Final Report

European Science Foundation (ESF) - GENOMIC RESOURCES

Exchange grant no. 4118

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

Sylvie Stucki
Laboratory of Geographic information systems
School of Architecture, Civil and Environmental Engineering
École polytechnique fédérale de Lausanne

Host:
Professor Michael W. Bruford
Division Organisms and Environment
Cardiff School of Biosciences
Cardiff University

11th June 2013

1 Purpose of the visit

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

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

2 Description of the work carried out during the visit

2.1 Pre-processing

Both datasets were filtered with a call rate of 95% for SNPs and individuals, the minimum allele frequencies (M. A. F.) were set to 1% for the first group and 5% for the second, resulting in 804 samples genotyped for 41,215 SNPs and 102 samples genotyped for 634,849 SNPs respectively (Purcell *et al.*, 2007).

The environment was characterised with the WorldClim dataset, which consists of minimum, maximum and mean monthly temperatures, monthly amount of precipitation and 19 derived variables at a resolution of 1 kilometre (Hijmans *et al.*, 2005). The topography was described with the digital elevation model SRTM which mesh is 90 meters (Farr *et al.*, 2007). The slope and curvature were derived from the altitude. Environmental data was prepared with SAGA GIS (www.saga-gis.org) and the values corresponding to the sampling locations were extracted in Quantum GIS (www.qgis.org). A total of 72 environmental variables were included in the analysis.

2.2 Population structure

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

2.3 Landscape genomics

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

The study also allowed for testing and improving three features of Samβada:

- Multivariate models are assessed against the simpler nested models, so the gain in prediction is worth the added complexity.

- The module for spatial autocorrelation was completed with the computation of significance levels for local Moran's I, a Local Indicator of Spatial Association (Anselin, 1995).

- The interface between Samβada and other bioinformatic softwares was improved by providing a module to recode PED and MAP files, a popular format for SNP data (Purcell, 2009), into input files for Samβada.

