chromosome.

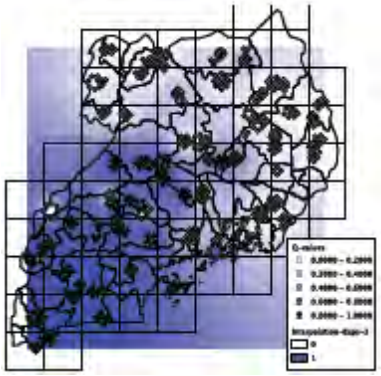
When the same approach was applied to the 800k dataset, no model was significant (p=0.01 before Bonferroni correction). This is explained by the large amount of binary variables (1,868,310) which lowered the significance threshold to $p=7.43\text{-}10^{-11}$ (score threshold: 42.4). No model had a significant p-value for the Wald test with 102 individuals. An alternative model selection was considered using False Discovery Rate (Benjamini and Hochberg, 1995) instead of Bonferonni correction. FDR controls the rate of false detections rather than the familywise error rate, thus it is more liberal than Bonferonni



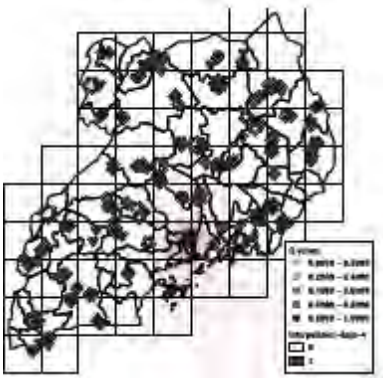
(a) Cluster 1 (22 ind.)



(b) Cluster 2 (423 ind.)



(c) Cluster 3 (330 ind.)



(d) Cluster 4 (11 ind.)

Figure 1: Population structure assessed by BAPS on 786 cattle and 38,597 SNPs for four clusters.
Each point stands for an individual. The darker the point, the higher the membership coefficient to the cluster. Points are arranged in circles around farm locations to avoid overlays. The background color interpolates individual' coefficients to show the regions were the populations are most commonly found.

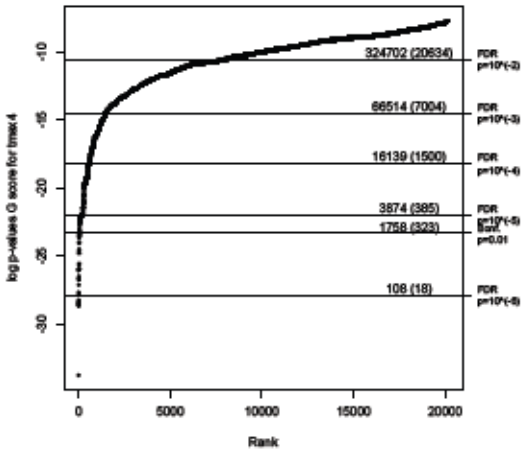


Figure 3: Distribution of pvalues for regression models with maximum temperature in April. Each horizontal line shows a possible threshold, either using Bonferroni (Bonf.) correction or False Discovery Rate (FDR).
The labels indicate the type of correction and the p-value for each level, along with the number of significant models and the
number of associated SNPs in parenthesis.

However no Wald score was significant with the considered version of FDR since it requires that at least one model passes the Bonferonni test. Therefore the analysis of results focused on the G score.

Figure 3 shows the distribution of the log p-values of G scores for all models involving the maximum temperature in April. This variable was commonly found as an accurate predictor for marker distributions. Significance threshold was set at p=0.01 with Bonferonni correction,

which lead to 1,758 significant models involving 323 SNPs. These loci were spread between chromosomes 5 (42 SNPs), 14 (4 SNPs) and X (277 SNPs). Latitude was often highly correlated with marker frequencies.

Gene mapping

The distributions of the significant loci on chromosome 5 are shown on fig. 4 for several thresholds. The most significant model involves the SNP BovineHD0500019261, this loci maps to the gene CHST11 which is involved in cartilage make up (Flicek et al., 2012).

The second cluster detected on chromosome 5 maps to an uncharacterised gene ENSBTAG00000033726 while the most significant SNP on chromosome X (BovineHD3000015663) is located near a conserved genomic region in 36 eutherian mammals.

Spatial autocorrelation

Logistic regression fit a global model for the presence of a marker, while spatial analysis provides information about local behaviours. Local Indicators of Spatial Association (LISA, Anselin, 1995) compare the value of a variable in each location with the weighted mean of its values in the neighbouring points. LISA are the local equivalent of spatial autocorrelation.

A bivariate LISA compares the value of a variable to the mean of another variable over the neighbouring points. Fig. 5 shows local correlation between the presence of the marker BovineHD0500019261_GG (allele GG) and the maximum temperature in April, computed with local Moran's I (LISA, Anselin, 1995).

The map shows a positive correlation in North and South Uganda separated by a non-significant region.

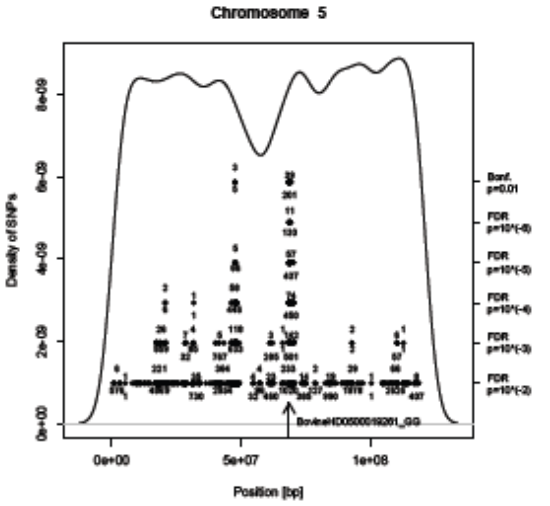


Figure 4: Solid line shows the overall SNPs density on chromosome 5. Horizontal plots represents the SNPs that were detected for different thresholds. These SNPs were grouped when they were closer than 2 106 bp. Each cluster is summarized by the number of SNPs it spans (below) and among these, the number of SNPs under selection (above). The vertical spacing between plots is arbitrary.
The arrow points out the SNP BovineHD0500019261.

Summary

The spatial distribution of marker BovineHD0500019261_GG is similar to the spreads of the Zebu and Ankole populations. Most environmental variables are also correlated with latitude. Two processes could explain these observations:

- Zebu and Ankole living areas overlap with the North-South environmental gradient in Uganda. The correlations measured between environmental variables and genetic markers are due to the demographic structure of Ugandan cattle.
- The spatial distribution of Zebu and Ankole in Uganda is influenced by natural selection, either by a climatic feature that follows a North-South gradient or by an unobserved environmental condition. The most likely candidate is the distribution of the tsetse fly, which transmits trypanosomiasis. A different resistance to this disease could explain the spatial distributions of these breeds.

The following analysis are ongoing to test these hypotheses:

- Separate studies of Ankole and Zebu populations: If the same loci are detected in both groups than in the overall analysis, these markers could result from a global adaptive process. If the study reveals different markers in Ankole and Zebu, these markers may show signatures of selection in each population.
- Multivariate studies including population membership (q-value) as a cofactor will allow to fit models where the population structure is taken into account.
- Comparison of breeds and markers distributions to parasites prevalence, especially the tsetse fly, to test whether they overlap.

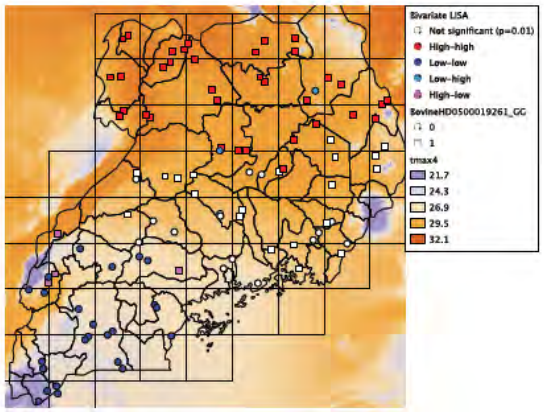


Figure 5: Figure 3: Bivariate local Moran's I between BovineHD0500019261_GG and the maximum temperature in April (background layer) for the 102 Ugandan cattle. This indicator measures the spatial correlation between the state of the marker and the temperature averaged over the 20 nearest sampling points. Dots shape indicate where the marker is present (square) or absent (circle) and their color shows the type of association (red=high-high, dark blue= low-low, pink=high-low and light blue=low-high, white=non-significant (p=0.01, 10'000 permutations). The sampling phase was planned following a regular grid to ensure an even spatial representation.

Future collaboration with host institution

The project NextGen is ongoing. Host and guest are still working on Ugandan cattle and will carry out in concert the analysis of whole genomes of Moroccan sheep. This visit was also the opportunity to tighten the links between host and guest for future collaborations.

Projected publications / articles resulting or to result from the grant

Samβada will be presented in a publication before its open-source release. This article will include the study on Ugandan cattle. The results obtained during the stay will also be presented in the next conference of the International Association of Landscape Ecology on 9-12th September 2013 in Manchester. The European Science Foundation will be acknowledged for its support on both occasions.

Other comments

This visit was very helpful and instructive for my research and I thank the European Science Foundation for making it possible. I am very grateful to Prof. Mike Bruford, Dr. Pablo Orozco-terWengel and all members of the laboratory for their warm welcome and our fruitful collaboration.

Project N°2

Genome wide association study for functional traits in cattle

Gábor Mészáros, post doc researcher at University of Natural Resources and Life Sciences, Vienna, Austria

Purpose of the visit

The main objective of the stay was to carry out a genome-wide association studies (GWAS) for a range of functional traits in cattle. GWAS looks for associations between genotypes of the SNPs and some trait of interest. Most of these traits are continuous they are likely to be controlled by many loci of small effects, or a mixture of a few genes with large effects and many genes of small effects. It is challenging to perform analysis of GWAS data because the number of SNPs (p) is much larger than

the sample size (n), commonly referred to as the "small n, large p" problem. A major difficulty in this problem is that the number and extent of spurious associations between predictors and response increase rapidly with increasing p.

There are several methodologies to perform genome wide associations, single SNP associations being the most common. In this case each SNP is considered as a single effect in a linear model, running as many times as many SNPs are considered in the analysis. Methodologies, such as lasso, ridge regression and elastic net with different penalty factors are simultaneously accounting for the effects of SNPs. Another possibility is to implement Bayesian variable selection, such as Bayesian lasso to find important regions for the traits of interest. Resulting effects for any of these methodologies might be caused by real association to the trait, or just be a product of "population structure" in the data set caused by e.g. differences in breeds, admixture levels or differences in ancestry for the analyzed data set. For this reason the implementation of a correction for population structure is needed to remove false positive results. Likewise, careful steps need to be taken not to delete associated markers, and so to avoid false negative results.

