Samβada in Uganda: landscape genomics study of traditional cattle breeds with a large SNP dataset

May 3, 2013

Introduction

Since its introduction [9], landscape genomics has developed quickly with the increasing availability of both molecular and topo-climatic data. The current challenges involve processing large numbers of models and disentangling selection from demography. Several methods address the latter, either by estimating a neutral model from population structure [3] or by inferring simultaneously environmental and demographic effects [6]. Here we present Samβada, an integrated software for landscape genomic analysis of large datasets. This tool was developed in the framework of NextGen with the objective of characterising traditional Ugandan cattle breeds using single nucleotide polymorphisms (SNPs) data.

Methods

Samβada uses logistic regressions to estimate the probability that an individual carries a specific genetic marker given the habitat that characterises its sampling site [8]. The genetic data is recoded as binary variables and their association to the topo-climatic data is assessed with log-likelihood ratio (G) and/or Wald tests [4]. Models are ranked according to their scores to ease post-processing analyses.

Large SNP panels and whole-genome sequences often require sharing the computational load. When requested, Samβada splits the molecular data to distribute processing and merges the results subsequently .

While global regression models assess the overall relationships in the data, spatial patterns of associations give information about local processes at work. Samβada can measure the level of spatial autocorrelation in both molecular and environmental datasets using local and global Moran's I [1].

Data

Blood and skin samples were collected form 102 Ugandan cattle along with their geographic coordinates. The samples were genotyped with the 800k BovineHD assays (Illumina Inc., San Diego, USA)), rendering 2.113.358 binary markers for analysis. The environment was described with 73 variables: monthly values of temperature and precipitation from WorldClim [7], and slope and aspect derived from the digital elevation model STRM3 [5].

Results

Models were assessed according to their G score, 1549 significant associations involving 323 loci were found ($p = 0.01$ before Bonferroni). Fig. 1 shows the distribution of p-values for models involving maximum temperature in April, a variable commonly found to predict allele frequencies. Most associations were found in chromosomes 5, 14, 20 and X. The most significant model involves the SNP BovineHD0500019261 on chromosome 5 (Fig. 2). A bivariate LISA map presents the spatial association between this marker and the mean temperature in April (Fig. 3).

Discussion

High-density SNP assays allow detecting genomic regions potentially involved in local adaptation. In our study, loci under selection are associated with latitude, and the most relevant local correlations were found in Uganda's North and South. This might indicate a demographic effect since cattle breeds differ between these regions, but it may also reflect local adaptation as many environmental parameters are correlated with latitude. The SNP BovineHD0500019261 maps to the gene CHST11 which is involved in cartilage make up.

Our study shows that landscape genomics can handle large molecular datasets. However the sampling size is critical (n=102) to assess model significance. Bonferroni correction might be too conservative for whole-genome sequencing and alternative approaches such as False Discovery Rate might be considered.

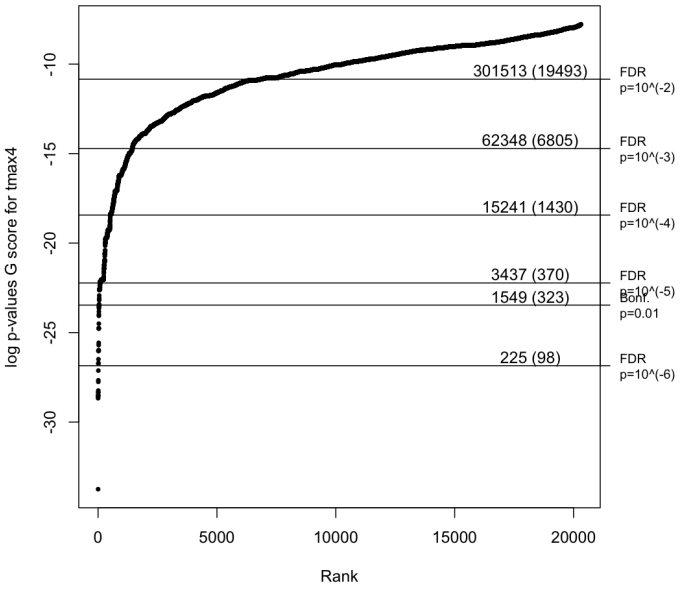

Figure 1: Distribution of p-values for regression models with maximum temperature in April. Each horizontal line shows a possible threshold, eigher using Bonferroni correction or False Discovery Rate [2]. The labels indicate how many models are significant at this level, the number of associated SNPs are in parenthesis.