3 Description of the main results obtained

3.1 Population structure

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

3.2 Landscape genomics

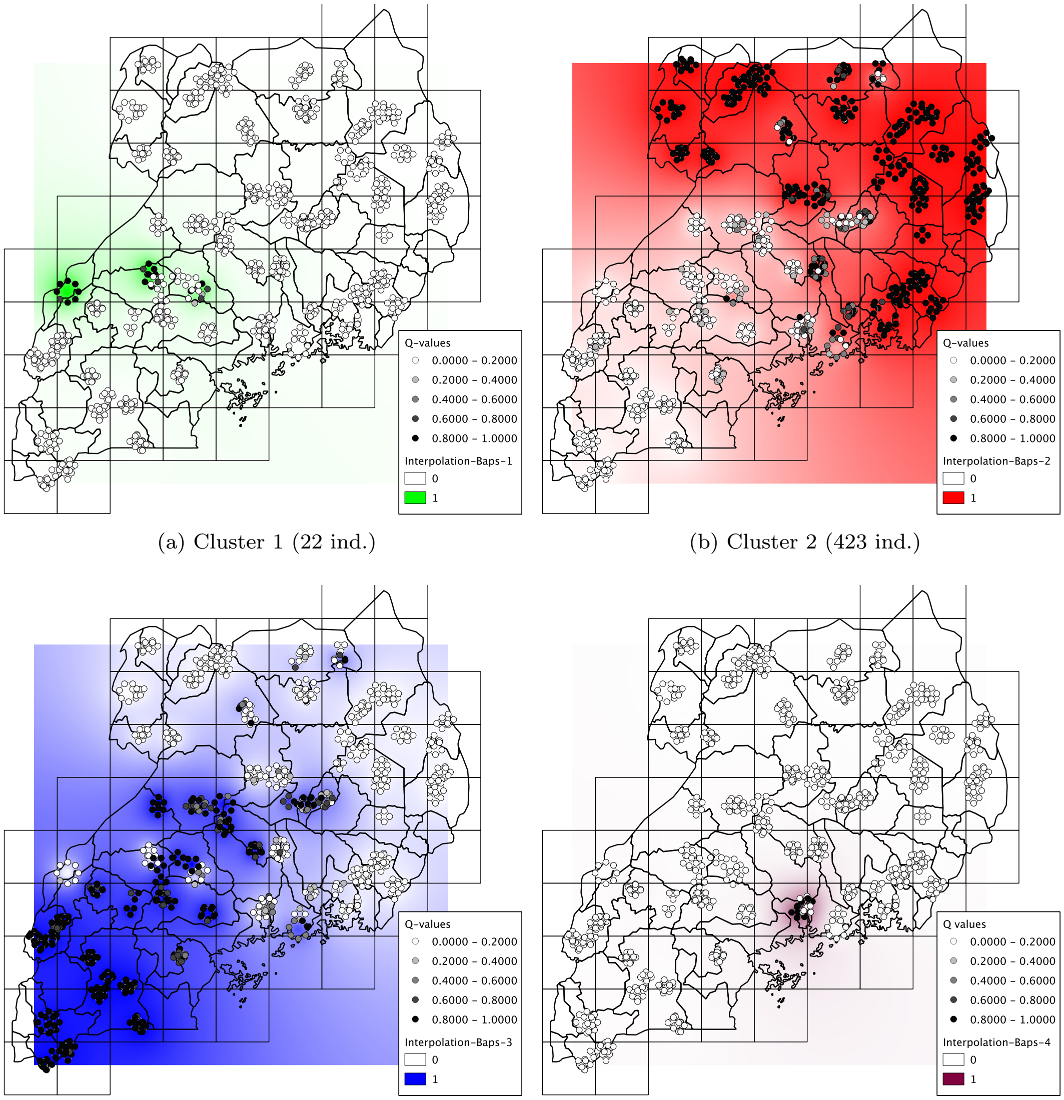
3.2.1 Detection of loci under selection

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

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

Table 1: Most significant models with 804 individuals and 41,215 SNPs (=120,869 binary markers). The first column is the marker name, formed by the SNP name and the allele, then come the name of the environmental variable, the log likelihood-ratio (G) score, and the Wald score. The last column shows the sign of the correlation between the marker frequency and the value of the environmental variable. These loci are located on the X chromosome.

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



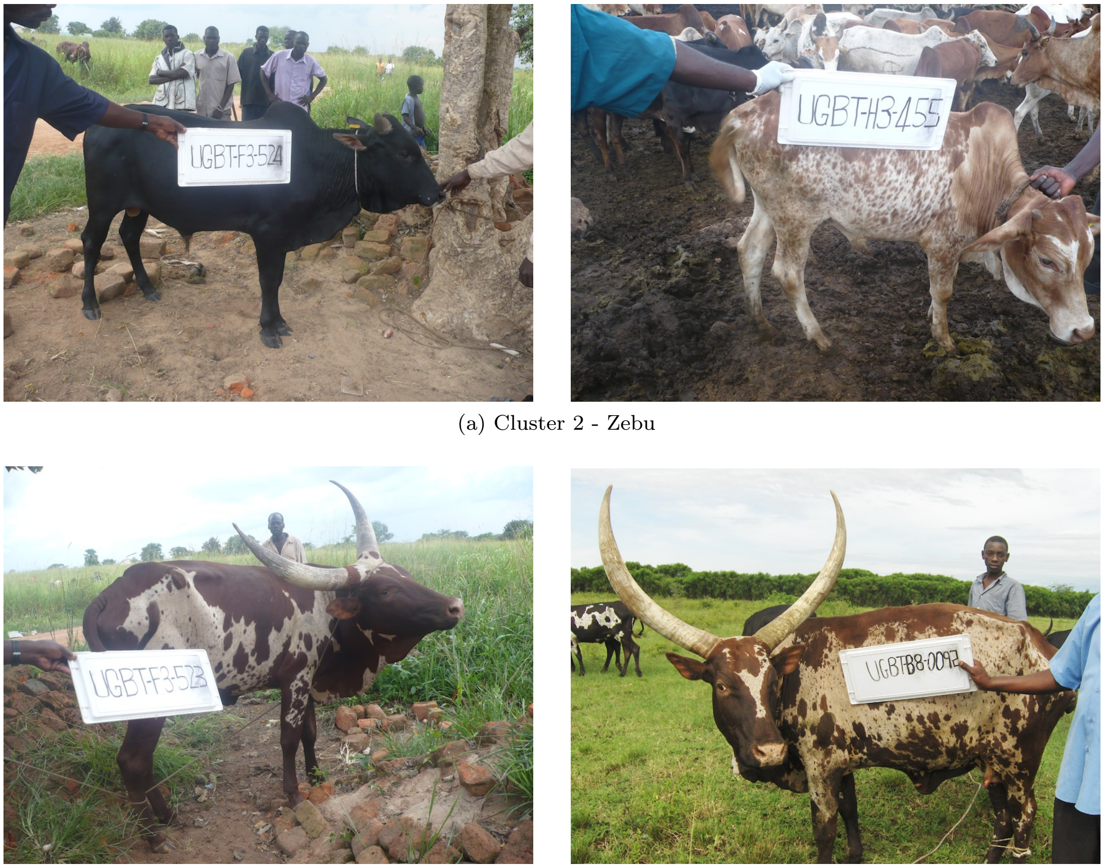
(a) Cluster 1 (22 ind.)