Description of the work carried out during the visit

The original plan was to use the 50k Illumina SNP chip genotypes for Pinzgau and Tyrol Grey cattle (~220 genotypes each) and the 50k genotype data for ~2.000 Austrian Fleckvieh bulls available for the "Genome wide association study for functional longevity and related traits in dairy cattle" project. After submission of the ESF application we got the permission to use 50k genotypes from the joint German-Austrian (DEA) genetic evaluation for ~6.000 Fleckvieh bulls. As the estimated breeding values, deregressed breeding values for all traits and all reliabilities were available for this big set, we decided to use the DEA set to perform genome wide analysis on a range of traits.

The PLINK (Purcell et al. 2007) software was used for quality control. Only SNPs that could be unambiguously mapped following the paper of Fadista and Bendixen (2012) were kept for the analysis. Also SNPs at sex chromosomes were removed. The R software (R Core Team, 2012) was used to perform the analyses and visualize the results. Due to the large number of genotypes involved the handling of the data and the actual analysis was challenging, especially when it involved correction for population structure. For this purpose the eigenvector decomposition with GEMtools R package was used. The computed eigenvectors were then fitted in the model together with the SNP effects. The computations were carried out using the Vienna Scientific Cluster. Unfortunately even this huge server allowed running only one or two analyses at the same time, depending on the methodology, supposedly due to memory limitations.

The paper "Evaluation of the lasso and elastic net in genome-wide association studies" previously submitted to Journal of Animal Breeding and Genetics was reviewed and we decided to improve it before resubmission. The GWAS on a real data set will be changed from longevity to fat content, to show the features of different methodologies. This however needs a re-analysis of each model type with and without population structure correction, which is currently under way.

For the "Genome wide association study for functional longevity and related traits in dairy cattle" project, deregressed breeding values for milk production, fat content, longevity, fertility, calving ease (maternal and direct), stillbirth rate (maternal and direct) and somatic cell count were analyzed using a single SNP association with and without population structure correction. The effect of using different

numbers of eigenvectors for population structure correction was also explored.

Description of the main results obtained

The significance values from each model were extracted and transformed to a negative logarithm, so the higher values denote higher significance level. All –log(p) values were plotted, distinguishing the chromosomes using different colors.

The number of eigenvectors used for population structure correction was studied in subsequent runs, when deregressed breeding values for fat content were used as phenotypes. For this trait there is a huge signal on chromosome 14, presumably DGAT, with –log(p) values up to 150. To show the changes in smaller peaks we limited the y axis to 40.

Figures 1-3 show the –log(p) for no population structure correction, 15 and 117 eigenvectors used. The three scenarios intended to compare situations with no, low and high number of eigenvectors in the model. The 15 eigenvectors were chosen based on numbers found in
the literature (Hao et al. 2010). The 117 determined by the GEMtools package as the number of significant dimensions for the ~6.000 genotypes.

Some of the peaks visible with 15 eigenvectors in the model vanish when using the high number of eigenvalues for population structure correction. Notable examples are on chromosome 5 and 11, smaller previously significant SNPs on multiple other chromosomes. From these results it is not clear if the approach gets rid of false positives or some of the SNPs become false negative when using many eigenvectors. A more detailed look into this problem is needed.

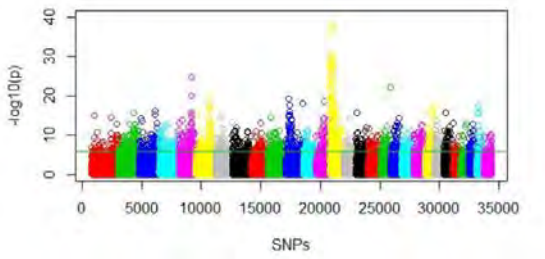


Figure 1 Single SNP analysis without correction for population structure

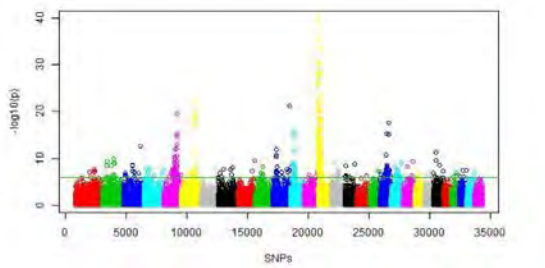


Figure 2 Single SNP analysis using 15 eigenvectors for population structure correction

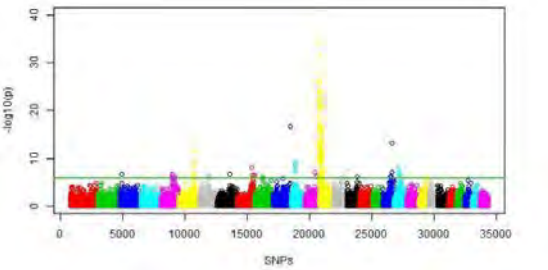


Figure 3 Single SNP analysis using 117 eigenvectors for population structure correction

In addition to the fat content presented above, single SNP analyses with and without population structure correction were conducted during the stay. Interestingly, for most of the traits there were SNPs above the Bonferroni line, which is a very conservative form of significance testing. Only some of the results are shown in this document, but based on the results obtained during the research stay a wide range of additional, more detailed analyses will be done.

The GWAS results for longevity are shown in figure 4. Previous analyses with a smaller data set of ~2.000 Fleckvieh bulls did not show any significant SNPs. With this data set we obtained several significant

SNPs indicating signals on chromosomes 13, 14 and 19.

Results from single SNP analysis for fertility after population structure correction are shown in figure 5. Also here we see some interesting SNPs, but these are much less clear, not supported by many neighboring SNPs as in the previous cases.

For direct stillbirth rate there are two very clear signals on chromosomes 14 and 21, and several additional significant ones on chromosomes 5, 6, 10 and 17. The graph is very similar for direct calving ease (not shown), with the major peaks at the same chromosomes. The results confirm analyses published by Pausch et al. (2011), using a smaller set of Fleckvieh bulls.

In general, the obtained results provide exciting research opportunities for the future.

One of the first steps will be to check the false positive rate using the qvalue R function, which identifies the important SNPs and corrects for the false discovery rate. A trial run for one of the traits was conducted during the stay. Depending on the significance threshold, different numbers of significant SNPs and false positives can be expected.

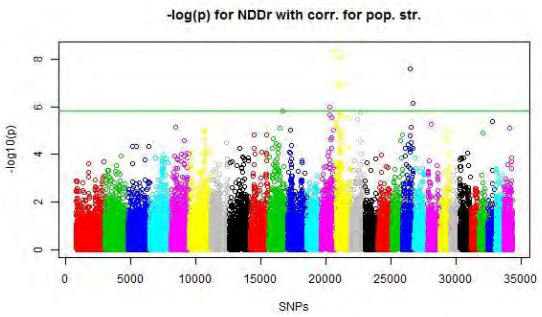


Figure 4 Single SNP analysis for longevity with population structure correction

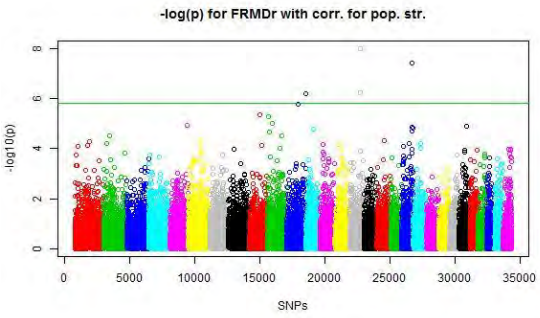


Figure 5 Single SNP analysis for fertility with population structure correction

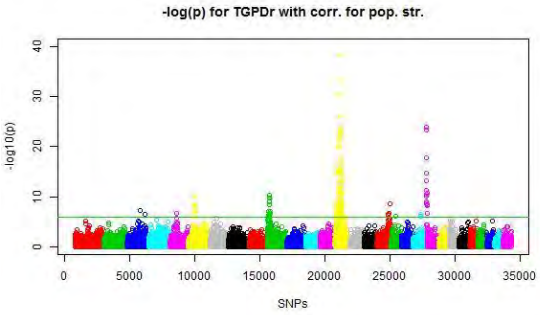


Figure 6 Single SNP analysis for longevity with population structure correction

The results from lasso, ridge regression and elastic net with diverse penalty factors are under way. In general the ridge regression is selecting the most SNPs, lasso the lowest number. The penalty factor in elastic net can be between 0 and 1 - the “proportion” of lasso in the analysis. If the penalty factor is 0 then the analysis is a ridge regression, if 1 the analysis is equal to a standard lasso. With increasing value of the penalty factor the number of selected SNPs is decreasing in our study, as expected.

We have also worked with the Bayesian lasso during the stay, using the bLASSO (Hans, 2010) R package. The run on the full data set was not possible due to computational constrains. Even with only 2.000 iterations and 1.000 burn-in steps the computation took about 3 days on the Vienna Scientific Cluster. After various convergence testing methodologies using the CODA R package we found that the number of iterations should be increased. Because of this we decided to go for a chromosome wise analysis using the

bLASSO, which is a more feasible option even with many more iterations. The significance threshold for the results is the 0.5 posterior inclusion probability (PIP). The results using 20.000 iterations on chromosome 14 are shown in figure 7. Although the number of iterations was much higher in this case, the follow up tests showed that some of the SNPs still did not converge. Additional increase and fine tuning of the parameters is needed.

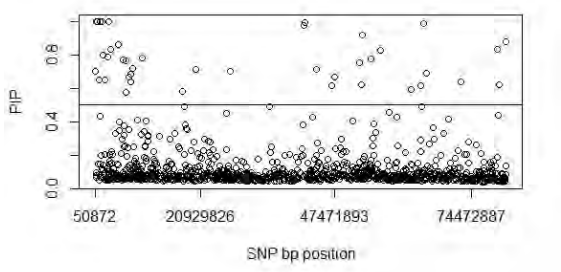


Figure 7 Posterior inclusion probability (PIP) for SNPs on chromosome 14 using fat content as phenotype

Future collaboration with host institution

The collaboration between the two groups at University of Natural Resources and Life Sciences, Vienna and Linköping University was non-existent before this project. This stay however opened new opportunities for the two groups to jointly work on statistical analysis of genotype data. A visit of Dr. Waldmann to BOKU is planned for summer 2013, with continued work on GWAS in cattle.

Projected publications/articles resulting or to result from your grant

An oral presentation of the results will be given (already accepted) at the 64th Annual Meeting of the European Federation of Animal Science in Nantes, dealing with diverse methodologies in GWAS. The paper "Evaluation of the lasso and elastic net in genome-wide association studies" will be resubmitted to a peer reviewed journal, currently we consider Frontiers in Genetics, upon completion of the results. One more paper is under preparation on genome wide association in functional traits in cattle.

Project N°3

Sequence-based characterization of farm animal biodiversity: revealing the genetic basis of body size variation in domesticated pigs.