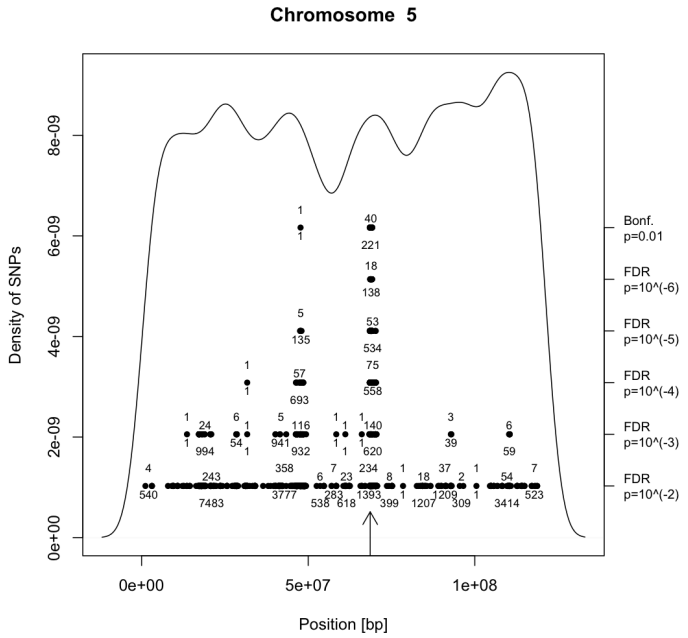

Figure 2: Solid line shows the overall SNPs density on chromosome 5. Horizontal plots represents the SNPs that were detected for different thresholds. These SNPs were grouped when they were closer than $2 \cdot 10^6$ bp. Each cluster is summarized by the number of SNPs it spans (below) and among these,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

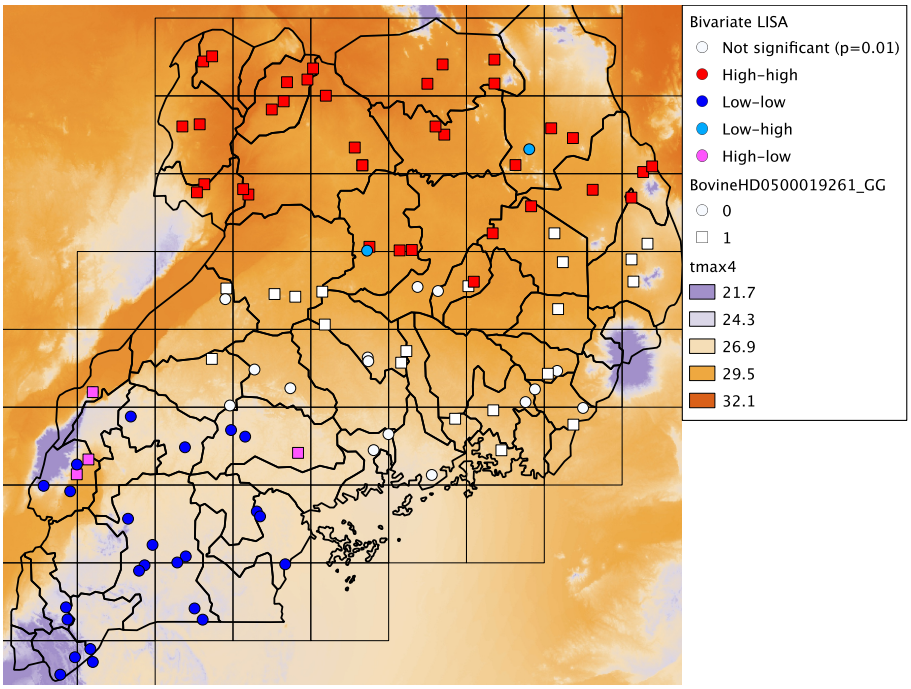

Figure 3: Bivariate local Moran's I between BovineHD0500019261_GG and the mean temperature in April (background layer) for the 102 Ugandan cattle. This indicator measures the spatial correlation between the state of the marker and the temperature averaged over the 20 nearest sampling points. Dots shape 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.

References

[1] 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.

[2] 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.

[3] Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**(4), 1411–1423.

[4] Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman & Hall, 3rd edition.

[5] 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).

[6] Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*.

[7] 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.

[8] 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.

[9] Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power of population genomics: from genotyping to genome typing. *Trends in Ecology and Evolution*, **4**(12), 981–994.