ESF Exchange Grant: "Detection of signatures of selection in wild and domestic livestock populations"

Filippo Biscarini[1*]

1. PTP (Parco Tecnologico Padano), Via Einstein - Loc. Cascina Codazza, 26900 Lodi (Italy)
*Contact author: filippo.biscarini@gmail.com

April 15, 2013

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.

1 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).

2 Description of the work carried out during the visit

2.1 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 focussed on the 16 samples from the Iranian pilot study to set up the methodology and workflow.

2.2 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 Cardiff 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.

2.3 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.

3 Description of the main results obtained

3.1 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 mouflon populations

3.2 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 different 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:

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

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:

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

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 iHS 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

3.3 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 ($p \approx q \approx 0.5$) 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.

3.4 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 different 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 different 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 differential 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

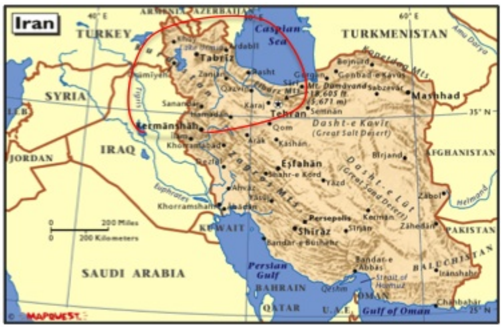
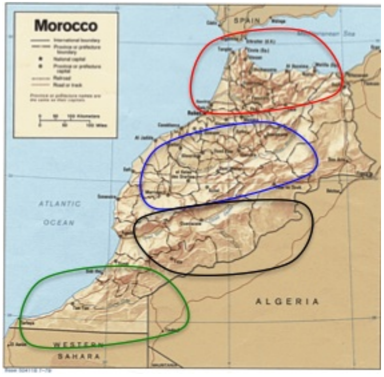


Figure 2: Different climatic regions of Morocco (coast, northern Atlas, southern Atlas, desert) from where domestic sheep and goats were sampled



(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.

3.5 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 ($\approx 2GB$) 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 22^{nd}.

4 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.

5 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.

References

Nextgen project. URL http://nextgen.epfl.ch/.

Sheep genome v. 3.1. URL http://www.ncbi.nlm.nih.gov/assembly/GCA_000298735.1/.

F. Biscarini, M. Del Corvo, A. Albera, P. Boettcher, and A. Stella. A web-interface for the detection of selection signatures: development and testing. In *Proceedings of the 64th EAAP annual meeting, Bratislava (Slovakia)*. EAAP, 2012.

J. W. Kijas, J. A. Lenstra, B. Hayes, S. Boitard, L. R. P. Neto, M. San Cristobal, B. Servin, R. McCulloch, V. Whan, K. Gietzen, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS biology*, 10(2):e1001258, 2012.

P. OROZCO-terWENGEL, M. Kapun, V. Nolte, R. Kofler, T. Flatt, and C. SCHLÖTTERER. Adaptation of drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular ecology*, 21(20):4931–4941, 2012.

C.-J. Rubin, M. C. Zody, J. Eriksson, J. R. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, 464(7288):587–591, 2010.

P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

A. Stella, P. Ajmone-Marsan, B. Lazzari, and P. Boettcher. Identification of selection signatures in cattle breeds selected for dairy production. *Genetics*, 185(4):1451–1461, 2010.

B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS biology*, 4(3):e72, 2006.