Christian Reimer, Department of Animal Sciences, Georg-August-University Goettingen

Purpose of the visit

The Göttingen Minipig is one of the smallest pig breeds in the world (SIMIANER AND KÖHN, 2010). Therefore it is a promising candidate to clarify the huge variety in body size of domesticated pigs. Nowadays, the whole genome sequencing technology gives us the possibility to have a detailed look on at animals genetic information at relatively low costs. We took the decision to sequence a group of our Göttingen Minipigs and some individuals of another miniature pig breed, the Berlin Minipig (MiniLEWE) and to combine this information with publicly available sequencing information from other domesticated pigs and some wild boars as well as with several outgroup species such as the African warthog or warty pigs from the Philippines. In order to find a experienced partner to work on this project, we aimed for a collaboration with Dr. Carl-Johan Rubin from the Biomedical Center of Uppsala University. This workgroup has rich experience with the analysis of sequencing data to find proofs of domestication in animals (RUBIN ET AL., 2012) as well as access to a very powerful server cluster. A three month research stay in Uppsala was arranged to conduct the

basic data preparation and to run first analyses to find differences between miniature and normal sized breeds. Simultanously, DNA samples of ten individual Göttingen Minipigs, two Berlin Minipigs, a DNA-pool from Berlin Minipigs and a pooled sample consisting of DNA from animals of the founder breeds of the Göttingen Minipig where queued at the sequencing facility in Uppsala.

Description of the work carried out during the visit

Unfortunately, the commissioned sequences where not ready until the end of the stay in Uppsala due to a long queue at the facility but will be processed during the summer. Alternatively we were able to download the whole genome of a lately sequenced Göttingen Minipig from another study (VAMATHEVAN ET AL., 2013) which was deposited in the European Nucleotide Archive and the genome of a Wuzhishan Minipig from China (FANG ET AL., 2012). As the representatives of normal sized pigs, we downloaded 37 domestic pigs, either from Europe or Asia, 11 wild boars and 6 warthogs as outgroup animals, which were underlying material of the studies by GROENEN ET AL. (2012) and RUBIN ET AL. (2012)

We used the susScr3 (build 10.2, ARCHIBALD ET AL., 2010) available in the UCSC genome browser as the reference genome. This reference was indexed using BWA (LI AND DURBIN, 2009). Afterwards the downloaded sequences were aligned against this reference with the short-read aligning algorithm of BWA (LI AND DURBIN, 2009). The resulting BAM files were sorted with samtools (LI ET AL, 2009) and PCR duplicates were marked with Picard-tools (PICARD, 2009). The resulting BAM Files were indexed with Picard accordingly.

We then evaluated the depth of coverage of the single samples with GATK (MCKENNA ET AL., 2010, DEPRISTO ET AL., 2011). Since different individuals had different average depth, we normalized the results from GATK, so that every individual had the same average depth across the whole genome and summarized them in windows along the genome with a custom made script.

The final SNP calling was done with the Unified Genotyper from GATK with default options for both, single nucleotide variants (SNV) and Indels. Statistics on the quality parameters were built in order to have a basis for deciding on an appropriate filtering. Afterwards we removed all outgroup animals and all SNPs in which a variant allele was only present in the outgroups. The first subsequent filtering only on SNVs with GATK VariantFiltration used the following options: SNP Cluster were removed, if there were more than 5 SNPs in a range of 20 basepairs. A SNP was removed, if either the BaseQualityRankSum, the MappingQualityRankSum or the ReadPosition-RankSum were lower than -6. In addition FisherStrandValues higher than 26 were removed. Filtering on mapping quality was not carried out, since we wanted to keep SNPs which were only present in the minipigs or just a few more domestics or wild.

As the final and most important filter we used a custom made script to exclude loci with an insufficient or extremely high depth. Therefor we calculated the distribution of the depth of coverage on chromosomes 3, 13, X and Y over all domestic pigs, wild boars and the minipigs. We decided to filter away all loci with a coverage lower than approximately half of the mean coverage, i.e. 150 X and all positions with a coverage of roughly two times the mean coverage without outgroups, i.e. 630 X.

The filtered dataset was annotated with genes from the Ensembl (FLICEK ET AL.,

2013) database using the software Annovar (WANG ET AL., 2010).

The next step was the calculation of FST values for the breed contrast of domestics against wild boars and minipigs to determine regions with high diversification between these groups. The FST statistic was calculated after the formula by WEIR (1996):

$$F_{ST} = \frac{\frac{n_i(p_i - p)^2}{(r-1)n}}{p(1-p)}$$

Where is the allele frequency of the first allele over all subpopulations, is the the allele frequency of the first allele within the subpopulation , is the number of individuals of a subpopulation , is the average subpopulations size and is the number of subpopulations. The term had to be corrected, because it overestimated the FST values in a systematical manner. We exchanged against , so that the maximum value to occur was 1. To reduce the effect of outliers and to reduce the number of data points, the resulting FST values were summarized in 40 and 10 kilobasepair windows which were 50 % overlapping. To identify windows with extraordinary high values, a treshold was calculated. The threshold was the lowest FST value of the top 0.1 % quantile of all windows. Windows with a higher value were identified.

To determine the composition of the sequenced animals from their ancestral breeds and to figure out a reasonable number of founder breeds, we used the program “Admixture” (ALEXANDER ET AL, 2009). Since “Admixture” is not able to take LD between SNPs into account properly and we had no knowledge of the actual LD structure within our samples, we performed the analysis for several marker distances (~8 kb, ~39 kb and ~117 kb) over all autosomes as well as for chromosome 1 with an average marker distance of ~1 kb.

Description of the main results obtained

In general, we obtained an extremely large and valuable high quality dataset for further analysis during the stay in Uppsala. It gives us the possibility to compare pigs from different continents, different levels of domestication and breeds with highly different properties with each other. On the other hand our project is lagging at the moment, because our target was to include a lot of individuals from miniature pig breeds. Since they will not be ready until August, we had to manage the first steps with only two minipigs from public sources and not of the same breed.

The analysis of the sequencing depth before filtering revealed, that the domestic, wild and outgroup animals were sequenced at an average depth of 6 up to 8 X. The two minipig genomes resulted from two assembly studies so that in total approximately 80 X were available. These samples were downsampled to an average coverage of at least 12 up to 15 X. While plotting the normalized data for every group (domestic, wild, mini, outgroup) we found regions of extremely high coverage values. The plot for chromosome 7 is shown exemplarily (Figure 1).

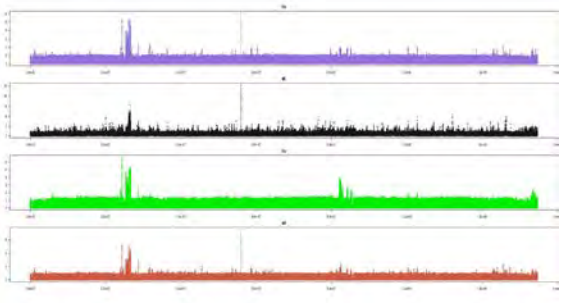


Figure 1: Sequencing depth of chromosome 7, divided by subpopulations. The plot shows the domestics, minipigs, outgroup animals and wild boars (top to bottom). It is remarkable, that there are regions, which show a different pattern between the groups, i.e. around 56000000 bp where the peak is not present in the outgroup animals or the region

around 25000000 bp where the pattern in minipigs looks different than in all other groups.

After the final filtering the SNP set contained about 30.1 million SNPs on 18 autosomes and in the unknown regions. The sex chromosomes where processed separately but not included in further analysis. **Table 1** shows the number of SNPs per chromosome and in total.

Table 1: Number of SNPs after filtering

Chromosome	No. of SNPs	Chromosome	No. of SNPs
chr_1	2936902	chr_11	1189587
chr_2	1920120	chr_12	865076
chr_3	1739426	chr_13	2307108
chr_4	1719589	chr_14	1846826
chr_5	1380430	chr_15	1687854
chr_6	1830387	chr_16	1184457
chr_7	1677898	chr_17	966280
chr_8	1797958	chr_18	870279
chr_9	1943120	chr_Un	990819
chr_10	1282471	**Total**	**30136587**

The calculation of F_{ST} statistics between the domestics, wild boars and minipigs revealed regions with a remarkable diversification. Taking the 0.1 %

most important windows into account, there were 116 windows. Figure 2 gives an overview for all autosomes.

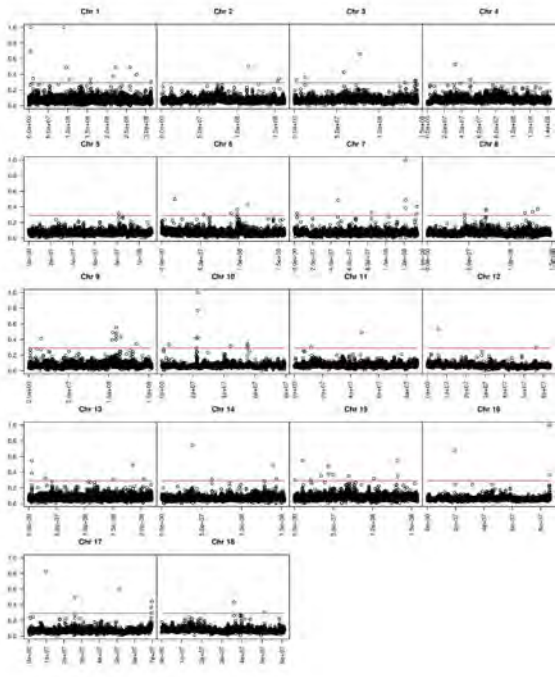


Figure 2: F_{ST} statistics for domestic pigs, wild boars and minipigs, summarized in 40 kb windows with 50 % overlap for all autosomes.

To determine the breed composition of the included individuals, we used the program "Admixture". We choose the cross-validation argument to decide for a correct K-value. Table 2 shows the attributes of the compared runs.

Table 2: Attributes of all "Admixture" runs

Region	Chr 1	Chr 1:18	Chr 1:18	Chr 1:18
No. of SNP icluded	300000	314445	62883	20956
Sum of chromosome length [kb]	315321	2450713	2450713	2450713
Average marker distance [kb]	1.051	7.793	38.972	116.945
Results of the Cross Validation				
CV error (K=1):	0.66975	0.67278	0.67466	0.67548
CV error (K=2):	0.55316	0.56820	0.57104	0.56675
CV error (K=3):	0.54821	0.57610	0.57267	0.56856
CV error (K=4):	0.59074	0.61469	0.61545	0.61073
CV error (K=5):	0.63071	0.64771	0.65411	0.65021
CV error (K=6):	0.63207	0.64958	0.66459	0.65760

As the cross validation results show there are only slight differences between the errors in different scenarios. The final results of the admixture analysis show that there is no real dependency on the marker density in our case. In all scenarios with K=2, "Admixture" detects a clear differentiation of Asian and European breeds. The Wuzhishan minipig clusters clearly with the Asian cohort, whereas the Göttingen minipig shows genetic admixture of Asian and European origins. When we choose K=3, a clear fractioning into european wild boars, Large White related strains and asian origins could be seen. It is remarkable, that breeds like the Landrace, Pietrain and Hampshire shared important genome sections with Large White and the remaining sections with wild boars, whereas the Durocs cluster perfectly with the European wild boars. In that case, the Wuzhishan minipig still showed perfect affiliation with the Asian breed, whilst the Göttingen minipig carries minor parts from the Large White and the European wild boar clusters. **Figure 3** shows the results for the analysis of all autosomes with 20000 SNP markers exemplarily.