(b) Cluster 2 (423 ind.)

(c) Cluster 3 (330 ind.)

(d) Cluster 4 (11 ind.)

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



(a) Cluster 2 - Zebu

(b) Cluster 3 - Ankole

Figure 2: Pictures taken during the sampling. The two largest clusters correspond to the Zebu and Ankole populations.

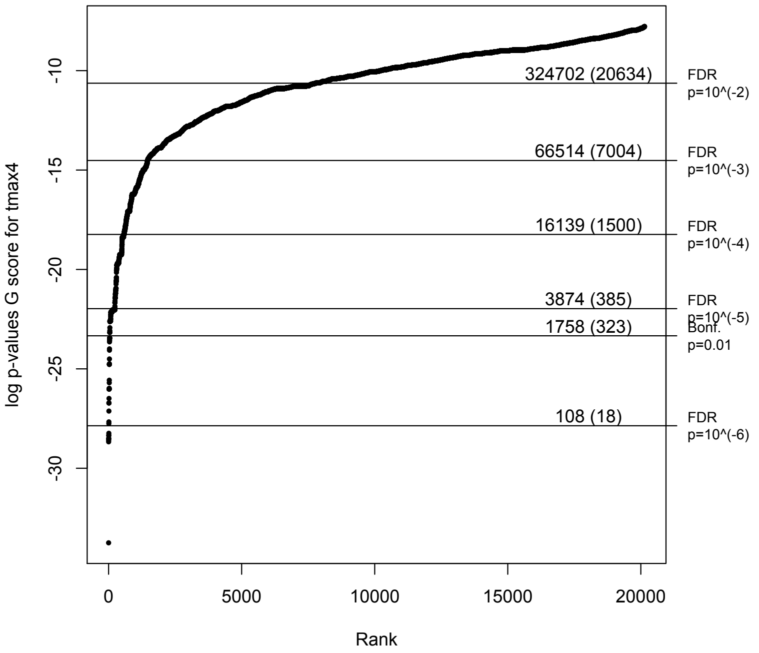


Figure 3: Distribution of p-values for regression models with maximum temperature in April. Each horizontal line shows a possible threshold, either using Bonferroni (Bonf.) correction or False Discovery Rate (FDR). The labels indicate the type of correction and the p-value for each level, along with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

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

Figure 3 shows the distribution of the log p-values of G scores for all models involving the maximum temperature in April. This variable was commonly found as an accurate predictor for marker distributions. Significance threshold was set at p=0.01 with Bonferonni correction, which lead to 1,758 significant models involving 323 SNPs. These loci were spread between chromosomes 5 (42 SNPs), 14 (4 SNPs) and X (277 SNPs). Latitude was often highly correlated with marker frequencies.

