EUROPEAN SCIENCE FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

Research Networking Programme
NEWSLETTER No. 6 August 2014

ESF GENOMIC-RESOURCES

Standing Committee for Life, Earth and Environmental Sciences (LESC)

Introduction

End of 2013, it was the last opportunity for the steering committee of GENOMIC RESOURCES to fund exchange grants in the domain of the management of Farm Animal Genomic Resources. Eight projects were selected and the corresponding stays took place during the last six months of the project, between January and June 2014. They are described in the current issue of the GENOMIC RESOURCES Newsletter.

What's new ?

Final conference

GENOMIC RESOURCES funded a final conference hold in Cardiff between June 16 and 18, 2014, organized by Michael Bruford, Pablo Orozco-terWengel and Mafalda Costa. The talks and integrative activities focused on four subtopics related to livestock conservation practice in a changing world, aiming at understanding rapid changes in conservation practices in light of policy developments, climate change and diversifying market demands: i) redefining the role of genome data in livestock conservation and prioritization; ii) identifying improved and more integrative analysis methods for livestock genomic, environmental and socio-economic data; iii) assess genomic resources for minority livestock species and breeds; iv) horizon-scanning for the twenty most important problems to be overcome for effective livestock resource conservation during the next decade. The final GENOMIC RESOURCES Newsletter (issue#7) will present the main outcomes of this scientific event.

Exchange Visits 2013-2014: final reports of the projects funded

Project N.1

Phenotypic study of captivity-induced plasticity in wild boars (Sus scrofa)

Maxime Garica, Austria, visited the Unit 'Physiologie de la Reproduction et du Comportement', INRA, France

Purpose of the visit

The 3-month exchange visit's purpose was to establish the wild boar vocal repertoire, using acoustic recordings and behavioral observations on a group of wild boars that were available at the host's institution. This represents the first step of a longer-term project and provides a strong and essential basis to carry on shared research between the 3 institutions involved: the Museum National d'Histoire Naturelle (MNHN), and the Institut National de la Recherche Agronomique (INRA) in France, and the Cogbio Lab_department of Cognitive Biology, University of Vienna, Austria. This longer-term project aims at investigating the slight changes brought by living in a domestication context as compared to wild. Knowing more about the vocal repertoire will allow a better understanding of the way wild boars perceive their environment and interact with each other, and help us gain insight concerning the behaviors of their closest relative, the domestic pig.

The long-term project, which depends entirely on this exchange visit's outcome, will investigate the ontogeny of the vocal apparatus and which are the phenotypic changes induced on acoustic communication by assessing the effect(s) of domestication (on animals living in semi-natural condition vs. in individual pens). Knowing more about this species repertoire is hence decisive. Eventually, the applied goal of this research is to use acoustics as a way to monitor suidae's internal state to a further extent than currently done. This will also make pig industry more knowledgeable concerning the required conditions for better housing quality and breeding programs. Thus from establishing the vocal repertoire as a first step, the final collaboration project will deal with animal welfare, husbandry conditions and species/breed conservation.

Description of the work carried out during the visit

It is important to note that an unexpected event occurred in the early stage of the visit: some captured animals were diagnosed positive to PRRS (Porcine Reproductive and Respiratory Syndrom).

While this disease does not prevent the animals from living a rather normal life, it threatened the long-term project because of sanitary conditions (animals have to eventually be CT-Scanned at INRA) and the group had to be euthanized.

The host's institution and members made it up to the situation by contacting a known wild boar local breeder who could supply new, disease-free animals. His parks were made accessible to Maxime Garcia (ESF Grant trustee) so that he could make the behavioral observations and acoustic recordings, purpose of the visit.

By studying the available population intensively, with a constant presence in the field, MG could get close enough to the animals to observe and record natural behaviors and vocalizations. A special situation of capture and following housing (in a stable owned by the breeder) allowed him to record sounds made impossible to get in the outdoor parks because of their production patterns (very low level and low frequency sounds).

Part of the Exchange visit's period also consisted in developing tools useful for subsequent analysis, and some advanced programming was done on a sound analysis software, to pursue with data processing (which has now started).

Description of the main results obtained

Even though data acquisition was made extremely hard because of non-controlled, natural conditions (wind, non-vocal noises made by the species of interest and other species, rain, anthropogenic noise, hardly recordable sounds due to production characteristics…), several thousands of calls could be recorded, among which more than 800 were assessed as of good quality for further analysis.

A primary analysis reveals that high pitched calls are expressed in negative valence situations. This confirms what has been found in studies carried on domestic pigs and welfare assessment.

The novelty of the results, from what comes out in this preliminary stage of the data analysis, is that not only what can be described as "screams" or "squeal" (high frequency calls) are emitted in negative situations. Even what has been called has "grunts" (low frequency calls) can be emitted in such context.

It seems that there is a continuum in the calls produced in this species, translating into their internal state: more specifically, a gradation in the signal can emerge, related to the state of arousal of an individual: potentially, an increasing negative valence could be correlated with a transition from low frequency to high frequency calls (e.g: from "grunts" to "squeals"/"screams") as well as with a transition from a voiced signal to a signal containing Non-Linear Phenomena (an occurrence in many mammal acoustic signals, which has been

suggested to be a cue to a high arousal state and attention grabbing).
A main and direct outcome from these observations is that studies using vocalizations to assess animal welfare should focus, not only on high-pitched calls (negative valence with high arousal), but also on lower pitched ones, that could as well indicate a negative valence state with lower arousal. By giving importance to such lower frequency calls, one could prevent the animals from experiencing a transition from 'negative valence and low arousal' to 'negative valence and high arousal'. Keeping a low level negative valence (versus high level negative) could already be a major achievement in implementing better housing conditions and breeding programs.

Future collaboration with host institution
This study was the necessary foundation for a longer-term collaborative project aiming at identifying which can be the subtle morphological changes related to the domestication process, in relation with the production and expression of vocalizations in wild boars. While some interesting and useful results are already foreseeable, this project is still at its primary stage and will last for two years, during which the host and the Exchange visit grant trustee will work together.

Projected publications / articles resulting or to result from the grant
One publication detailing the composition of wild boars vocal repertoire is planned once the analysis will be completed (publication expected within 8 months to a year from the end of the study).
Eventually, the current study will serve a long-term project between the three institutions (MNHN & INRA, France – CogBio, Austria) and the ESF will then be acknowledged accordingly.

Other comments
The results of the present study will be presented at the European Conference on Behavioural Ecology (ECBB2014), being held in Prague from July 17th to July 20th.
I am truly grateful to Dr. Yann Locatelli (host researcher), who introduced me to personal from INRA (Unit 'Physiologie de la Reproduction et du Comportement') with whom he works.
We could also discuss about long-term collaboration, possibly concerning the time by which my PhD will be completed.
Finally, I would like to thank the ESF for this Exchange visit grant. I would also like to thank Prof. W. Tecumseh Fitch, director of the Cognitive Biology Department, University of Vienna, whose university grant covered the costs that exceeded the amount provided by the ESF grant.

Project N.2

Assessing the demographic history of Old World camelids *(Camelussp.)* through whole genome sequencing

Robert Fitak from Austria visited Cardiff University School of Biosciences, Cardiff, UK.

Purpose of the visit

The objective of this study was to apply two recently developed methods that infer the demographic history of populations from genetic data to a dataset of Old World camels (Camelussp). This study was an initial component of a larger research program by our group to understand the evolutionary history of Old World camels, specifically to identify and conserve camel genomic diversity from both an agricultural and wildlife perspective. Each perspective requires characterization of both neutral levels of genomic variation in addition to loci influenced by selection, such as those responsible for local adoption and/or economically relevant traits. Unfortunately, locally reduced levels of genetic variation which are often a signature of selection can,

at times, be indistinguishable from the effects of certain demographic processes (e.g. bottlenecks, inbreeding, etc). Despite the dogma of the selection vs. demography dilemma ('selection acts on relatively small regions whereas demography affects the entire genome'), the stochastic nature of the coalescent can still lead to inconsistent patterns among loci. Nonetheless, understanding and accounting for the demographic history of a population is imperative when concluding which loci may be under the coercion of selection (Rosenberg and Nordborg 2002).