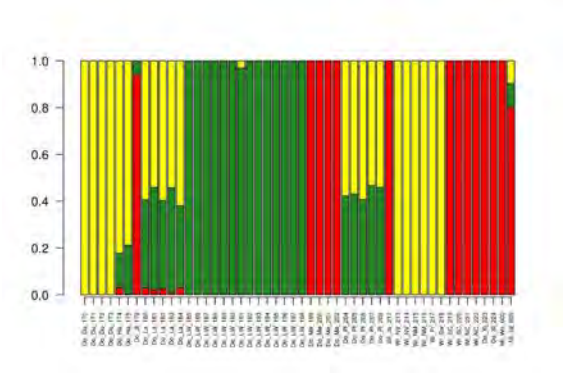


Figure 3: Genetic admixture of the project animal with K=3 estimated from 20000 markers equally distributed over all autosomes. (Group encoding: Do=domestic, Wi= Wild boar, Mi=Minipig

Breed encoding: Du=Duroc, Ha=Hampshire, Ji=Jiangquhai, La=Landrace, LW=Large White, Me=Meishan, Pi= Pietrain, Ja= Japanese wild boar, NV/NM= Wild boar Netherlands, Fr=Wild boar France, Sw= Wild boar Switzerland, SC/NC=Wild boar South/North China, Xi=Xiang, Wu= Wuzhishan, Gl= Göttingen Minipig).

Future collaboration with host institution

As mentioned before, we are awaiting another 10 whole genome sequences from our Göttingen Minipigs, two from Berlin Minipigs and two from DNA pools from Berlin Minipigs and some founder animals respectively, which are being processed in Uppsala at the moment. When the facility in Uppsala finished sequencing, the data will be transferred to the Uppmax server cluster where it will be handled with the same pipeline described in section 2.. With the additional data we will be able to do more powerful analysis, since we were limited by the low number of only two minipigs up to now.

Within this collaboration we scheduled a meeting of Prof. Simianer, Dr. Rubin and Christian Reimer in Göttingen in September to discuss the approach for the new data and for Dr. Rubin to give a presentation on sequencing techniques to students and scientific stuff. In return, Prof. Simianer and Christian Reimer will visit Uppsala in October/ November for a short stay to present first results and to discuss further proceeding.

Christian Reimer will return to Uppsala for a one month stay in November/ December to conduct final analysis and to prepare the publication of our results.

Projected publications / articles resulting or to result from the grant

The results presented in this report will also be shown on the Annual meeting of the German Society of Animal Breeding in Göttingen on September 5th and 6th.

We are waiting for the new data to come to decide for a publication in a peer reviewed journal.

Other comments

The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2010044.

Project N°4

Transcriptome analysis to characterize proli_c sheep *(Ovis aries)* breeds.

Kisun Pokharel, MTT Agrifood Research Finland

Purpose of the visit

The objective of my visit was to learn transcriptome analysis methods. In my project, I am analyzing miRNA and mRNA data from di_erent sheep breeds with varying litter sizes. Genes affecting fertility traits in sheep are complex and in addition, the sheep genome is poorly annotated. Therefore, I wanted to get overview of the methods that are employed in analysing sheep transcriptome. Also, I wanted to get exposure of good practices in working with big genomic data - how to manage big data and get the most out of it.

Description of the work carried out during the visit.

During my exchange period, I discussed results from mRNA data analysis and annotated transcripts resulting from assembly of reads that were mapped against sheep reference genome. I also made de novo assembly of RNA-Seq data. In addition, I designed a pipeline to analysis miRNA data.

De novo assembly

Trinity [1] program was employed for de novo assembly of sheep transcriptome. Trinity combines three different software modules namely Inchworm, Chrysalis, and Butterfly. Inchworm assembles the RNA-seq reads into unique sequences of transcripts also known as contigs. Chrysalis clusters contigs and constructs de Brujin graphs. Finally, Butterfly processes each graphs and generates full-length transcripts. Mate-ends and paired-ends from two data set were combined separately. The trinity job was run in Taito supercluster at CSC-IT Center for Science, Finland.

Transcriptome annotation

Blast2GO (B2G) [2] was used to annotate assembled reads resulting from genome/transcriptome based assembly as well as de novo assembly. The program makes blast search for query sequences and extracts gene ontology (GO) terms for each sequences. However, when we have large transcriptome data, it is useful to make local blast and import blast results into B2G. In addition, results from InterProScan can also be imported for motif analysis as it takes significant amount of time to query sequences against the database.

Contigs assembled by Trinity as well as transcripts from sheep-genome based assembly were blasted against local non-redundant (nr) database in Taito supercluster at CSC. Blast results were imported into Blast2GO program and were mapped against Gene Ontology (GO) terms and annotated against reliable functions.

Description of the main results obtained

A total of 277425 transcripts were assembled by Trinity with N50 value of 4568. The length of assembled transcripts range from 200nts to 41905nts. Due to the large number of assembled contigs, the transcripts were divided into smaller chunks based on the sequence length. All transcripts were blasted against local non-redundant (nr) protein database using blastx. The blast results were imported into Blast2Go for further annotation and analysis.

Mapping reads against sheep reference assembly resulted 32,503 transcripts, out of which 23,249 were annotated. A total of 1,047 transcripts did not find any hits in sequence database. Although majority of sequences were found in UniProtKB database, two other databases which consisted queried sequences were mouse genome database (MGI) and rat genome database (RGD). 21,468 transcripts also have functional classification in InterPro database.

Future collaboration with host institution

Researchers from both host and guest institution participated in a one-day meeting at host institution on the last day of my stay. Research projects from both parties were presented and discussed in detail. As there were many common research questions, both parties agreed to strengthen collaboration. I am planning to make another visit during Autumn 2013 for working with larger data set.

Projected publications/articles resulting from or to result from the grant

In addition to the results herein described, there are some tasks that need to be accomplished in next couple of months. A draft-paper has already been composed and we are planning to submit the work within next few months.

Other Comments

Prof. Goran Andersson, my host researcher, not only provided valuable suggestions regarding my project work, he

also introduced other fellow bioinformaticians and group leaders from various genomics projects which helped me in many ways. Although one month was not enough for analysing next generation sequencing data, I had a very fruitful visit in terms of expanding the knowledge I had before. The results from this visit will be added to the manuscript that was in preparation before leaving to host institute.

My sincere appreciation goes to Prof. Goran Andersson for providing such a good learning environment, fruitful suggestions and collaboration. I would also like to thank Prof. Eric-Boncgam Rudolf, Dr. Fernando Lopes Pinto, Shahid Manzoor, Prof. Anna Nasholm, Dr. Anna Maria Johansson and Sangeet Lamichhaney for all their help and discussions. Last but not least, I would like to acknowledge the Advances in Farm Animals Genomic resources Research Networking Programme and the European Science Foundation for providing an exciting opportunity.

Project N°5

Detection of signatures of selection in wild and domestic livestock populations

Filippo Buscarini, Parco Tecnologico Padano, Lodi, Italy

Introduction

With the Research Networking Programme "GENOMIC-RESOURCES", the European Science Foundation (ESF) promotes advances in research and innovation in farm animal genomic resources, and contributes to the education of scientists in cutting edge approaches to the characterization, economic evaluation, management, exploitation and conservation of Farm Animal Genetic Resources (FAnGR). By funding a number of Exchange Visits (ESF Exchange Grants), of a duration between 2 weeks and 3 months, "GENOMIC-RESOURCES" endeavours to accomplish its mission through the promotion of active collaboration and training between scientists from Universities and Research Centres in different European countries. Applying for the 2013 Call for ESF Exchange Grants, I had the opportunity to spend two weeks at the School of Biosciences of Cardiff University (Wales, United Kingdom). I carried out my research visit from March 11th to March 22nd 2013 in the conservation genetics research group of Prof. Michael W. Bruford.

Purpose of the visit

The purpose of the visit was to develop a research methodology for the analysis of whole genome sequences for the detection of signatures of selection in wild and domestic livestock populations. This work forms a component of the research activities being developed within the EU collaborative project "NEXTGEN" ([nex]). The increasing availability of genomic data has made it possible to genetically characterize different populations and to have insights into their phylogenetic relationships and evolutionary history (see, for instance, [Rubin et al., 2010]). With current technologies, high density panels of genetic markers and whole genome sequences are available, yielding vast numbers of polymorphisms and enabling a more comprehensive and accurate representation of the genetic architecture of individual species. This data should enable the detection of genetic variation and polymorphisms in all populations, and lead to the reduction of ascertainment bias (due to the use of heterologous reference sequences when mapping polymorphisms) in demographic and genetic inference. SNP chips typically contain 50 to 800 thousand single nucleotide polymorphic loci, whereas whole sequences contain up to some tens of millions of polymorphisms. In the context of the European FP7 project

"NEXTGEN", samples from wild populations of Mouflons *(Ovis orientalis)* and Bezoars *(Capra aegagrus)* have been collected from Northern Iran. Mouflons and Bezoars are considered the ancestors of domestic sheep and goats respectively, and Northern Iran is (one of) their putative domestication sites. Additionally, samples from domestic populations of sheep *(Ovis aries)* and goats *(Capra hircus)* have been collected in Iran and in 4 different climatic regions of Morocco (coast, northern atlas, southern atlas, desert). Within the NEXTGEN project, the demographic history and genetic characterization of the goat and its relatives and ancestors is being studied. The focus of my academic visit was to select and apply one method for the unambiguous detection of signatures of selection related to the domestication of sheep and goats, and their adaptation to different environments. This requires a comparative analysis of the genome of wild and domestic animals in order to identify loci affected by different selective pressures and that are likely to be involved in the processes of domestication and adaptation. A variety of methods can be used to detect genomic signatures of selection, which usually compare allele frequencies in different populations. Some such methods are the estimation of heterozygosity (e.g. [Rubin et al., 2010]), Fst (e.g. Kijas et al. [2012]) and CLLs (composite log-likelihoods, [Stella et al., 2010]) along the genome in a set of populations. We chose to focus on the estimation of EHH (Extended Haplotype Homozygosity, Sabeti et al. [2002]), which examines the variability along the genome within a population, by comparing the presence and extent of stretches of homozygous haplotypes to the overall homozygosity of the genome under analysis. The results obtained will be used to find loci involved in the domestication process (such as for example the TSHR gene found in chickens, [Rubin et al., 2010]), and to infer different selection pressures experienced by wild and domestic populations in relation with the adaptation to different environments. Directional selection leads to fixation of haplotypes or alleles in the different populations. However, balancing selection, which actively maintains diversity in the population, might also be relevant in the processes of adaptation to the environmental conditions and domestication (for instance, as might be the case with the high variability maintained at the MHC -major histocompatibility complex- loci).

Description of the work carried out during the visit

Material

The first step was to check what data were available and to obtain them. Data from a pilot study on 8 domestic sheep (O. aries) and 8 mouflons (O. orientalis) from northern Iran (see 1) were available for download from the FTP site of the NEXTGEN Project. These data had already been assembled by EMBL-EBI, and from the BAM files (reads aligned against the reference genome of the sheep v. 3.1 [she]) VCF (Variant Call Format) files had been produced. Data on domestic sheep and goats from Morocco, and from domestic goats and bezoars from Iran are still being produced, and only partially available. Therefore, we focused on the 16 samples from the Iranian pilot study to set up the methodology and workflow.