3.2.2 Gene mapping

The distributions of the significant loci on chromosome 5 are shown on fig. 4 for several thresholds. The most significant model involves the SNP BovineHD0500019261, this loci maps to the gene CHST11 which is involved in cartilage make up (Flicek *et al.*, 2012). The second cluster detected on chromosome 5 maps to an uncharacterised gene ENSBTAG00000033726 while the most significant SNP on chromosome X (BovineHD3000015663) is located near a conserved genomic region in 36 eutherian mammals.

3.2.3 Spatial autocorrelation

Logistic regression fit a global model for the presence of a marker, while spatial analysis provides information about local behaviours. Local Indicators of Spatial Association (LISA, Anselin, 1995) compare the value of a variable in each location with the weighted mean of its values in the neighbouring points. LISA are the local equivalent of spatial autocorrelation. A bivariate LISA compares the value of a variable to the mean of another variable over the neighbouring points. Fig. 5 shows

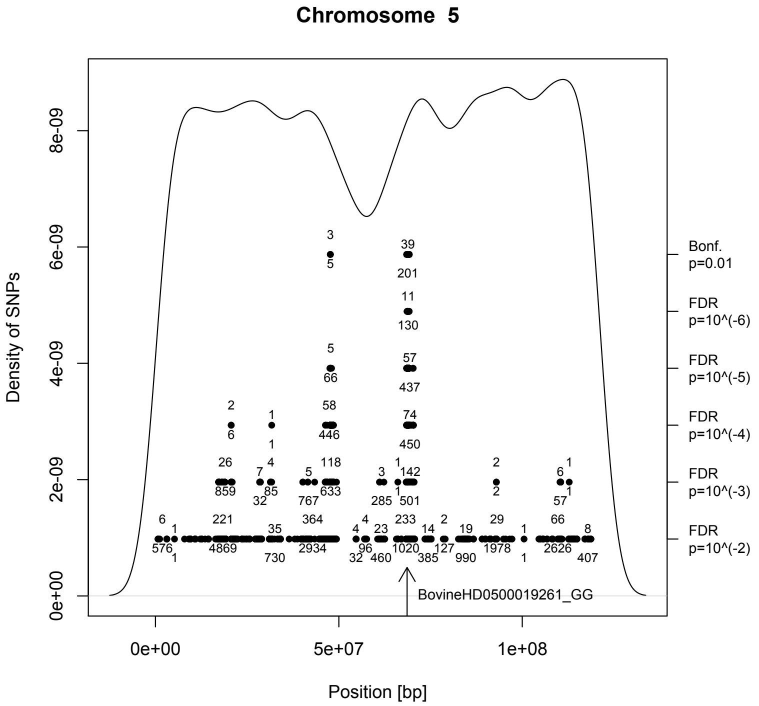


Figure 4: Solid line shows the overall SNPs density on chromosome 5. Horizontal plots represents the SNPs that were detected for different thresholds. These SNPs were grouped when they were closer than $2 \cdot 10^6$ bp. Each cluster is summarized by the number of SNPs it spans (below) and among these, the number of SNPs under selection (above). The vertical spacing between plots is arbitrary. The arrow points out the SNP BovineHD0500019261.

local correlation between the presence of the marker BovineHD0500019261_GG (allele GG) and the maximum temperature in April, computed with local Moran's I (LISA, Anselin, 1995). The map shows a positive correlation in North and South Uganda separated by a non-significant region.