In addition to inferring the demographic history of our various camel populations, we found it necessary to investigate the effects of two important technical concerns often encountered when working with whole-genome resequencing data. First, does the level of divergence between an individual's sequenced genome and the reference genome that was used for mapping reads have a significant impact on the observed demographic history? And second, does the removal of polymorphisms from repetitive regions also impact the observed demographic history?

Description of the work carried out during the visit

The dataset consisted of whole-genome shotgun sequences from three species of Old World camel: C. dromedaries (n = 9), C. ferus (n = 9), and C. bactrianus (n = 7). Paired-end sequencing of each individual was performed on a single lane of an IlluminaHiSeq with a mean insert size of 500 bp. Reads were 3' trimmed to a minimum phred base quality score of 20 and minimum length of 50 bp using POPOOLATION v1.2.2 (Kofler et al. 2011). Trimmed reads were mapped to the C. ferus CB1 genome assembly (Genbank ID GCA_000311805.2) using BWA v0.6.2 (Li and Durbin 2009). For C. dromedarius, reads were also mapped to an in-house de novo assembly of the dromedary genome. All alignments were filtered to contain only high quality (MQ ≥ 20), unique (MarkDuplicates; PICARD v1.89, http://picard.sourceforge.net), unambiguously mapped, and properly paired reads (SAMTOOLS v0.1.19, Li et al. 2009). All single nucleotide polymorphisms (SNPs) and corresponding genotypes used in downstream analyses were identified using the GATK Best Practices Pipeline (Van der Auwera et al. 2013).

To model the demographic history, we first used the Pairwise Sequentially Markovian Coalescent model (PSMC; Li and Durbin 2011). Analyses were performed according to default parameters with a minimum and maximum coverage cutoff per base set to 5 and 40, respectively, and with 100 bootstrap replicates. Results were plotted assuming a generation time of 5 years for camels. To remove the potential effects of repetitive elements on the distribution of heterozygous sites, we constructed a consensus genome sequence for each individual using the 'FastaAlternateReferenceMaker' walker in GATK using our cleaned genotype data (VCF file). Subsequently, the repetitive regions in the resulting genome were masked using the annotations available for the CB1 assembly. The PSMC analysis was then repeated as described above.

Because the method implemented in PSMC lacks power to detect recent (less than ~2000 years ago in camels) changes in effective population size (Ne) due to the lack of recombination events (Li and Durbin 2011), we used the software SNeP (Barbato et al., under development) to infer more recent demographic history. SNeP uses the extent of linkage disequilibrium (LD) between markers at various distances to infer Ne¬¬. For each species of camel we included SNPs found in the 5000 largest scaffolds that were polymorphic (minimum allele frequency ≥ 0.05), had a genotyping rate ≥ 0.9, and did not deviate from Hardy-

Weinberg equilibrium (HWE; $p \geq 0.0001$) using PLINK v1.9 (https://www.cog-genomics.org/plink2). Because the computational time necessary for SNeP increases exponentially with the number of SNPs, we determined the threshold where increasing the number of SNPs did not produce significant decreases in variance of Ne. We generated 10 random subsets of 5,000, 10,000, 20,000, 50,000, and 100,000 SNPs in C. ferus, phased the genotype data using BEAGLEx v3.3.2 (Browning and Browning 2007)and ran SNeP using the method of Corbin et al. (2010). The previous pipeline was replicated 10 times in each species for the threshold number of SNPs identified.

Description of the main results obtained

The mean coverage of each genome alignment was approximately ~15X. In total, across all three species of camels 4,960,087 SNPs were called. Insertion/deletion polymorphisms were excluded from downstream analyses. The results of the demographic history recreated from raw genome alignments using PSMC can be found in Figure 1.

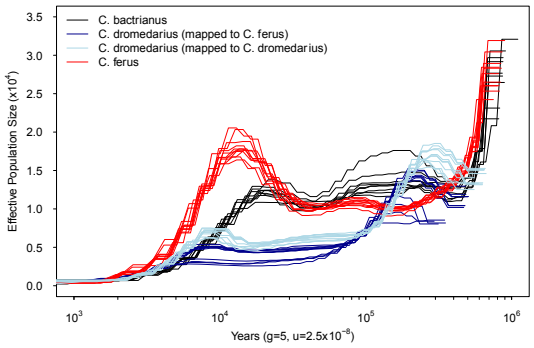

Figure 1:*Demographic history of three camel species up to 1,000 years ago reconstructed using PSMC (Li and Durbin 2011). Each line represents the estimated effective population size (N¬e) for an individual genome according to the colors depicted in the legend. Bootstrap replicates are not shown for clarity, but variance is quite large more recently than 103 and older than ~5x105 years ago.*
Abbreviations: g = generation time, u = mutation rate.

All three species appeared to have suffered an extended period of relatively small Ne between 200k – 20k years ago. This population reduction may be a result of the last glacial maximum associated with the late Pleistocene between 120k – 12k years ago.A similar pattern has been observed in other mammals from the Northern Hemisphere (e.g. Zhao et al. 2012). In the wild camel, C. ferus, a rapid population expansion occurred potentially following the last glacial maximum but then abruptly ended and a second bottleneck followed, probably eventually resulting in the small Ne observed today. The domestic C. bactrianus shared a similar ancestry with C. ferus until 20k years ago, where it experienced a less substantial expansion and eventual reduction. The most recent bottleneck, however, appeared to have been earlier and less severe than in C. ferus. Although speculative at this point, contact with early humans in the region may have initiated this decline prior to domestication ~6k years ago. Finally, in dromedaries, a similar pattern also emerged, albeit the events occurred more recently and N¬e was consistently smaller than in the other two species. It is possible that parameters such as generation time and mutation rate are quite different in C. dromedaries compared with the other species, and these factors need to be investigated further.

We also investigated the potential effects of two technical concerns: 1) mapping to a relatively divergent reference genome and 2) the effect of removing SNPs from repetitive regions. In the former, we mapped our C. dromedarius reads to both the C. ferus reference genome and a de novo dromedary genome in preparation in our laboratory (Figure 1 dark blue and light blue lines, respectively). Despite a divergence time of ~5 million years ago between C. dromedarius and C. ferus, the effects on historical Ne were rather minimal. Mapping to the more divergent genome appears to consistently underestimate Ne, but the exact mechanism for this remains to

be determined. Several factors such as genome assembly quality, repeat structure, SNP-calling algorithms, etc. may influence this result. In the latter, the removal of repetitive regions prior to analysis with PSMC results in similar patterns but with several important differences (Figure 2). First, the absolute estimates of effective population size are much lower than the previous estimates and uninformative prior to ~13k years ago; probably a result of the fewer number of heterozygous bins thus fewer coalescent events. Second, the most recent population decline in both C. ferus and C. bactrianus occurred several thousand years more recently than in Figure 1, congruent with the current archaeological estimates of initial domestication.

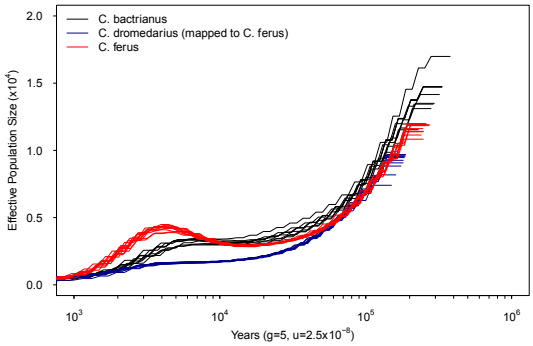

Figure 2: *Historical effective population sizes of each camel species recreated using PSMC (Li and Durbin 2011) and a set of genome-wide SNPs filtered to exclude annotated repetitive regions. Each line represents an individual genome according to the colors depicted in the legend. Bootstrap replicates are not shown for clarity, but variance is quite large more recently than 103 and older than ~5x105 years ago.*
Abbreviations: g = generation time, u = mutation rate.

Because PSMC lacks resolution in the recent (last few thousand years) past, we used the program SNeP to explore patterns in Ne using LD. Using just C. bactrianus, we found that 20k SNPs was sufficient to estimate Ne and increasing the number of SNPs did not gain any resolution in demographic history but rather resulted in large increases in computational time (Figure 3).