Methods

The estimation of iHS (integrated Haplotype Score, [Voight et al., 2006]), based on the EHH (Extended Haplotype Homozygousity) methodology ([Sabeti et al., 2002]), has been used for the detection of signatures of selection in domestic sheep and mouflons. The research group led by Prof. Michael W. Bruford in Cardi has extensive experience in evolution, demographics and genetics, and they have

expertise in the estimation of EHH and iHS. The estimation of EHH and iHS is suited for the detection of selective sweeps, i.e. regions of the genome that have experienced (or are still experiencing) a positive selective pressure in favour of a mutation that confers fitness advantage to its carriers. By looking at the length of the conserved homozygous haplotype around the mutation, recent and ancient selection events can be differentiated: a long haplotype is indicative of rapid fixation and therefore of recent selection. The iHS statistic compares the distribution of haplotype frequency around a locus between the ancestral and derived allele, relative to the genome as a whole, indicating whether and which one of the two alleles has been favoured by selection. By comparing allele information in wild and domestic sheep, it is possible to monitor the evolution of the frequency of the ancestral and derived alleles in the different populations, which will help shedding light on the domestication and evolutionary history of the sheep.

Under neutral evolution, new variants require a long time to reach high frequency in the population, and LD around the variants will decay substantially during this period owing to recombination. As a result, common alleles will typically be old and will have only short-range LD. Rare alleles may be either young or old and thus may have long- or short- range LD. The key characteristic of positive selection, however, is that it causes an unusually rapid rise in allele frequency, occurring over a short enough time that recombination does not substantially break down the haplotype on which the selected mutation occurs. A signature of positive natural selection is thus an allele having unusually long-range LD given its population frequency. The decay of LD, and therefore the relative scale of short and long range LD, is dependent on local recombination rates. A general test for selection on the basis of these principles must therefore control for local variation in recombination rates. The integrated Haplotype Score (iHS) is suited for this purpose.

Software

For the estimation of EHH and iHS the C++ programme iHS written at the University of Chicago (Pritchard Lab) was used. The R programming environment for statistical analysis, the high-level programming language Python and the Linux shell scripting languages have been used for the rest of the analyses.

Description of the main results obtained

Description of data

Whole-genome sequences from 8 Iranian sheep (O. aries) and 8 Iranian mouflons (O. orientalis) were aligned against the sheep reference genome v. 3.1 for detection of polymorphisms. Table 1 shows the total number of polymorphisms called in the two groups. In O. orientalis 3.5 million more polymorphisms were found than in O. aries.

	whole-genome sequences		
	alignment	sample size	n. of polymorphisms
O. orientalis	sheep genome	8	31,833,834
O. aries	sheep genome	8	28,306,478

Table 1: N. of samples and total n. of polymorphisms in the sheep and mou on populations.

Estimation of iHS

For the estimation of the iHS, the following steps were followed:

1.differentiate between ancestral and derived allele at each SNP locus. In Voight et al. [2006] this was done using the chimpanzee genome as an outgroup for the human genome. In this study, we were compelled to use O. orientalis as putative ancestral sequence, and defined the ancestral allele at each locus as the most frequent allele in the mouflon population.

Table 2 summarizes the rules followed to discriminate between the ancestral (coded as 0) and derived (coded as 1) alleles for di erent SNP genotypes in the O. orientalis and O. aries subpopulations. The frequency of the major allele was indicated as p, that of the minor allele as q;

2.estimate the decay of EHH (stretches of homozygosity) as a function of the distance from the core allele (either ancestral or derived). In plots of EHH, the area under the curve is expected to be much greater for alleles under rapid selection (due to a slower decay of homozygosity);

3.calculate the iHH (integral of the Haplotype Homozygosity, on each side of the core SNP): integral of the observed decay of homozygosity (iHHa, i HHd for ancestral and derived alleles respectively);

4.calculate the unstandardized integral Haplotype Score:

$$iHS = \ln\left(\frac{iHHa}{iHHd}\right) \quad (1)$$

When the rate of decay is similar for the ancestral and derived allele, the ratio iHHa/iHHd is≈ 1, and iHS (the natural logarithm of the ratio) approaches 0. Large negative values indicate unusually long haplotypes carrying the derived allele, and viceversa.

5.standardization of iHS. iHS is standatdised by subtracting its expected value and dividing by its standard deviation:

$$standardisediHS = \frac{\ln\left(\frac{iHHa}{iHHd}\right) - E\left[\ln\frac{iHHa}{iHHd}\right]}{E\left[\ln\frac{iHHa}{iHHd}\right]} \quad (2)$$

The standardised iHS measures how unusual the haplotypes around a given SNP are, relative to the genome as a whole.

6.the iHS statistics calculated at each SNP locus are then averaged over a sliding window of SNPs spanning x kbps (100 kbps in Voight et al. [2006]). The reason is that selective sweeps tend to produce cluster of extreme iHS values, while under a neutral model extreme his values are randomly scattered along the genome. This helps avoiding spurious signals of selection.

0 = ancestral allele 1 = derived allele					
O. orientalis	AA	AT	AA	AT	AT
	A=0 q=1	p=0	A=0	A=T=0	A=T=0
		p=q=0.5 discard!			
O. aries	AA	AT	AT	AC	GC
	A=0	p=0 q=1	A=0 T=1	A=0 C=1	A=T=0 G=C=1

Table 2: Rules to determine the ancestral and derived alleles

Initial results

For the initial analysis, aimed at setting up the work flow of analysis and the informatics pipeline for the computations, data from only one chromosome were selected. Chromosome 26 was chosen, which is one of the smallest chromosomes of the sheep genome, thereby reducing computation issues. Table 3 summarizes the polymorphisms found on chromosome 26. The vast majority of the polymorphisms present on the chromosome is represented by single nucleotide polymorphisms (SNPs); there are, however, about 9% of other polymorphisms (INDELS -Insertion-Deletions-,CNVs -copy number variations-, and other structural variations), both in O. aries and O. orientalis. A low level of homozygosity (compared to that normally found in commercial sheep populations) was found in the analysed samples: 6.65% in O. orientalis and 4.65% in O. aries. This might be due to the fact that no sheep sample from Iran was used to build the sheep reference genome and the result may be due to ascertainment bias. More polymorphisms have been found in the O. orientalis genome than in that of O. aries, which is consistent with the strong directional selection pressure that domestic sheep are likely to have undergone.

chromosome 26				
	polymorphisms	SNPs	INDELs co. (%).	homozygous SNPs (%)
O. orientalis	590,386	536,442	53,944 (9.14%)	35,656 (%6.65)
O. aries	526,166	476,757	49,409 (9.39%)	21,722 (%4.56)
diff	-10.88%	-11.13%	-8.41%	-39.08%

Table 3:

The number of SNPs common to the O. aries and O. orientalis samples was 335,273: for 17,440 of such SNPs the frequency of both alleles was approximately the same (p q 0:5) and it was not possible to determine which allele was ancestral. The remaining 317,833 SNPs were used for iHS estimation (see Table 4)

	O. orientalis	O. aries	Common SNPs	p=q (ancestor)	left for iHS
n. SNPs	536,442	476,757	335,273	17,440	317,833

Table 4: Available SNPs for iHS estimation on chromosome 26

With such large numbers of SNPs, computation time certainly proved to be an issue. Preliminary results show that, on a MacBook Pro with and Intel Core(TM)2 Duo CPU at 2.53GHz, 48 hours were needed to complete the analysis for chromosome 26 of O. orientalis. However, computation speed can be considerably increased by running the analysis on a server or a distributed CPU cluster, and by using an optimised version of the ihs C++ programme.

Experimental design

Whole-genome comparison between signatures of selection detected -through the estimation of iHS- in O. aries and O. orientalis individuals sampled from the domestication centre in northern Iran (Figure 1), will give us first indications on the domestication process of the sheep. However, although they come from the same geographic region, they live in di erent environments: Iranian mouflons live in the wild, while Iranian domestic sheep are farmed and kept in a more controlled environment. Thus, domestication signals may be confounded by signals of adaptation to the two different environments. This ambiguity may potentially be overcome using also the samples of O. aries from 4 di erent geographic regions in Morocco (coast, northern Atlas, southern Atlas and desert: see Figure 2).

By making multiple pair-wise comparisons between domestic sheep from different environments and the Iranian mouflons, we may be able to distinguish between signals of domestication and adaptation. The di erential signatures of selection appearing in all comparisons may be considered true signals of domestication, while those specific of domestic sheep from any given region can be interpreted as signals of adaptation to the different environments. The Cochran Mantel Haenszel Figure 1: Area of northern Iran from where the sheep and mouflons for the pilot experiment were sampled

Figure 2: Different climatic regions of Morocco (coast, northern Atlas, southern Atlas, desert) from where domestic sheep and goats were sampled

(CMH, see OROZCO-terWENGEL et al. [2012] for an illustration) test will be used to simultaneously assess the significance of signatures of selection from multiple comparisons: this is analogous to the Fisher test for pairwise comparisons (contingency tables) but extended to the case of multiple comparisons. The CMH test is used to test multiple 2x2xk contingency tables for independence of marginal sums across k replicates.

Software developed

While setting up the workflow for the analysis, some relevant software has been developed. A shell script was written to efficiently handle and divide large vcf files (≈ 2GB) containing all polymorphisms in the genome of O. aries and O. orientalis. A python programme was then written to determine the allelic state (ancestral or derived) and prepare the input files for the iHS estimation. The ihs C++ programme for the estimation of iHS was modified and optimised in order to make it more efficient and faster: preliminary results promise to reduce computation time by as much as 40%.

An overview of the project, the work done so far and the preliminary results obtained, was illustrated to staff and students in the Bruford group during an informal lab meeting seminar held at the School of Bioscience of Cardiff University on Friday March 22nd.

Future collaboration with host institution

The short placement helped strengthening the relationship with the host institution, fostering current collaborations and opening up possibilities for future collaborations. First, the work on the domestication of sheep started during the placement is being continued. The workflow for the estimation of iHS is being finalised and will then be applied to the entire dataset from the Iranian pilot study (8 sheep, 8 mouflons, the 52 autosomes of the sheep genome, sex chromosome excluded -at least initially). The method will be then applied to whole-genome sequences coming from sheep sampled in Morocco, to complete the study on domestication and adaptation of O. aries. When data on domestic goats and their wild counterparts, the bezoars, are available, the same line of work will be applied also to this species, with the aim of shedding light on the domestication process of the goat. Additionally, the bioinformatic pipeline and software thus developed, will be added to a user-friendly web-interface for the detection of signatures of selection recently developed at PTP ([Biscarini et al., 2012]), which currently implements the method of CLL.

Projected publications / articles resulting or to result from the grant

The work initiated during my stay at Cardiff University is going to produce the following projected publications:

1 article on the detection of signatures of selection for domestication and adaptation to different environments in sheep;

1 article on the detection of signatures of selection for domestication in goats;

1 article on the development of software and user-friendly web.interface for the detection of signatures of selection in animal and plant populations;

The European Science Foundation (ESF) will be duly acknowledged in all publications resulting from the grant.