3.3 Summary

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

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

The following analysis are ongoing to test these hypotheses:

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

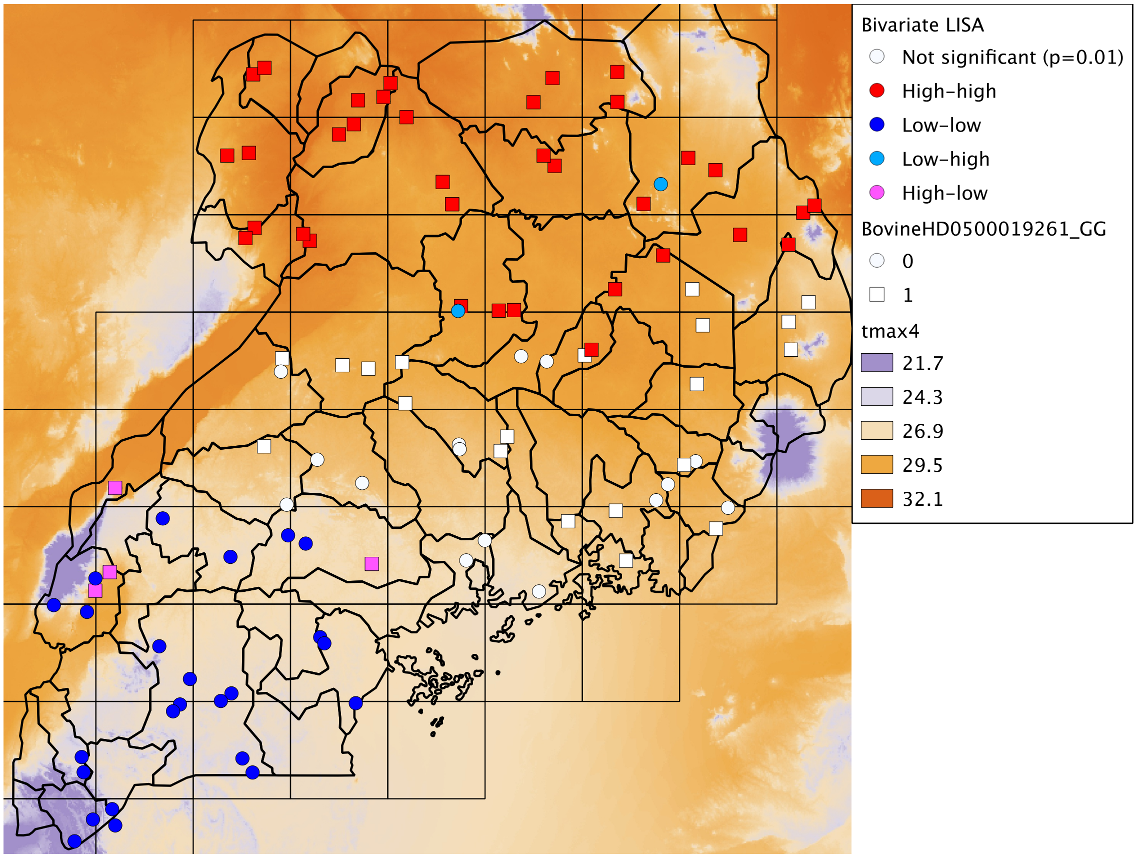


Figure 5: Figure 3: Bivariate local Moran's I between BovineHD0500019261_GG and the maximum temperature in April (background layer) for the 102 Ugandan cattle. This indicator measures the spatial correlation between the state of the marker and the temperature averaged over the 20 nearest sampling points.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

- Comparison of breeds and markers distributions to parasites prevalence, especially the tsetse fly, to test whether they overlap.

4 Future collaboration with host institution

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

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

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

6 Other comments

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

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**(9), 1655–1664.

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, **27**(2), 93–115. GISDATA (Geographic Information Systems Data) Specialist Meeting on GIS (Geographic Information Systems) and Spatial Analysis, Amsterdam, Netherlands, Dec 01-05, 1993.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **57**(1), 289–300.

Corander, J. and Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, **15**(10), 2833–2843.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, **45**(2).

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. J. (2012). Ensembl 2012. *Nucleic Acids Research*, **40**(D1), D84–D90.

Hijmans, R., Cameron, S., Parra, J., Jones, P., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal Of Climatology*, **25**(15), 1965–1978.

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

Purcell, S. (2009). Plink 1.07. `http://pngu.mgh.harvard.edu/purcell/plink/`.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**(3), 559 – 575.