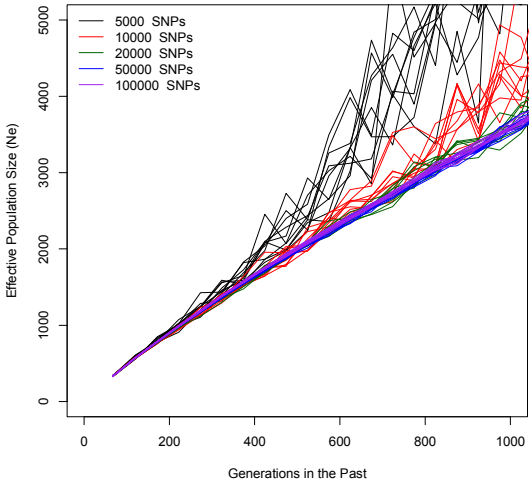

Figure 3: *Estimates of Ne from SNeP. Each line represents a different permutation of the number of SNPs indicated by the color in the legend. There is little improvement when using 20k SNPs or more.*

Using permutations of 20k SNPs, we observed a similar population trajectory indicating a gradual decline in all three camel species over the last 1,000 generations (~5,000 years) (Figure 4).

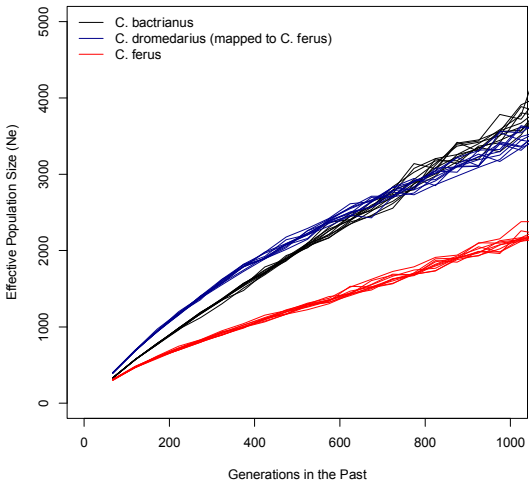

Figure 4: *Estimates of Ne from SNeP for each species of camel (see legend for colors). Each line represents a different permutation of 20k SNPs.*

This is in agreement with the estimates of PSMC and expectations of domestication. However, using SNeP, C. ferus consistently had a smaller Ne than the other species, which was not observed in the PSMC results. This methods, however, assumes

accurate estimates of allele frequencies to infer LD within a population, and our dataset, which is based on a small sample size (n = 7 or 9), may be too small to reliably estimate LD. Furthermore, it is possible our dataset contains individuals from structured populations, which may also confound calculations of LD.

Future collaboration with host institution

The interaction and collaboration with the host was a fruitful experience. We plan on staying on close contact because both groups are working on similar questions and analyses in two different groups of camelids. Additionally, we will remain in close contact in order to further test the utility of the software SNeP, which is being developed in the Bruford Lab. Finally, the collaboration will continue in a larger project investigating the patterns of demographic history in vertebrates in general from whole genome sequencing.

Projected publications/articles resulting or to result from your grant

Our future plan is for these results to incorporated into two future publications, one focused on our new dromedary genome assembly and a second on demographic history and selection in the other two species (C. ferusand C. bactrianus). In both cases the ESF and the host institution will receive proper acknowledgement for their support.

Other comments

I am extremely grateful to Dr. Mike Bruford and the Cardiff University School of Biosciences not only for their support during my stay, but providing important information necessary when encountering visa complications. I would also like to thank all members of the Bruford Laboratory, especially Dr. Pablo Orozco-terWengel and Mario Barbato for the help I received from them. Finally, thanks to JukkaCorander and the CSC — IT Center for Science Ltd in Finland for use of their computational resources.

References

- Browning, S.R. and Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81, 1084-1097.
- Corbin, L.J., Blott, S.C., Swinburne, J.E., Vaudin, M., Bishop, S.C. and Woolliams, J.A., 2010.Linkage disequilibrium and historical effective population size in the Thoroughbred horse.Animal Genetics 41, 8-15.
- Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C. and Schlotterer, C., 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6, e15925.
- Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
- Li, H. and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences.Nature 475, 493-496.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009.The Sequence Alignment/Map format and SAMtools.Bioinformatics 25, 2078-2079.
- Rosenberg, N.A. and Nordborg, M., 2002.Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.Nature Reviews Genetics 3, 380-390.
- Van der Auwera, G.A., Carneiro, M.O., et al. 2002. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline.

Current Protocols in Bioinformatics, 11:11.10:11.10.1–11.10.33.

- Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., Hu, Y., He, W., Zhang, S., Fan, W., Zhu, L., Li, D., Zhang, X., Chen, Q., Zhang, H., Zhang, Z., Jin, X., Zhang, J., Yang, H., Wang, J., Wang, J. and Wei, F., 2012. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nature Genetics 45, 67-71.

Project N.3

The Role of Farm Animal Genetic Resources for Sustainable Intensification

Ahmadi Bouda Vosough from UK visited the University of Louvain-la-Neuve, Belgium

Purpose of the visit

The objective of my visit was to gain a closer insight about the theoretical background and analytical methods and approaches that are used by social and natural scientists at the Université catholique de Louvain (UCL) in relation to sustainable agricultural systems. Particularly I was interested in comparing a number of available and tested socio-economic decision-support methods with respect to their capacity in incorporating various dimensions of sustainable intensification and to identify and discuss their application to farm animal genetic recourses (FAnGR).

Description of the work carried out during the visit

I spent a period of 4 weeks (28th April to 23rd May 2014) at UCL with Professor Philippe Baret in his lab in Louvain-la-Neuve. During the first week of my visit, I conducted a literature review focusing on 'sustainable intensification' (e.g. Godfray & Garnett 2014), 'ecological agriculture' (e.g. Weiner 2003), 'transitions and systems changes' (e.g. Geels & Schot 2007) and 'economic theory of biodiversity preservation' (Weitzman 1998).

In the second week, I read and learned about 'cognitive mapping approach' that is considered as a decision-support tool that was recently applied to analysing systems of practices in social-ecological systems by Vanwindekens et al (2013) and Vanwindekens et al (2014). Also in the second week, I met Dr Muriel Tichit (director of research at INRA) and discussed about the details of my research particularly the components of sustainable intensification.

In the third week, I visited the Biodiversity governance (BIOGOV) research unit of UCL and met the head of the unit, Professor Tom Dedeurwaerdere, and his fellow researchers. In my one day visit to this research unit I became familiar with a range of research topics including: environmental ethics, philosophical views on disagreement in science and Putnam's epistemological shift, geographical indication and intellectual property, convention on biological diversity (Nagoya protocol), environmental justice, governance of collective actions (e.g. farmers collective actions for biodiversity) and the impact of specialised knowledge brokers on the adaptation of greening measures. Based on the activities in the previous three weeks, I wrote and submitted an abstract to Frontiers journal entitled "Comparing decision-support systems for sustainable intensification: an application to FAnGR".

In the fourth and last week of my visit I started to expand the submitted abstract to be able to discuss the first draft with Prof Baret while still in UCL.

Description of the main results obtained

The conducted literature review, the scientific meetings attended and the discussions I had with a number of professors and researchers at UCL provided

me an insight on the concept of sustainable intensification and its related criticism as well as on application of decision-support tools to socio-ecological systems and issues such as animal genetic resource conservation issue. These insights are presented and discussed in our forthcoming paper. The abstract of this paper is presented below):

Sustainable intensification (SI) is a multifaceted concept incorporating the ambition to increase or maintain the current level of agricultural yields while reduce negative ecological and environmental impacts by using a broad range of production methods and consumption patterns. Integrated analytical methods such as econometric methods, optimisation models, non-market valuation, and many other methods have been used to support decision making processes at different levels of agricultural systems. However, their capability in adapting to a holistic view of agricultural systems (in oppose to a reductionist view) to fulfil objectives of SI varies considerably. Further, these methods often consist of set of values, objectives and implicit assumptions that may be inconsistent or in conflict with merits and objectives of SI. These potential conflicts will have consequences for adoption and up-take of agricultural research and technologies such as genetic technology in pursuit of SI. Interdisciplinary research that integrates natural and social sciences is needed to provide guidance on feasibility, practicality and policy implementation for SI. The objectives of this paper are to compare a number of available and tested socio-economic decision-support methods with respect to their capacity in incorporating various dimensions of SI and to identify and discuss their application to FAnGR.