Project N°6

Estimation of effective population size from linkage disequilibrium in a closed herd of ancient Iberian pigs.

Maria Saura, INIA (Instituto Nacional de

Investigación y Tecnología Agraria y Alimentaria, Madrid, Spain

Purpose of the visit

The effective population size (Ne) is a crucial parameter in conservation genetics, as it is related with the genetic variability and gives an idea of the genetic status of the populations. Traditionally, Ne has been estimated from the rate of inbreeding or the rate of coancestry using pedigree information (Toro et al. 2000). With the increasing availability of high-density genome-scan data, interest has grown in exploring whether more reliable and accurate estimate of genetic variability parameters might be derived on the basis of genomic marker data. This molecular information allows us to obtain both current and ancestral estimates of Ne from linkage disequilibrium (LD), providing valuable knowledge on the ancient history of the populations (Hayes et al 2003, Tenesa et al 2007). LD over long distances reflects recent Ne, whereas LD over small distances reflects the Ne in the more distant past (Hayes et al. 2003). The pattern of historical Ne can also help us to understand the impact of selection on the genetic variation present in populations and can lead us to an in-depth understanding of inbreeding processes (Corbin et al 2010).

One of the varieties of Iberian pigs that is now in serious danger of extinction is the Guadyerbas strain. It is one of the most ancient surviving Iberian strains and currently is the only representative of the black hairless genetic type. The strain has been conserved in an experimental herd as a genetically isolated population since 1944 (Toro et al 2000). Accurate genealogical and performance information for reproductive traits has been recorded since the foundation of the herd and this makes it a genetic resource of exceptional value.

The aim of this visit was to apply the methodology developed and implemented by Albert Tenesa (Tenesa et al 2003, 2007) and John Woolliams (Corbin et al 2010, 2012) regarding the estimation of Ne from LD patterns to our data of Iberian pig. We have available genotypes obtained from the Illumina Porcine SNP60 Genotyping BeadChip for 220 individuals.

Description of the work carried out during the visit and main results

The work I have performed during the visit can be summarized in two main epigraphs:

1. Estimation of linkage disequilibrium in the population of Guadyerbas animals
2. Estimation of ancestral effective population size from LD measures in this population

1. Estimation of LD

The measurement of LD used in this study is the squared correlation coefficient between SNP pairs (r^2) (Hill and Robertson 1968), computed as:

$$r^2 = \frac{p_{AB} - p_A p_B}{p_A p_a p_B p_b}$$

LD (r^2) was calculated for all syntenic marker pairs within distance bins up to 50Mb. LD declined with increasing distance between SNP pairs, with high LD values at distances lower of 1Mb (r^2= 0.6). Figure 1 shows that the most rapid decline was seen over the first 0.9Mb, with the mean r^2 decreasing by more than half over this period.

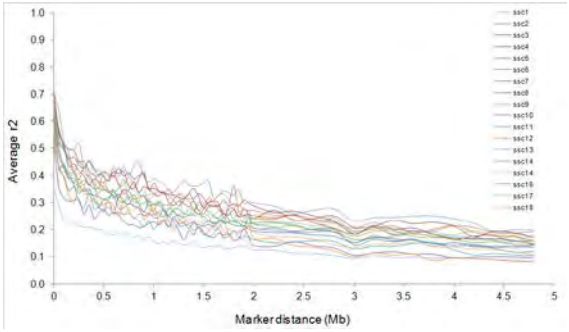


Figure 1. Linkage disequilibrium across autosomes as a function of the genomic distance. Average r^2 is represented according physical distance into bin sizes of 0.05Mb bins from 0 to 2Mb and 0.2Mb bins from 2 to 5Mb.

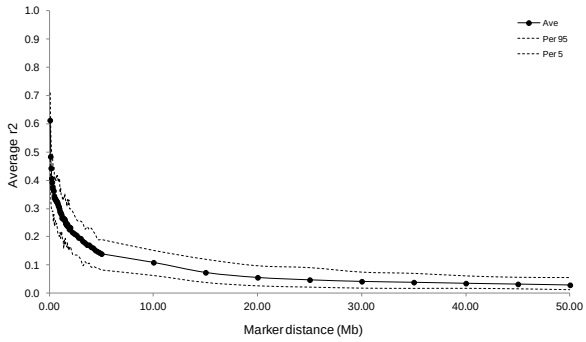


Figure 2. Linkage disequilibrium across the genome as a function of the genomic distance. Average r^2 is represented according physical distance into bin sizes of 0.05Mb bins from 0 to 2Mb; 0.2Mb bins from 2 to 5Mb; and 5Mb bins from 5 to 50Mb. Percentiles 5 and 95 are also represented.

In order to determine the extension of LD, we calculated r^2 between randomly sampled markers (30 SNPs per chromosome). This value was 0.064±0.024, which was similar to that observed for syntenic markers at distances greater than 15Mb (Figure 2). Both the magnitude as well as the extension of LD was much higher than for other porcine breeds (Nsengimana et al. 2004, Harmegnies et al. 2006).

From all the pairwise r^2 comparisons within chromosome, the accumulated proportion of pairwise $r^2 = 1$ achieved 13% for the whole genome. Average LD for different bin distances per chromosome are summarized in Table 1.

Chrom	0.5Mb	1Mb	5Mb	10Mb	50Mb
ssc1	0.24	0.21	0.13	0.10	0.06
ssc2	0.42	0.37	0.24	0.18	0.09
ssc3	0.46	0.41	0.28	0.22	0.11
ssc4	0.49	0.44	0.28	0.20	0.07
ssc5	0.40	0.36	0.22	0.16	0.07
ssc6	0.48	0.42	0.28	0.21	0.08
ssc7	0.38	0.33	0.21	0.16	0.07
ssc8	0.50	0.45	0.28	0.21	0.11
ssc9	0.41	0.37	0.24	0.18	0.08
ssc10	0.32	0.28	0.17	0.12	0.05
ssc11	0.37	0.31	0.19	0.14	0.06
ssc12	0.38	0.33	0.18	0.13	0.06
ssc13	0.52	0.47	0.31	0.25	0.16
ssc14	0.47	0.42	0.28	0.21	0.10
ssc15	0.44	0.40	0.26	0.20	0.10
ssc16	0.42	0.36	0.22	0.18	0.08
ssc17	0.47	0.41	0.26	0.20	0.09
ssc18	0.39	0.34	0.20	0.14	0.06
Average	0.42	0.37	0.24	0.18	0.08
SD	0.0688	0.0648	0.0479	0.0397	0.0255

Table 1. Average r^2 for different distances bins of 0.5, 1, 5Mb and 50 Mb for each chromosome.

In order to explore the LD within chromosome, we identified LD blocks with a r^2 value higher than 0.9 (Figure 3). The largest blocks were observed in the chromosomes 9 (122 SNPs), 13 (two blocks of 299 and 376 SNPs), 14 (428 SNPs) and 17 (256 SNPs). One of the blocks detected in chromosome 13 contains a QTL related to prolificacy (Noguera et al 2009), which could be related with the decrease in the number of individuals across generations (data not shown).

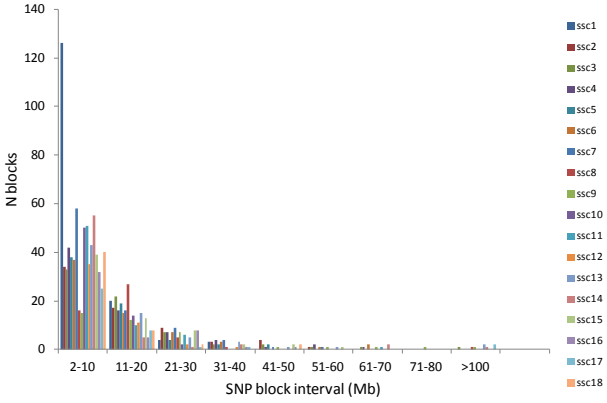


Figure 3. Haplotype blocks arranged according to the number of SNPs within a block for each chromosome. Notice that chromosomes 8, 9, 13, 14, 17 show haplotype blocks including more than 100 tag SNPs (notice that a tag SNP refers to the most representative SNP within a block for the r^2 defined).

Figure 4. Mean LD estimates at different physical distances pooled over all the autosomes and estimated at three minimum threshold levels of cutoff for MAF.

2. Estimation of effective population size

LD is function of Ne and recombination rate (c) and is usually measured as the correlation of the frequency of alleles at two chromosomal loci as decribed by Sved 1971,

$$E[r^2] = \frac{1}{4N_e c}$$

Including mutation (α) and finite experimental sample size (Hill 1981), the relationship between LD and Ne is summarized in the expression,

$$E[r^2] = \frac{1}{\propto + kN_e c} + \frac{1}{n}$$

Exploring these equations Hill predicted that LD between marker located at close distances reflects the ancient history of the population, while LD at larger distances describes more recent events.

Chromosome-specific Ne was estimated for all autosomes from estimated LD (r^2) between pairs <100 Kb apart (discarding SNPs at <5 Kb in order to avoid the influence of gene conversion), following Tenesa et al (2007).

We used the recombination rates (*c*) from the study of Tortereau et al 2012 (in order to avoid dependence between LD and *c*). We adjusted the estimates of Morgan/Mb provided by the authors to the corresponding distances between marker pairs of 100 Kb and then estimated Ne separate for each chromosome by fitting the nonlinear regression model:

$$y_i = \frac{1}{\propto + \beta c_i} + e_i$$

where $y_i = (r^2 - 1/n)$ and parameters α and β were estimated iteratively using least squares. Effective population size was thus estimated from the regression coefficient.

The main advantage of this method is the possibility of investigating the change on Ne over time, as LD between loci at a determined recombination rate reflects the ancestral Ne 1/2c generations ago (Hayes et al 2003, assuming linear growth).

We thus explored the number of generations in the past that the method allowed according to our data. Using small distances (<100 Kb), average estimates per chromosome reflected a value of Ne of 291 individuals for 1426 generations in the past (Table 3). Pig domestication occurred 10000 years ago, with the expected consequence of the reduction of Ne in the domestic populations. The next step will be to explore LD at higher distances with the aim of estimating Ne from the foundation of the herd (1944) to the present. In addition, the availability of the complete pedigree of the herd will allow us to compare different estimates of Ne.

Future collaborations

During the stay new ideas about the project have emerged, including to perform estimates of *Ne* from genealogical data and coalescence analysis. Concerning the former approach, the fact of having the complete genealogy of the herd will allow us to test the validity of coalescence analysis applied to livestock breeding.

In addition, during my stay at The Roslin Institute, I got involved in a project from Albert Tenesa, about complex traits in human populations analyzed with the Human SNP chip. I applied the same methodology they are implementing in their human population to our porcine data. This new project will entail a new collaboration with Albert´s group that is going on at the present.

Project publications resulting from the grant

We expect to publish two different papers corresponding to each one of the projects that will be also presented in congress during next year.

