Research Networking Programmes

Short Visit Grant ☐ or Exchange Visit Grant ☒

(please tick the relevant box)

Scientific Report

The scientific report (WORD or PDF file – maximum of eight A4 pages) should be submitted online within one month of the event. It will be published on the ESF website.

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

Application Reference N°: 4577

1) 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 increasinglybeing 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 preselected, 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.

2) 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.

3) 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.

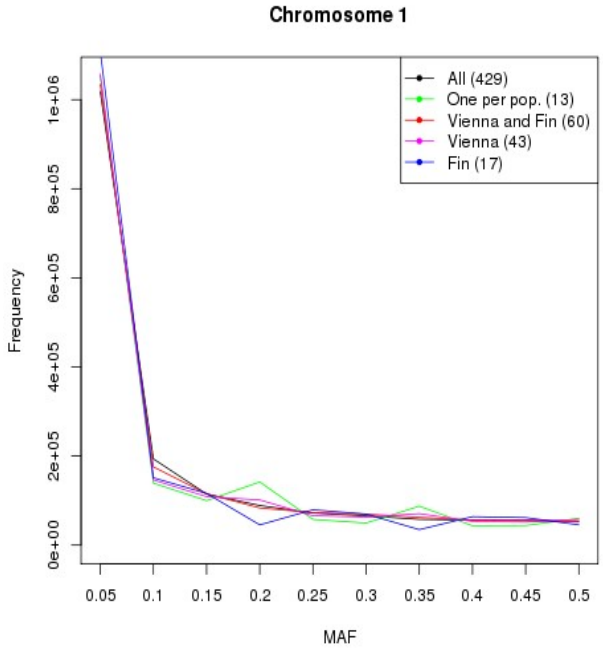

Figure 1:Minor allele frequence histogram. Histogram contains MAF for more populations than were used in this project, which can be ignored for this report. Vienna refers to BSW and Fin to FAY.

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. 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 subsetted the NGS based SNP calls based on SNPs that are genotyped on either of the commerical 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.

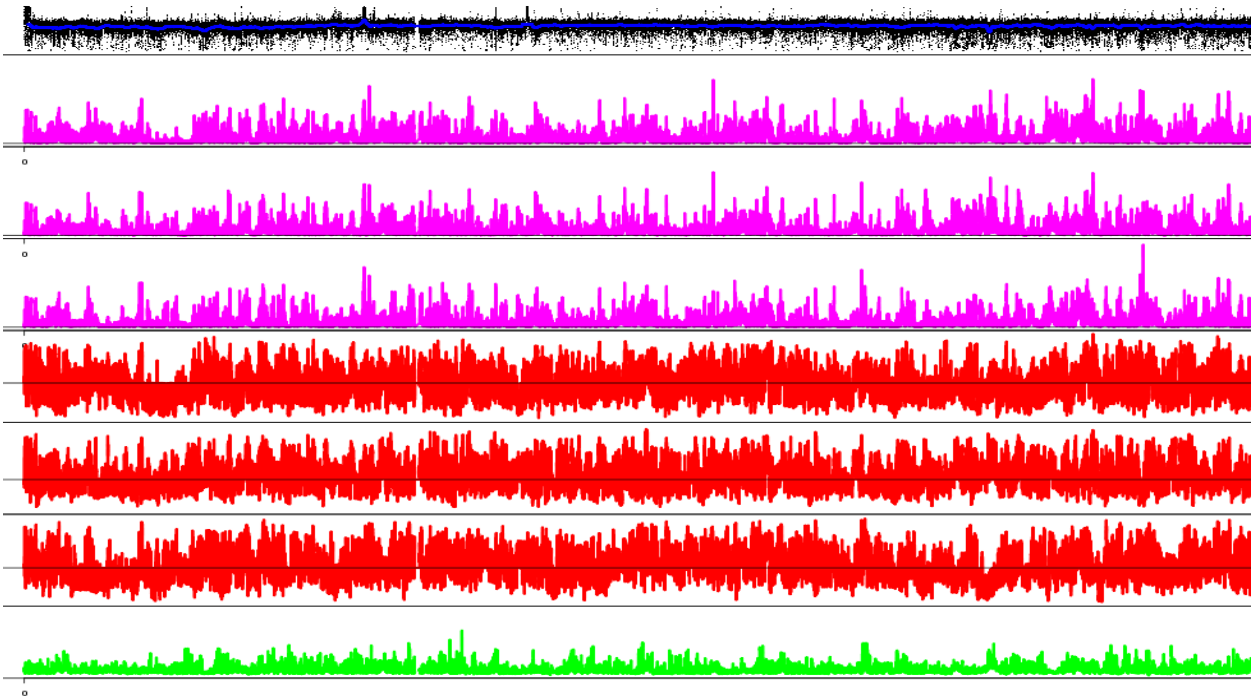

Figure 2:Visualization of (smoothed) read depth (first line), Pi (Line2: Joined populations, Line3: Brown Swiss, Line4: Finnish Ayrshire) and Tajima D

(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 in case of the CLR, 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.

4) Future collaboration with host institution (if applicable)

Please, see point 5

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

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.

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.

6) **Other comments (if any)**

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): 585–95.