Future collaboration with host institution

We identified and agreed on two main areas of: 1- PhD/MSc student exchange and 2- developing joint proposals that both parties (UCL and SRUC) could collaborate in future work.

Projected publications / articles resulting or to result from the grant

The above mentioned abstract was accepted, on 30th May, and we were invited to submit a full article for peer-review by 30 September 2014. The work on this paper is currently under progress

Other comments

I would like to express my sincere appreciation to my host researcher Professor Philippe Baret whose scientific knowledge and broad view and experience was a great support for me during my visit to UCL. Despite of time constraint with Philippe's help, I managed to come up with a research plan and take the first initial steps of writing a joint paper. I am also very thankful to staff and researchers at Faculty of biological, agronomical an environmental engineering, as well as Earth and Life Institute of UCL for their hospitality and knowledge exchange. In particular I am thankful to Sophie T'Kint, Antoinette Dumont, Dr Julie Van Damme, Dr Frédéric Vanwindekens and Prof Tom Dedeurwaerdere for all their kind helps and hospitality. I am also very thankful to the European Science Foundation for providing this exciting opportunity.

References

- Geels, F. W., and Schot, J. (2007). Typology of sociotechnical transition pathways. Research policy, 36(3), 399-417.
- Godfray H. C.J. and Garnett T. (2014). Food security and sustainable intensification, Phil. Trans. R. Soc. B 5 April 2014 vol. 369 no. 1639 20120273.
- Vanwindekens, F., Stilmant, D., Baret, P. (2013). Development of a broadened cognitive mapping approach for analysing systems of practices in social-ecological

systems. In: Ecological Modelling, Vol. 250, p. 352-362.

- Vanwindekens, F., Baret, P., Stilmant, D. (2014). A new approach for comparing and categorizing farmers' systems of practice based on cognitive mapping and graph theory indicators. In: Ecological Modelling, Vol. 274, p. 1-11.
- Weiner, J. (2003). Ecology–the science of agriculture in the 21st century. The Journal of Agricultural Science, 141(3-4), 371-377.
- Weitzman, M. L. (1998). The Noah's ark problem. Econometrica, 1279-1298.

Project N.4

Identification of Signatures of Selection in Cattle from Next-Generation Sequencing Data

Daniel Fisher from Finland visited the Institut für Populationsgenetik, University of Veterinary Medicine in Vienna

Purpose of the visit

Genetic characterization of livestock populations has traditionally been based on very limited sets of potentially neutrally evolving microsatellite markers. A recent paper by Orozco-terWengel et al. (2011) showed that much larger numbers of loci are needed for reliable demographic inference.

With the onset of high throughput genotyping technology, genetic diversity studies are increasingly being done with much bigger sets of loci involving a couple of thousand single nucleotide polymorphisms (SNPs), genotyped on commercially available SNP chips such as the Illumina 3k, 6k and 50k bovine SNP chip. However, SNPs on commercially available SNP chips are pre-selected, based on provider defined criteria, and therefore do not represent unbiased sets of polymorphisms in the genome.

Ever decreasing costs and higher output of Next-Generation Sequencing (NGS) technologies allows us to do population genetics with unprecedented large-scale genetic data, investigating literally all polymorphisms (SNPs, insertion - deletion polymorphisms and structural variants including copy number polymorphism at single base pair resolution) to determine levels of genetic variation in and genetic differentiation between populations and to study evolutionary forces affecting genetic variation such as mutation, natural selection and genetic drift.

Description of the work carried out during the visit.

I spent a period of approx. 3 weeks (16. March until 8. April) at the Institut für Populationsgenetik, University of Veterinary Medicine in Vienna,in order to work together with Dr. Marlies Dolezal.

During this time we developed pipelines to do population genetic analysis for genome-wide SNPs called from Illumina paired-end next generation whole genome sequence data in Brown Swiss (BSW) and Finnish Ayrshire (FAY) dairy cattle breeds. This data is available to us from the FP7 funded project QUANTOMICS contract n. 222664-2. SNPs had been called as part of run 3 of the 1000 bulls consortium in a multi sample setting with samtools v0.1.18. To reduce false positive calls and to phase the data beagle v3 (Browning and Browning (2011)) was run. This resulted in 28.1 Mio. SNPs genome-wide for BS and FAY.

Description of the main results obtained

Within population analysis - Summary statistics: We used-chromosome wise minor allele frequency (MAF) histograms (see Figure 1 for an example) as a first means of data inspection.

As allele frequency distributions looked reasonable assuming that the majority of SNPs are evolving under neutrality we did not apply any further data filtering. We then calculated classical population genetic parameters Pi (Nei and Li (1979)), Tajima's D (Tajima (1989)) using VCFtools v0.1.12

option –TajimaD and -window-pi. Values of Pi and Tajima's D are strongly influenced by the window sizes in which they are calculated.

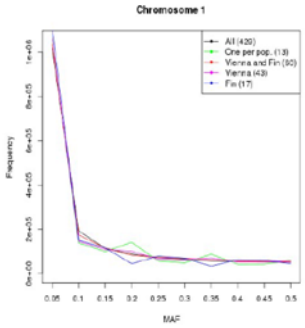

Figure 1:Minor allele frequence histogram. Histogram contains MAF for more populations than were used in this project, which can be ignored for this report. Vienna refers to BSW and Fin to FAY.

Due to lack of sound theory on how to estimate an optimal window size we empirically evaluated statistics calculated at different window sizes of 100, 500, 1000, 2500, 5000 and 10000 base pairs. We used visualizations as shown in Figure 2 to determine the best possible window size. Windows of size 1kb appeared to have a high signal to noise ratio and gave best agreement between test statistics and hence was chosen as window size for all further analyses. To evaluate the amount of noise due to false positive SNP calls we sub-set the NGS based SNP calls based on SNPs that are genotyped on either of the commercial bovine high density SNP arrays (777k Illumina BovineHD BeadChip and Affymetrix's GeneChip(R) Bovine Genome Array) The drawback of this approach is the lack of rare variants as commercial chips suffer from the so called ascertainment bias. We refer to this subset as the HD-variants.

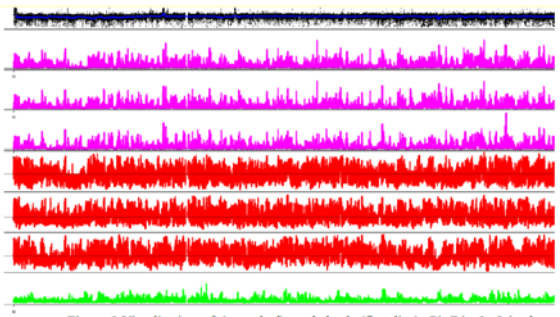

Figure 2:Visualization of (smoothed) read depth (first line), Pi (Line2: Joined populations, Line3: Brown Swiss, Line4: Finnish Ayrshire) and Tajima D (Line5: Joined populations, Line6: Brown Swiss, Line7: Finnish Ayrshire)visualization.) and Fst between Brown Swiss and Finnish Ayrshire (Line8) for window size 1000.

Statistics relying on site frequency spectra:
Nielsen et al (2005) proposed a composite Likelihood Ratio (CLR) test to identify selective sweeps from the data. However, calculating the CLR genome-wide with the tool sweepfinder (http://people.binf.ku.dk/rasmus/webpage/sf.html) is a computationally heavy task. SweeD (Pavlidis et al.2013) offers a parallel version of the original sweepfinder implementation that overcomes the computational burden and enabled us to calculate the CLR test statistic also for different window sizes. Also, sound theory for choosing the right window-size is not available. Again, we visualized the effect of different window sizes onto the CLR statistic and found that also here a window of 1000 is ideal.
All scripts were developed in a generic way to be easily applicable on other similar datasets. As part of the run3 data from the 1000 Bulls project we also have access to more populations and will later apply above developed functions to this data, too.