Project N°7

Impact of sample size, marker density and population structure on the estimation of effective population size from the SNP chip information: a simulation approach

Dr. Stamatina Trivizaki, from the Institute of Animal Genetic Improvement of Nea Mesimvria, Thessaloniki (Greece).

Purpose of the visit

The conservation of genetic diversity is a priority factor in rare small populations (Meuwissen, 2001). The effective size (Ne) is one of the most important issues in population genetics, given its usefulness as a measure of the long-term performance of the population regarding both diversity and inbreeding and, therefore, to characterize the risk status of livestock breeds (Duchev et al., 2006; FAO, 1998). Recently, the average of effective size across populations has been proposed to assess the genetic diversity in livestock (Villanueva et al., 2010). Ne helps to explain the observed extent and pattern of genetic variation in a population from a retrospective point of view and to predict the loss of genetic variation and the survival of small breeding populations from a prospective point of view (Wang 2005). Ne is also necessary in the application of genomic selection of populations, since the accuracy of the breeding values depends on the linkage disequilibrium between the QTL and SNP, and a higher Ne means higher number of markers is needed (Meuwissen et al., 2001). However, estimates of Ne vary with the methodology used to asses it, thus limiting its ability to characterize the risk status of a population or to predict the precision of the genomic selection. Therefore, assessing the reliability of the estimates of Ne is a very important challenge.

The basic methods used to estimate Ne can be divided into demographic, pedigree-based or marker-based approaches. Though, demographic methods provide crucial Ne estimates in some situations, they are based on simplified population models and use limited population data. In pedigree-based methods, inbreeding rate is estimated from pedigree records, which in turn is used to estimate Ne. However, these methods require reliable and complete pedigree data over several generations, which is often lacking even for breeds in developed countries (Konig and Simianer, 2006). Since 2007, assays to generate dense genome-wide single nucleotide polymorphism (SNP) data became available for cattle. These developments offered new possibilities regarding the application of marker-based methods for estimating Ne with linkage disequilibrium (LD) information (Flury et al., 2010, Qanbari et al., 2010).

Sved (1971) and Hill (1981) proposed that LD would come exclusively from genetic drift for neutral unlinked loci in an isolated population with random mating. This phenomenon could be used to estimate Ne by exploiting that the variance of LD or the correlation of gene frequencies r between loci is a known function of the population size (Hill, 1981). Based on simulations and real data, the same relation between Ne, segment length c and expected LD was confirmed for a multi locus measure of LD called chromosome segment homozygosity (CSH) (Hayes et al., 2003). Generally, the Ne is estimated (2c)-1generation ago where c is either distance (in Morgans) between two markers for which LD is estimated (Hill, 1981) or the chromosomal segment length for CSH (Hayes et al. 2003).

With many breeding programs now incorporating genomic information at great expense, simulation provides a useful tool for providing information about the potential that different analysis methods have to increase the accuracy of estimating breeding values and to compare the alternative structures of breeding programs,

at low cost while overcoming difficulties that origin from data accessibility and availability. The application of simulation for LD estimation of Ne and the effect of a series of population characteristics and inputs to the estimation accuracy is a preliminary approach that would provide information about the best methodology to be further applied in real population data. The aim of this short visit is multi-dimensional as it targets not only in acquiring useful results according to the proposal but also to create a solid research background that could be further developed through future collaboration and invest in the expansion of the scientific horizon of both the applicant and the host group. The first step to this direction is to investigate the possibilities and capacities of bioinformatics tools, mathematic equations and genomic background information in estimating Ne from LD. The genomic dimension of LD approach of Ne is quite innovative and still unexplored in many aspects. Several steps have been made to the direction of estimating Ne in historical populations, while several formulas have been proposed for this reason from Sved (1971), Weir and Hill (1994), Waples (2006) and Ober et al. (2013)(Table 2). In this short visit research plan the main focus was different than the approach attempted by previous scientific groups on the estimation of Ne. In our case study, the interest is focusing on recent historical population, examining Ne estimation in the previous 100 generations (recent past).To this extend the effect of 1) sample size and (2) marker density on LD estimation of Ne were examined.

Description of the work carried out during the visit

Simulated Data

The simulation method, implemented in a software package called AlphaDrop (Hickey and Gorjanc, 2012). This package is used to simulate genomic data and phenotypes with flexibility in terms of the historical population structure, recent pedigree structure, distribution of quantitative trait loci effects, and with sequence and single nucleotide polymorphism-phased alleles and genotypes. The system is calling a combination of coalescent and gene drop methods to simulate sequence, SNP, and QTL and it is packaged in a Fortran 95 program called AlphaDrop, which calls the Markovian Coalescence Simulator (MaCS) (Chen et al. 2009). In order to perform simulation according to the project requirements, collaboration with Assistant Professor Gregor Gorjanc (University of Ljubljana) was established. Dr. Gorjanc demonstrated the capacities of the program in terms of the parameters included and modifications availability due to the needs of the project.

Three different bovine populations simulated having the same characteristics but scanned for BeadChips of different size. Each individual in the population is represented by 29 chromosomes. AlphaDrop is setting up data structures in terms of SNP chips and pedigree. It then calls MaCS, which simulates a sample of haplotypes with sequence information for each chromosome according to the specified ancestral population, mutation and recombination rates that in this case were both 10-8 (approximately one mutation and one recombination event per Morgan). Pedigree was internally created in AlphaDrop with each generation consisting of 50 sires, one dam per sire and two offsprings per dam, holding the population size per generation constant to 100 individuals. The population simulated for effective population size of 200.

The BeadChips used are: (a) BovineSNP50 v2 DNA Analysis BeadChip that contains 54.609 highly informative SNPs uniformly distributed across the entire genome of major cattle breed types, (b) BovineHDBeadChip with more than

777.000 SNPs that uniformly span the entire bovine genome and (c) IlluminaBovineLD Genotyping BeadChip with 6.909 SNP across bovine genome. Considering the chromosomal length at 1 Morgan the distribution of SNP by each BeadChip is presented in the following table (1).

Table 1. SNP distribution in total bovine genome by different BeadChips provided.

BeadChip	Number of SNP in the entire bovine genome	Number of SNP per chromosome	
		approximately	used in analysis
IlluminaBovineLD Genotyping BeadChip	6.909	238	200
BovineSNP50 v2 DNA Analysis BeadChip	54.609	1.883	2.000
BovineHDBeadChip	777.000	26.793	20.000

The full sequence and phased data were additionally programmed to be outputted.

Haplotype data were simulated in AlphaDrop for 100 generations ago.

The size of the output files is 2,1G for BovineLD, 24G for BovineSNP50 and 78G for BovineHD. In order to shrink output file size and adjust to available computational capacities, data was retracted using Linux commands for the most remote generation 90 to 100 generations ago to be further analysed.

Pedigree estimation of Ne

The software package, GRain v2.1 (Baumung et al.,) have been demonstrated and used in the assistance of Professor Curik. This software intended to enable and promote testing of various hypotheses with respect to purging and heterogeneity of inbreeding depression.

The program withdraw information from the pedigree output file created internally in AlphaDrop and estimate the inbreeding coefficients that were applied to the estimation of inbreeding Ne per generation according to (1).

(1)

Where ΔF = the increment in inbreeding per generation (the rate of inbreeding)

Analysis of LD data and estimation of Ne

An R script has been accommodated in collaboration to Dr. Gorjanc for the estimation of both LD and Ne among different BeadChip sizes either from haplotypes.

In order to approach this estimation all correlations between SNPs involved were estimated. According to Ne theory, physical SNP positioning in the genome could be translated to genomic distance according the formula (2).

(2)

Where Gen= generations ago, c= genetic distance in Morgan (Hill, 1981).

The position of SNP in the genome is provided by AlphaDrop output positioning file. As the main interest focuses on the recent past, we targeted on physical distances between 500 and 40.000 Kbp which corresponds to 1,25 generation up to 100 generation ago. The physical distance has been cut into bins of size 500 Kbp. All the LD values estimated from simulated data were assigned to the corresponding bin and the average, median and CV of LD within each bin was estimated. In order to approach Ne three different mathematic formulas were tested (Table 2). Physical distances are

transcribed into genetic distances and into their relevant count of generations.

Table 2. Mathematic equations for the LD estimate of Ne.

Literature reference	Equation
1. Sved (1971)	(3)
2. Weir and Hill (1994)	(4)
3. Ober et al. (2013)	(5)
$E(r^2)$= expectation correlation between SNPs, Ne= effective population size, c= genetic distance (in Morgans), S= census population sample size	

While the first three formulas are approaching Ne using haplotype data, the last derivation is approaching Ne from genotypic data. The precursor formula presented by Sved although it had been revised and updated it is still widely used in Ne research.

Specific routines and loops were added to the R script in order to estimate Ne in different simulated single generation or in sequential series of generations of AlphaDrop. Moreover, R script was modified to estimate Ne retrieving information from the total available population or a random sample of the population per generation. Additionally a routine for excluding Minor Allele Frequencies (MAF) SNPs modified R script according to literature suggestions. SNPs with MAF smaller than 0,01 were excluded from the calculations.

After setting different options, three simulated sets of haplotype data created by scanning with different density markers and analysed by taking into consideration the exclusion of MAF, total population (100 individuals) or subsets of 50 and 25 individuals and estimates in generations 90 to 100 of AlphaDrop. Ne was approached by all three formulas referred above.

Bioinformatic tools and additional knowledge acquired

During the short visit in the University of Zagreb the applicant had the chance to acquire basic knowledge in the use of Linux operating system and commands. The use of Linux is approached in PC terminal but also in University server. Additional computational skill has been practiced by the introduction to R statistical computing and furthermore the use of more sophisticated routines that required for data processing. Dr. Trivizaki explored the capacities of MaCS and AlphaDrop simulation programs and practiced estimations using VanRad and GRain programs for inbreeding coefficients.

Moreover, the applicant had the opportunity to attend a short course in artificial neutral network theory lectured by Dr. Hayrettin Okut, an introduction course to Bayesian statistics by Dr. Gregor Gorjanc and a short term course in Conservation Genetics by Dr. Ino Curik.

Description of the main results obtained

Simulated Data analysis

R script application and estimation of Ne in simulated data created by HD BeadChips couldn't be accessed because of lack in computer capacity to allocate 20000 SNP matrix in data reading.

Pedigree estimation of inbreeding Ne

For the simulated data population inbreeding Ne approximately 200 was estimated by the application of GRain.

LD estimation of Ne

Considering Ober formula for estimating Ne from LD, that is the most recent proposed for estimation in remote past generations, it is clear that Ne is overestimated in the recent 200 past generations. According to this formula Ne is "indicating" values of 300 with increasing variation around the mean value for each bin as the sample size taken into account is getting smaller. Though when applying the bigger chip (2000 SNP/chr) instead of 200 SNP/chr, Ne values are shrinking around the mean in each bin, demonstrating less dispersion.

In the 100 more recent past generations Sved and Weir-Hill formulas show better fitting to the expected value of Ne 200. The use of total population information gives higher precision to the Ne values, while use of smaller divisions of the population underestimate Ne values for Sved and slight overestimate for Weir-Hill. These results are more visible using denser molecular marker that contributes to the elimination of Ne values around the mean. Actually, Weir-Hill formula seems to have the most accurate approach of Ne value estimated from inbreeding (pedigree) for this time frame.