Between population analysis:
Moreover, we also calculated between-population statistics, namely the fixation index Fst (Hudson et al. (1992)). The fixation index was also calculated using the window size of 1000 and the values of Fst were then added to the Figure 2. The fixation index Fst can lead as well to an exhaustive amount of false positive signals and for that reason we tried out different window sizes and compared the signals, we received. We examined the Fst signals for window sizes between 1000 and 10.000 using a dense grid of step-size 2000 and also here, 1000 turned out to be the ideal window size.

Future collaboration with host institution
Further analyses to detect positive selection using Extended Haplotype Homozygosity (EHH) and Integrated Haplotype Score (iHS) with the software selscan (Szpiech, Hernandez 2014) as well as the interpretation of the derived results is currently ongoing.

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

The results and pipelines will be presented at the "Livestock Genomic Resources in a Changing World" conference in Cardiff (17.6- 19.6.) meeting, if accepted. The abstract title is similar to the project title "Identification of Signatures of Selection in Cattle from Next- Generation Sequencing Data". Results based on the work done during the visit will be also presented at the Livestock Genomics meeting in Cambridge in September 2014. Finally, we aim to publish results based on this work in peer-reviewed journals.

References

- Browning, B.L., and Browning, S. R. (2011): A fast, powerful method for detecting identity by descent. The American Journal of Human Genetics 88:173-182.
- DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. (2011): A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 43:491-498.
- Hudson, R.R., Slatkin, M., Maddison, W.P. (1992): Estimation of Levels of Gene Flow from DNA Sequence Data. Genetics 132 (2): 583–9.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A. (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-303.
- Nei, M., Li, W.-H. (1979): Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases. PNAS 76 (10): 5269–73.
- Nielsen, R., Williamson, L., Kim, Y., Hubisz, M., Clark, A., et al. (2005): Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575
- Orozco-terWengel, P., Corander, J., Schlötterer, C. (2011): Genealogical lineage sorting leads to significant, but incorrect Bayesian multilocus inference of population structure. Molecular Ecology 20: 1108–1121; doi: 10.1111/j.1365-294X. 2010.04990.x
- Pavlidis, P., Živković, D., Stamatakis, A., Alachiotis, N. (2013): SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol. doi:10.1093/molbev/mst112.
- Szpiech, Z. A., Hernandez, R. D. (2014): Selscan: an efficient multi-threaded program to perform EHH-based scans for positive selection. arXiv:1403.6854.
- Tajima, F. (1989): Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123 (3): 585

Project N.5

Maintaining fitness and diversity using genomic measures of coancestry

Mirte Bosse from The Netherlands visited The Laboratoire d'Eco-anthrophologie et ethnobiologie, Muséum National d'Histoire Naturelle, CNRS, France

Purpose of the visit

Conservation programmes aim at optimizing the probability of survival of the population in the programme, which is obtained by maintaining maximum values of genetic diversity [de Cara et al 2011]. However, when molecular data is used to achieve this goal, deleterious variants could be kept in a population, which can have a negative effect on the overall fitness [de

Cara et al 2013]. Recently, a measure of coancestry based on shared regions of the genome has been proposed as a compromise to maintain both fitness and genetic diversity when the population in the programme has some inbreeding load [de Cara et al 2013]. These predictions for management based on genealogical, molecular or IBD segments have been tested with simulated data, but have so far not been implemented with real data. The publication of the pig reference genome in 2012 [Groenen et al 2012] created the opportunity to analyze pig genomes in greatest detail. Stretches of shared coancestry are occurring often in the genome of individual pigs [Bosse et al 2012], and therefore pigs are an excellent model to test the different proposed management strategies. By using re-sequence data of two highly divergent pig populations, we tested the hypothesis that the management strategy based on shared genome segments can maintain both diversity and fitness. This exchange visit enabled a collaboration between Angeles de Cara and Mirte Bosse in which their expertise on in silico population management and pig genomics were combined.

Description of the work carried out during the visit

Within this project, we used re-sequence data, genotype data and pedigree information from two very different pig populations: a commercial line from the European Pietrain breed and a zoo population from the endangered species Sus cebifrons. Five Cebifrons and 11 Pietrain individuals were re-sequenced to ~10x depth of coverage each and aligned to the Sus scrofa reference genome build 10.2 [Groenen et al 2012]. In addition, 47 Pietrain pigs were genotyped on the Illumina porcine 60K iselect beadchip [Ramos et al 2009]. Before the start of in silico management, we examined the background of both populations for their most important characteristics. The 5 re-sequenced Cebifrons individuals and 11 Pietrain pigs were compared in terms of their nucleotide diversity and distribution of variable sites over the genome. A filtered genotype matrix including ~100.000 variable sites was constructed from the Cebifrons re-sequence data, and for the Pietrain population the genotypes from the SNPchip were used. For all re-sequenced individuals, we checked whether their genome contained deleterious mutations with the Variant Effect Predictor (VEP) and these alleles were added to the genotype matrices.

Because sequence data is available for only a subset of the Pietrain dataset, we included deleterious positions in all individuals and assigned randomly the allele at these positions, so that each individual roughly contained the same proportion of deleterious variants. All chromosomes were phased separately with shape it to reconstruct the haplotypes that were present before the start of the in silico management. The status of both populations before the management started was recorded in terms of observed heterozygosity, mean fitness and shared coancestry.

Since Sus cebifrons is an endangered species, we conducted an analysis of their past effective population size in order to gain more insight into their demographic history.

The past Ne of the Cebifrons population was examined with two independent methods using the distribution of variation in individual genomes derived from re-sequence data.

We used the pairwise sequential Markovian coalescent (PSMC, Li and Durbin 2011) on all Cebifrons individuals and the runs of homozygosity method described by MacLeod et al (2013) in one of the males to compare the performance of the methods.

Because Pietrain is a commercial pig breed, we screened the Pietrain population for signatures of selection. Although maintenance of variation is important in these populations, diversity in some regions should be reduced because haplotypes linked to particular commercial traits are present. We used the R rehh package [Gautier and Vitalis 2013] to check for extended haplotype homozygosity in the Pietrain 60K dataset and used PLINK [Purcell et al 2007] to screen the genomes for runs of homozygosity.

Using our own Fortran code based on [de Cara et al 2013], we managed both populations for 10 generations, maintaining the same population size and sex ratio as in the initial population. All management strategies were replicated in silico 100 times for each population. Management was based on optimizing diversity by using optimal contributions based on three different measures of coancestry: 1) genealogical coancestry as obtained from pedigree information, where available (when not available, we assumed those individuals unrelated); 2) molecular coancestry as obtained from IBS status of all markers; 3) coancestry based on measurements of IBD segments between individuals.

Recombination events per chromosome were drawn from a Poisson distribution based on the mean recombination rate for each chromosome as described in [Tortereau et al 2012].

After each generation, we measured the observed heterozygosity, expected heterozygosity, average coancestry and the average fitness of the population. Finally, we checked the persistence of selection signatures in the Pietrain after 10 generations of each management strategy.

Description of the main results obtained

Both populations showed signatures of inbreeding in their genomes, but referring to different time points. We assessed the demographic history of the Cebifrons individuals with two independent methods (MacLeod method displayed in Figure 1). Two major bottlenecks can be observed, which coincide with past fluctuations in sea level and glaciations [Frantz et al 2013]. Both methods show roughly the same patterns of population expansion and reduction, suggesting that the Cebifrons population has been small for a substantial period of time.

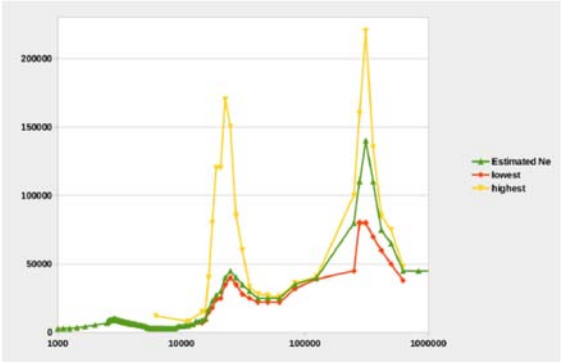

Figure 1. *Past effective population size for Cebifrons population. Ne obtained for one male Cebifrons with the ROH-based method as described by MacLeod. The original estimate is in green, confidence intervals are between the red and yellow line. Time (years in the past) is displayed on the x-axis and the y-axis indicates the corresponding effective population size. We assumed a generation time of approximately 5 years.*

The effect of each management strategy was similar in both populations. Using molecular coancestry in the management maintained the most diversity, while managing based on genealogical coancestry maintained the least diversity (example for Cebifrons in Figure 2). The segment-based coancestry management resulted in intermediate levels of heterozygosity in the population, with longer segments containing less variation than short segments. When diversity levels are compared between the Cebifrons and Pietrain populations, it can be concluded that the decay of variation using the genealogical coancestry method is much stronger in Cebifrons than in Pietrain. This

result is expected because the Pietrain has a larger effective population size.

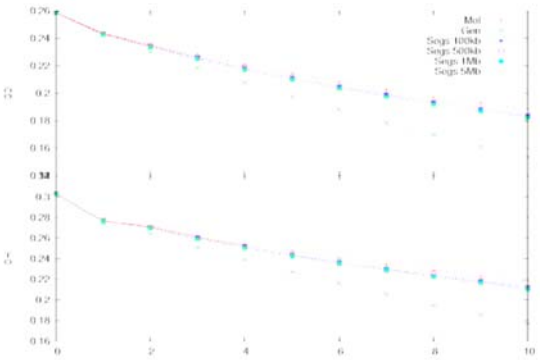

Figure 2. *Variantion in the Cebifrons population during three management strategies. Genetic diversity and heterozygosity in the Cebifrons population during three management for 10 generations: Molecular-based management, genealogical-based management and management based on segments of 4 different lengths.*

The results indicate that management based on molecular coancestry will maintain the highest diversity in the population. However, this strategy does not incorporate the possible negative effects from deleterious alleles that are maintained. When management is applied based on avoiding long segments of coancestry, the diversity is higher than based on genealogical coancestry, but the decrease of fitness should not be as severe as using molecular coancestry in the management. This is illustrated in the decrease of fitness over 10 generations in the Cebifrons populations, when the different management strategies are applied (Figure 3).

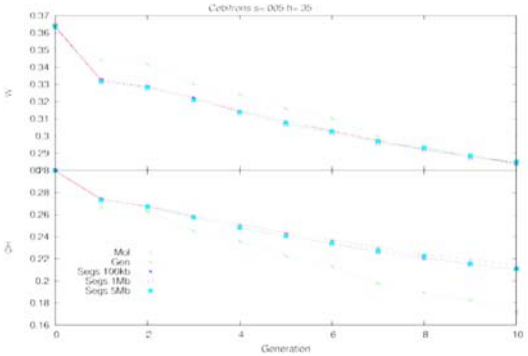

Figure 3. *Decay of fitness and observed heterozygosity in the Cebifrons population. Average fitness (W) and observed heterozygosity (OH) is measured in the original population and per generation after management with the molecular, genealogical and segment-based method.*

The Pietrain individuals contained some extended regions of homozygosity within their genome. In some instances, this pointed towards signatures of selection as inferred from the extended haplotype homozygosity test (Figure 4). However, we observed that the signatures of selection as inferred from clustering of long haplotyes are reduced in the Pietrain population after management.

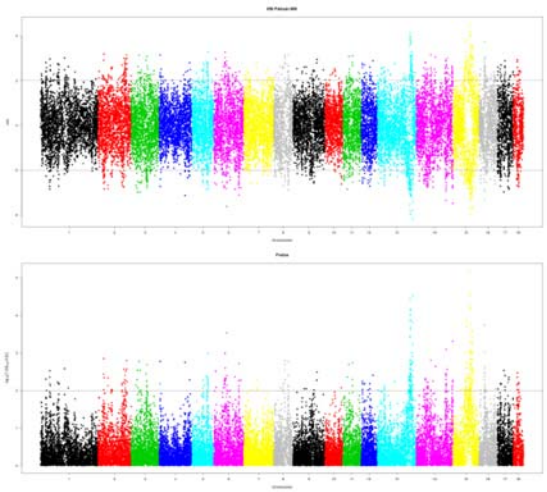

Figure 4. *Extended haplotype homozygosity per chromosome for the Pietrain population. 4A displays the raw iHS signal before management over all chromosomes in the Pietrain population. 4B shows the p-value of the iHS signal before management for each marker. Values >2 are defined to be significant.*

When artificially selected variants should be maintained in particular regions of the genome, the management strategy should be adapted to the requirements of the breeding goal.

Future collaboration with host institution

We expect exchanging visits between both labs in order to complete this specific project and depending on available data from the breeding companies, to pursue a detailed study of the demographic and selective history of some breeds.

Projected publications / articles resulting or to result from the grant

The aim is to submit our research paper to a high-impact journal within a few weeks after the exchange. Naturally, the ESF will be acknowledged for this travel grant in the resulting publication.

References

- Bosse M, Megens H-J, Madsen O et al. (2012). Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. PLoS Genetics 8(11): e1003100.doi:10.1371/journal.pgen.1003100
- de Cara et al. (2013) Using genomic tools to maintain diversity and fitness in conservation programmes Mol Ecol 6091–6099, December 2013
- de Cara, Fernandez, Toro and Villanueva (2011), "Using genome-wide information to minimise the loss of diversity in conservation programmes", J Animal Breeding and Genetics 128:456-464.
- Frantz L, Schraiber J, Madsen O et al. (2013). Genomic sequencing provides fine scale inference of evolutionary history. Genome Biology 14:R107.
- Gautier, Vitalis. rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure
- Groenen MAM, Archibald AL, Uenishi H et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.
- Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.
- MacLeod I et al. (2013) Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. Mol Biol Evol. Sep 2013; 30(9): 2209–2223.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR et al. (2007) PLINK: a tool set for wholegenome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
- Ramos AM, Crooijmans RPMA, Affara NA et al. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One 4: e6524.
- Tortereau F, Servin B, Frantz L.A.F et al (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content BMC Genomics 13:586 doi:10.1186/1471- 2164-13-586

Project N.6

Tracking the Regional Taurine Introgression in the Current Nelore Brazilian Population

Ana Maria Perez Obrien, from Vienna, Austria, to the Istituto di Zootecnica, Università Católica del Sacro Cuore, Piacenza, Italy

Purpose of the visit

This exchange visit was performed with the purpose of estimating admixture levels in the Brazilian Nelore population and with the intention of localizing genomic regions of predominantly taurine origin. Further on, there was interest in exploring the possible functional importance of discovered regions for the production and adaptation of this breed in the Brazilian environment.

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

During this visit to Piacenza - Italy, admixture levels and different population structure parameters were analyzed using high density (777k) Single Nucleotide Polymorphism (SNPs) genomic markers genotyped for a large array of individuals belonging to different cattle breeds. Initial

quality control of the genotypes was performed removing all non-autosomal SNPs, and considering SNP and individual genotyping rates for all genotypes corresponding to individuals of the same breed together. Later the genotypes for all breeds were merged and SNPs showing minimum minor-allele frequencies in the complete sample were removed. The estimations of the levels of autosomal admixture in the Brazilian Nelore population were performed with the ADMIXTURE software, using a wide array of different breeds as reference populations to explore the behaviour of the estimates when changing the number of individuals, the breeds included and to decide further on the breeds to be included for the final estimations. Other population genetic studies were performed to asess the levels of differentiation among the studied breeds including breed-pairwise Wright's-Fst and a Principal Component Analysis (PCA) based on the Genomic Kinship Matrix.