(Table 3)

All formulas demonstrate important amount of outliers in recent generations in the past when genome is scanned with 200 SNP/chr. creating a lot of noise and make it difficult to identify any existing trend of Ne.

Another point of interest that is evident in processing these data sets is the influence of bin size selected. It is obvious that while addressing to more recent generations in the past the bin values of reference are more in number and are gradually decreasing when moving to most remote generations. This suggests that more recent generation LD values and the relevant Ne estimated from them, will be more detailed analysed. For example, 30 bins equal 30 intervals in genomic distances that refers to Ne values when approaching the first two generations in the past while only few bins are capturing Ne information for generations 28 to 200.

Future collaboration with host institution

The visit to University of Zagreb, Department of Animal Science had been a great opportunity to establish collaboration with Professor Curik and the members of his scientific team. The present project is very promising to its contribution to genetic conservation perspectives though it is estimated to be at its primitive stages. The three month collaboration is considered very fruitful as it established the foundation to future collaboration while both sides had agreed in expanding their research in a common direction that would contribute to scientific achievements. Moreover, a working group in recent past estimation of effective population size from molecular data is being formed with the contribution of Assistant Professor of the University of Ljubljana, Dr. Gorjanc.

Projected publications / articles resulting or to result from the grant (ESF must be acknowledged in publications resulting from the grantee's work in relation with the grant);

Based on the feedback of the current study and future research results an article on the topic will be submitted to a high quality peer reviewed scientific journal. Acknowledgement on the funding institution will be expressed in the coming manuscripts and adhere to other principles as required by the ESF GENOMIC-RESOURCES programs.

Other comments

The aim of this study as it was primarily proposed for ESF grant was an opportunity to assess the effect of several factors (marker density, sample size and population structure) on genomic data (LD) estimation of Ne. While starting accommodating this features the applicant, Dr. Trivizaki and host Professor, Dr. Curik realized the importance of exploring the area of estimating recent past Ne. Observations in this field as it is presented in section 3 revealed issues in question that hadn't been taken into consideration initially like evaluating behaviour of mathematic formulas in recent past Ne and computational restrictions. Taking into account possibilities of further exploring the area of Ne estimated by SNP data information, the scientific team is very much concentrated in delivering high quality results, especially since the current attempt is funded by an organization as reputed as ESF. The scientific team is already working on the direction proposed to ESF taking into consideration the key points attributed by this short visit.

On behalf of the applicant, I would like to express my gratitude to ESF for supporting the current study. This opportunity contributed in my involvement to the field of conservation genetics but also to the establishment of scientific network. I would like to thank Dr. Gorjanc for his assistance and teaching guidance in several bioinformatics tools. Furthermore I would

also like to thank the scientific team working in the Department of Animal Science at University of Zagreb for their assistance in minor and major importance issues not only in the science but also to everyday life.

Last but not least, I would like to thank Professor Curik for giving me the opportunity to participate in his team and get involved in such a crucial topic as Ne. His guidance and support cannot be appreciated more. I hope that our future collaboration will bring me in a position to compensate for the knowledge, the assistance, the opportunities and the personal contribution.

Recommended Bibliography

1. FAO ANIMAL PRODUCTION AND HEALTH - GUIDELINES 14
   In vivo conservation of animal genetic resources - http://www.fao.org/docrep/018/i3327e/i3327e00.htm

2. Emily Jane McTavisha,, Jared E. Deckerb, Robert D. Schnabelb, Jeremy F. Taylorb, and David M. Hillisa - New World cattle show ancestry from multiple independent domestication events - PNAS | Published online March 25, 2013 | E1405

3. Yuri Tani Utsunomiya, Ana Maria Pe´ rez O'Brien, Tad Stewart Sonstegard, Curtis Paul Van Tassell, Adriana Santana do Carmo, Ga´bor Me´sza´ros, Johann So¨ lkner, Jose´ Fernando Garcial - Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods - PLOS ONE | www.plosone.org - May 2013 | Volume 8 | Issue 5 | e64280

4. Jessica L. Petersen, James R. Mickelson, E. Gus Cothran, Lisa S. Andersson, Jeanette Axelsson, Ernie Bailey, Danika Bannasch, Matthew M. Binns, Alexandre S. Borges, Pieter Brama, Artur da Caˆmara Machado, Ottmar Distl, Michela Felicetti, Laura Fox-Clipsham, Kathryn T. Graves, Ge´ rard Gue´ rin, Bianca Haase, Telhisa Hasegawa, Karin Hemmann, Emmeline W. Hill, Tosso Leeb, Gabriella Lindgren, Hannes Lohi, Maria Susana Lopes, Beatrice A. McGivney, Sofia Mikko, Nicholas Orr, M. Cecilia T Penedo, Richard J. Piercy, Marja Raekallio, Stefan Rieder, Knut H. Røed, Maurizio Silvestrelli, June Swinburne, Teruaki Tozaki, Mark Vaudin, Claire M. Wade, Molly E. McCue - Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data - PLOS ONE | www.plosone.org - January 2013 | Volume 8 | Issue 1 |

5. O. Thalmann et al.- Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of Domestic Dogs - DOI: 10.1126/science.1243650 - Science 342, 871 (2013) - European Origin of Domestic Dogs

6. M. Simcic, J.A. Lenstra, R. Baumung, P. Dovc, M. Cepon & D. Kompan - On the origin of the Slovenian Cika cattle - J. Anim. Breed. Genet. ISSN 0931-2668

7. Maja Ferenčaković, Johann Sölkner and Ino Curik - Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors - Ferenčaković et al. Genetics Selection Evolution 2013, 45:42 - http://www.gsejournal.org/content/45/1/42

8. T.C. Bray, S.J.G. Hall, M.W. Bruford - Admixture analysis in relation to pedigree studies of introgression in a minority British cattle breed: the Lincoln Red - Article first published online: 5 AUG 2013 – Journal of Animal Breeding and Genetics - DOI: 10.1111/jbg.12047

9. Sean Hobana, Jan W. Arntzenb, Giorgio Bertorellea, Josef Bryjac, Margarida Fernandesd, Katie Frithe, Oscar Gaggiottif, Peter Galbuserag, José A. Godoyh, Heidi C. Hauffei, A. Rus Hoelzele, Richard A. Nicholsj, Sílvia Pérez-Esponaj, Craig Primmerk, Isa-Rita M. Russol, Gernot Segelbacherm, Hans R. Siegismundn, Marjatta Sihvonenk, Per Sjögren-Gulveo, Cristiano Vernesii, Carles Vilàh, Michael W. Brufordl, - Conservation Genetic Resources for Effective Species Survival (ConGRESS): Bridging the divide between conservation research and practice - Journal for Nature Conservation -Volume 21, Issue 6, December 2013, Pages 433–437

10. Sean M. Hoban, Heidi C. Hauffe, Sílvia Pérez-Espona, Jan W. Arntzen, Giorgio Bertorelle, Josef Bryja, Katie Frith, Oscar E. Gaggiotti, Peter Galbusera, José A. Godoy, A. Rus Hoelzel, Richard A. Nichols, Craig R. Primmer, Isa-Rita Russo, Gernot Segelbacher, Hans R. Siegismund, Marjatta Sihvonen, Cristiano Vernesi, Carles Vilà, Michael W. Bruford - Bringing genetic diversity to the forefront of conservation policy and management - Conservation Genetics Resources -June 2013, Volume 5, Issue 2, pp 593-598

11. H. M. Pereira, S. Ferrier, M. Walters, G. N. Geller, R. H. G. Jongman, R. J. Scholes, M. W. Bruford, N. Brummitt, S. H. M. Butchart, A. C. Cardoso, N. C. Coops, E. Dulloo, D. P. Faith, J. Freyhof, R. D. Gregory, C. Heip, R. Höft, G. Hurtt, W. Jetz, D. S. Karp, M. A. McGeoch, D. Obura, Y. Onoda, N. Pettorelli, B. Reyers, R. Sayre, J. P. W. Scharlemann, S. N. Stuart, E. Turak, M. Walpole, M. Wegmann - Essential Biodiversity Variables - Science 18 January 2013: Vol. 339 no. 6117 pp. 277-278 - DOI: 10.1126/science.1229931

Funding

ESF Research Networking Programmes are principally funded by the Foundation's Member Organisations on an *à la carte* basis. **GENOMIC-RESOURCES** is supported by:

- Fonds zur Förderung der wissenschaftlichen Forschung (FWF), FWF Austrian Science Fund, Austria
- Fonds National de la Recherche Scientifique (FNRS), Belgium
- Fonds voor Wetenschappelijk Onderzoek - Vlaanderen (FWO), The Research Foundation - Flanders, Belgium
- Nacionalna zaklada za znanost, visoko školstvo i tehnologijski razvoj Republike Hrvatske, National Science Foundation for Science, Higher Education and Technological Development, Republic of Croatia
- Suomen Akatemia, Biotieteiden ja ympäristön tutkimuksen toimikunta, Academy of Finland, Research Council for Biosciences and Environment, Finland
- Deutsche Forschungsgemeinschaft (DFG),German Research Foundation, Germany
- Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), The Netherlands Organisation for Scientific Research,The Netherlands
- Norges Forskningsråd,The Research Council of Norway, Norway
- Forskningsrådet för miljö, areella näringar och samhällsbyggande, Swedish Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS),Sweden
- Schweizerischer Nationalfonds (SNF), Swiss National Science Foundation, Switzerland
- Biotechnology and Biological Sciences Research Council (BBSRC), United Kingdom
- Institut National de la Recherche Agronomique (INRA) - France

GENOMIC-RESOURCES Steering Committee

Dr Stéphane Joost (Chair) Ecole Polytechnique Fédérale de Lausanne, Switzerland – Dr Göran Andersson, University of Uppsala, Sweden – Prof Philippe Baret Université Catholique de Louvain, Belgium – Prof Michael W. Bruford, Cardiff University, United Kingdom – Prof Nadine Buys, Katholieke Universiteit Leuven, Belgium – Prof Ino Curik, University of Zagreb,Croatia – Dr Juha Kantanen MTT Agrifood Research Finland, Finland – Dr Johannes A. Lenstra, University of Utrecht, The Netherlands – Prof Theo Meuwissen, Norwegian University of Life Sciences, Norway – Prof Jutta Roosen, Technische Universität München, Germany – Prof Johann Sölkner, University of Natural Resources and Applied Life Sciences, Austria – Prof. Michele Tixier-Bichard, INRA, France

Advisory Expert: Prof Paolo Ajmone Marsan, Università Cattolica del Sacro Cuore, Italy

Project coordination : Elena Murelli

ESF Liaison : Dr Maria Manuela Nogueira, Science | Eléonore Piémont, Administration

1 quai Lezay-Marnésia • BP 90015
67080 Strasbourg cedex • France
Tel: +33 (0)3 88 76 71 00 • Fax: +33 (0)3 88 37 05 32
www.esf.org