Apart from the Nelore population other Brazilian Cattle was also used for analyses, including three Creole breeds, Caracu, Curraleiro and Pantaneiro, which are candidate populations for the possible origin of the taurine introgression according to historical records, and Gir, another Indicine breed of current importance for milk production in the tropical areas of the country. Other breeds used for analyses included the European taurine breeds Angus, Holstein, Brown Swiss, Fleckvieh, Hereford, Piedmontese, Marchigiana, Chianina and Romagnola; the African taurine breed N'Dama; and two Pakistani indicine breeds Lohani and Tharparker. Inside the Nelore population, only young animals born after 2000 were kept for analysis to represent the current population and a division of the breed into two sets was performed, corresponding to a major division in the breeding programs into "pedigree" and "production" type. The pedigree type animals belong to programs using only breeding stock which is registered as pure by pedigree and with a strong focus on type traits such as conformation to breed standards. The production type set includes individuals bred in commercial programs, not registered as pure, and with a main focus of the breeding strategies in growth and other productive traits of importance in beef production. These two sets were compared to assess if there was a strong differentiation among these two types of cattle inside the breed, and whether larger levels of admixture existed in the commercial population as compared to the pedigree lines.

For the Nelore breed the levels of admixture were estimated for each autosomal chromosome as a first approach for the detection of regions with higher than expected taurine introgression. For the two chromosomes with the highest estimated admixture levels, regional admixture was explored using three approaches: 1) Allele frequency deviations, 2) SNP-wise Fst estimations, and 3) HAPMIX, a local admixture estimation software. For the first approach the allele with the highest frequency in the taurine populations was defined as the reference allele for analysis, and subsequently the frequency of this allele was estimated for the taurine reference, the pure indicine reference and the Nelore population. Using the estimated frequencies a search was performed for regions of continuous SNPs showing allele frequencies more similar to the taurine frequencies than to the indicine frequencies. In the second approach the single-SNP Fst values were explored to search for regions in the chromosome where continuos SNP's exhibited extreme Fst values, located in the highest 1 and 5% distribution of Fst values. For the third approach haplotypes were constructed for two reference populations composed of pure European taurine breeds and pure indicine breeds, and the genotypes of the Nelore population were used to estimate local levels of taurine admixture along the chromosome.

One more analysis was performed to evaluate the capacity of the mitochondrial SNPs included in the used SNPBeadChip to separate mitochondrial haplogroups. For this step ll mitochondrial SNPs were extracted and quality control performed to exclude SNPs and individuals with low genotyping rate and not exhibiting variation in the sample. These SNPs were then used to build haplotypes for each individual using the fastPHASE software and frequency of each haplotype for each breed was estimated.

Description of the main results obtained

Population Structure

The PCA results clustered the individuals inside each breed close together with minor mixings of individuals inside the Brazilian Creole populations, and clustering the European Angus, Holstein, Brown Swiss and Fleckvieh breeds very close together. The first component of the PCA explained the division into taurine and indicine (sub)species, while the second component divided the European and African taurine breeds, and the Nelore and Gir indicine breeds. The Fst results showed the lowest level of differentiation between Nelore and Gir, and the highest level of differentiation between the Hereford and the Brazilian Creole Curraleiro.

Admixture analyses performed using 2 to 4 assumed ancestries (K) separated the taurine and indicine ancestries, then the African taurine, and lastly the Nelore and Gir breeds were assigned different ancestries. Overall minor levels of taurine admixture were observed in Nelore and Gir with 0.5% average and 0.12% average European taurine ancestry respectively for each breed, and 0.23% in Nelore and 0.12% in Gir of estimated average African taurine ancestry. Indicine and African taurine ancestry was also observed in some taurine breeds, with the three Italian breeds exhibiting less than 10% indicine admixture, and an average 17% African taurine ancestry, and the Piedmontese breed showing 10% African taurine ancestry. The Brazilian Creole breeds showed a major component of European taurine ancestry, between 10 and 30% of Indicine ancestry and an average of 12% African ancestry, with the highest levels of indicine ancestry been observed in the Pantaneiro breed.

Local Admixture

The estimates for average chromosomal admixture levels found for the Nelore breed were systematically higher in the average sum of all chromosomes and for every single chromosome than those obtained using the complete set of genome-wide autosomal SNPs. Additional analysis removing SNPs in high correlation (i.e. exhibiting r^2 linkage disequilibrium levels higher than 0.2) were performed but the results obtained were still higher for the chromosome estimations as compared to the whole autosomal set. The results from the local ancestry approaches did not point out to specific regions of the genome having higher taurine ancestry. In general SNPs showing allele frequencies more similar to the taurine reference and SNPs showing high levels of Fst were randomly distributed across the analyzed chromosomes, and no specific regions with continuous SNPs showing either characteristic could be identified. The estimations obtained from the software used for local Admixture, HAPMIX, also did not point out to specific regions of higher taurine ancestry and in general the levels of admixture for the complete chromosome estimated through this method were higher than the levels obtained by using the ADMIXTURE and STRUCTURE software using the same SNP set (SNPs in each one of the two analyzed chromosomes).

Mitochondrial haplotypes

The analysis of the frequency of haplotypes indicated that the mitochondrial SNPs included in the SNP Beadchip used have low power to distinguish the known main cattle mitochondrial haplogroups and most

of the SNPs are removed through quality control as they show no variation between breeds. One of the haplotypes was found in common and exhibiting the highest frequency in all taurine populations, including the African taurine N'Dama, and the Nelore and Gir populations. The reference indicine groups used, including a set of Nelore imported from India, Tharparker and Lohani breeds, exhibited an almost completely fixed haplotype, which was not present in any of the taurine breeds, but two of the indicine animals were found to have the highest frequency haplotype in the taurine breeds. The indicine specific haplotype was found in the Nelore and Gir with a much lower frequency. These results indicate that the SNPs cannot differentiate the taurine African and European groups, and that they might be able to separate the major division of the haplogroups into taurine and indicine, but this needs to be confirmed by the use of mitochondrial sequence data to determine if the haplotypes found correspond to the specific haplogroups, and if the individuals from the indicine reference showing a different haplotype truly possess a taurine typical haplogroup.

Conclusions

The focus population of this study, the Nelore breed, showed very small levels of taurine introgression in the autosomal genome with less than 1% average taurine admixture, while the evidence from the mitochondrial haplotypes supported by historical records and molecular mitochondrial analysis reported in the literature, corroborate the introduction of taurine animals at some point in the history of the formation of the breed derived mainly from the use of taurine females. While our results from mitochondrial haplotypes are not conclusive, they support the taurine nature of the mitochondrial DNA in a large proportion of the young Nelore individuals, while the autosomal results indicate an almost exclusive indicine ancestry. Together both results indicate that even though there was taurine introgression in the Brazilian Nelore population, the breeding strategies adopted in Brazil, focused in the expansion of the indicine ancestry and "purification" of the breed, have reduced the taurine ancestry in the genome of this breed to minimum levels. The efforts in localizing regions of the genome in the Nelore that exhibit taurine origin were unsuccessful possibly due to the very low levels of taurine ancestry observed which can make the currently available approaches with SNP data unsuitable for this specific analysis. Most approaches to local admixture estimations up to date have been successfully applied in populations showing between 10 and 90% of one ancestry, mostly in populations of two known ancestries, which was not the case for this specific population**.**

Future collaborations

Further analysis of selected results of this investigation is currently being performed in collaboration with the host institution. This includes a deeper analysis of the admixture levels in the Brazilian Creole breeds which were only used as a reference population in the current study; given the higher proportions of taurine and indicine ancestry estimated in this breeds during the project, and the differences observed in the different breeds included, efforts to localize regions with high levels of african taurine and indicine introgression that are candidate selection signatures seem more promising for application in this breeds. Further investigation in Nelore is foreseen in the near future using full genome sequence data from Nelore and other breeds, in collaboration with the USDA and UNESP Brazilian partners. Monthly skype meetings have been taking place previous to the visit and have continued after the visit, were the current state of the project and results are discussed, and they include BOKU, Unicatt-Piacenza and UNESP collaborators, with ocassional participation of other

collaborators on specific topics. All efforts on the project will continue to be coordinated by the Brazilian partners from UNESP, who are working on increasing the data availability for the Brazilian creole breeds and leading the collection of sequence data. The final manuscript was drafted and submitted in collaboration with all previously mentioned partners, and results were discussed using e-mail correspondence.

Project publications resulting from the grant

Publication of the main results obtained, in the form of a short communication, has been submitted to GSE journal (Genetics Selection Evolution) and is currently under review. ESF was properly acknowledged in the submitted manuscript.

Other comments

The genotypes used for this study were provided by: United States Department of Agriculture – USA, Zebu Genomic Consortium - Brazil, EMBRAPA - Brazil, ZuchtData - Austria and Università Cattolica del Sacro Cuore – Italy.

I would like to thank the "European Science Foundation", and the "Advances in Farm Animal Genomic Resources" Research Networking program for the funding of this research grant and giving me the opportunity to expand my knowledge and scientific training. Thank you very much to Prof. Paolo Ajmone-Marsan and everybody in his working group at Università Cattolica del Sacro Cuore in Piacenza for all their help, kindness and support during this exchange visit. Finally, I would like to acknowledge and thank Tad Sonstegard, Curt VanTassell, Jose Fernando García, Marcos Vinicius B. da Silva, Yuri Tani Utsunomiya, and my supervisor Johann Sölkner, for their collaboration and support on the development of this research.

Project N.7

Genotyping by sequencing: new tool for population genomics of native and aquacultured Mediterranean mussel, Mytilus galloprovincialis

Anamaria Štambuk from Croatia visited the Patrik Nosil Lab at the University of Sheffield, UK

Purpose of the visit

Main purpose of my visit to Patrik Nosil lab at University of Sheffield (United Kingdom), was to implement genotyping by sequencing approach as suitable and versatile NGS tool into population genomics of aquacultured Mediterranean mussel Mytilus galloprovincialis, which hereafter becomes new species in the field of livestock/aquaculture genomics.

Description of the work carried out during the visit

During my two weeks visit (February 9th to February 23th) the Patrik Nosil lab on the University of Sheffield, we have used 200 milion reads generated by RAD-seq (Restriction site associated DNA sequencing) methodology on Illumina HiSeq sequencing platform to create pseudoreference Mytilus galloprovincialis consensus genome sequence, we have mapped barcoded RADseq Illumina sequences of native and aquacultured mussel individuals collected at the aquaculture sites, called SNPs and performed conversion of genotype likelihoods to genotype probabilities.

0. Preparation of samples (previously conducted): DNA isolation and RAD tag sequencing library preparation DNA isolation protocols were optimized in order to obtain DNA of the highest quality. Due to the high susceptibility of mussel DNA toward degradation, various extraction kits and protocols were optimized, using different mussel tissues and tissue conservation methods. On the end, DNA extracted from the muscle tissue preserved

in 96% EtOH using Sigma GenElute extraction kit, with homogenization step performed in liquid nitrogen and omitting vortexing was used.

Library preparation: In short, genomic DNA was digested with restriction enzymes (EcoRI and MseI). Custom made Illumina barcodes and adaptor sequences were ligated to the digested fragments. Only EcoRI restriction site was barcoded. These fragments were amplified by PCR using custom made Illumina primers. Amplified fragments ware polled into 30 uL of Illumina sequencing library and sent for sequencing. Sequencing was performed at NCGR (National Centre for Genomic Resources, USA) were fragments were size selected to the 300-500 bp, and sequenced in one lane of the Illumina NGS platform (HiSeq 2000 with V3 reagents).

Before starting the actual sequencing data analyses I received extensive few days bioinformatics training enabling me to work more efficiently in Unix environment and especially to successfully handle and process large amount of NGS data.

1. Quality control of raw data

Sequences were downloaded and checked for the raw data quality using FastQC tools. In total app. 200 million reads were generated out of one Illumina lane, and average quality of sequences was high.

2. Demultiplexing the data

Adaptor sequences and protector base were trimmed using custom written Perl scripts. Sequences that were too short, sequences without the barcodes or with irregular barcodes were discarded (3.1% sequences). 205 653 103 sequences were analyzed, 199 212 204 sequences contained barcodes and 6 440 865 sequences didn't contain barcodes.

104 762 sequences had a 3' MseI adapter sequence and 34 sequences were too short after removing the 3' adapter (<10 + 6 bp). 199 212 204 sequences were retained for further analyses.

3. Quality control

FastQC was run on the trimmed data. Average quality of reads was high (38), as was the overall quality, and GC content was 36%.

4. Splitting the reads

Custom made Perl scripts were used to identify and remove individual identifier (barcode) sequences from fastq files. All the sequences were assigned to the individuals and splited into separate bz2 files for each individual. Average number of reads per was 703930 bp per individual.

5. De novo assembly

Rainbow tools (Chang et al. 2012, Bioinformatics) were used to create pseudoreference Mytilus genome using Illumina RAD sequencing reads. Scripts were modified for one end sequencing protocols. We have arbitrary chose first 10 million reads and these were assembled into the contigs, yielding 459178 Mytilus contigs of length between 84-86 bp. Minimum overlap of sequences in contigs was 45 bp, and maximum 4 mismatches were allowed. Then, we made consensus seuqence by concating these contigs and used it to create first pseudoreference Mytilus genome by inclusion of arrays containing 30 N in-between two contigs. Total length of the pseudoreference was 52 805 760 bp. Afterwards, a second de novo assembly was performed using all 200 mil Mytilus reads generating pseudoreference genome of total length 472 757 920 bp.

6. Mapping the reads to the pseudoreference

Bowtie2 was used in combination with samtools to map all fastq sequence files to the pseudoreference Mytilus consensus genome assembled from 10 million reads, and sorted all the sequences in bam.bai files. Alignment sequences rate was around 95% for most of the individuals. This resulted in 199 212 204 sequences mapped and assembled onto the Mytilus pseudoreference genome, with the average

coverage depth of 438x per genomic region across all individuals.

7. SNPs calling

Samtools mpileup and bcftools incorporated in custom made Perl scripts were used to identify variable sites, and call single nucleotide polymorphisms (SNPs). We performed basic filtering, excluded all the insertions and deletion, all SNPs with the depth exceeding 10000, with the quality score below 20, and all the SNPs that were scored in less of 40% individuals. We have retained only biallelic SNPs. Final set of 180439 out of 343174 variable sites (SNPs) were retained after filtering.

8. Caculating genotype likelihoods and genotype probabilities

Custom made Perl scripts were used to convert phred-scaled genotype likelihoods to genotype probabilities using prior probability of each genotype for a given population allele frequencies.

Description of the main results obtained

The main purpose of this scientific exchange is fulfilled, as customized genomic tools have been developed for Mediterranean mussel, Mytilus galloprovincialis. The implementation of RAD-seq methodology and computational analytical tools in the populations genomics of aquacultured Mytilus galloprovincialis was really successful, yielding high quality genotype data and adding completely new species in the livestock genomics field.Although I am still processing data and did not reach the final step of having all population genomic results in hand, my two weeks visit to Patrik Nosil lab at the University of Sheffield already resulted in producing high quality genome wide genotyping data for Mytilus galloprovincialis. Currently, I am applying all the methods learned during my two weeks visit to address the questions of genomic differentiation between aquacultured and native mussels. Except introducing new species in the livestock genomic arena, to the best of our knowledge this is first application of GBS approach in the livestock genomics. We have shown that this methodology can be tailored to non model species to generate huge amount of genome wide SNP data at just small fraction of cost of the whole genome sequencing. In terms of genotyping by sequencing data analyses and especially in terms of acquiring new knowledge this visit to Patrik Nosil lab exceeded all my expectations.

Future collaboration with host institution

This collaboration is still ongoing, while research on aquacultured vs native mussels is active in terms of population differentiation data analyses.

Therefore, I am still learning many new methods in population genomics analyses within the frames of this collaboration.

Currently, I am also applying for the position of postdoctoral researcher on the UKF funded project led by dr. Nosil.

Projected publications / articles resulting or to result from the grant

Due to the high quality genoting data obtained, this collaboration is on good way to result in high quality scientific publication on the genomic differentiation between aquacultured and native population of Mediterranean mussel, Mytilus galloprovincialis. Naturally, ESF will be acknowledged for this travel grant in the resulting publication.

Other comments

I am truly greatful both to ESF and host laboratory of Dr. Patrik Nosil for the opportunity to gain valuable new analytical skills, and overall large amount of new knowlegde in the fast developing and exciting research field of population genomics.

My travel and life expenses that exceeded the amount limits of the ESF travel grant specific categories were covered from the UKF funded project led by Dr. Patrik Nosil.

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
- Institute 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 Sweden, 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 | Chantal Durant, 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