Detecting selection in population trees: an extension of the Lewontin and Krakauer test with an application to pig.

M. Bonhomme[*], C. Chevalet[*], B. Servin[*], S. Boitard[*], J. Abdallah[†], S. Blott[‡] and *M. SanCristobal*[*]

Introduction

The development of methods aiming at detecting molecular signatures of selection is one of the major concerns of modern population genetics. Broadly, such methods can be classified into four groups: methods focusing on (i) the interspecific comparison of gene substitution patterns, (ii) the frequency spectrum and models of selective sweeps, (iii) linkage disequilibrium (LD) and haplotype structure, and (iv) patterns of genetic differentiation among populations (for a review see Nielsen (2005)).

One approach of detecting loci under selection (outliers) with population genetic data is based on the genetic differentiation between loci only influenced by neutral processes (genetic drift, mutation, migration) and loci influenced by selection. Lewontin and Krakauer's test - LK test - for the heterogeneity of the inbreeding coefficient (F) across loci was the first to be developed with regard to this concept (Lewontin and Krakauer, 1973). The neutral model was a star-like evolution of populations with the same pattern of evolution. This has led to some criticisms in the literature, as well as improvements such as more sophisticated neutral models (Excoffier et al., 2009), conditional distribution on heterozygosity (Beaumont and Nichols, 1996), Bayesian models with McMC estimation (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008).

The key point is a good specification of the neutral model under which the null hypothesis will be based. It should be as close as possible to the real evolutionary history of the set of populations under study. In livestock, the divergence time between breeds in intraspecific studies is small, the history has a complex tree-like pattern, with negligible migrations between breeds. We propose here to take account of this kind of complex demographic history specific to livestock in a new test inspired from the original LK approach. Its application is very fast, making it a method of choice for massive SNP data.

[*] UMR444 Laboratoire de génétique cellulaire, INRA Toulouse, BP52627, F-31326 Castanet Tolosan cedex, France

[†] Department of Animal Production, Faculty of Agriculture, An-Najah National University, Nablus, P.O. Box 7, Palestine

[‡] Centre for Preventive Medicine, Animal Health Trust, Lanwades Park, Kentford, Newmarket, Suffolk, UK, CB8 7UU

Material and methods

The LK test. Consider L biallelic loci genotyped for a large set of individuals structured in n populations. Lewontin and Krakauer (1973) focused on the distribution of the F_{ST} statistic per locus, and proposed a test statistic denoted here by T_{LK}. Let $p = (p_1, ..., p_j, ..., p_n)'$ be the n-vector of allelic frequencies of the first allele (say) in the n populations. The quantity F_{ST} is defined as

$$F_{ST} = \frac{\frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})^2}{\bar{p}(1-\bar{p})} \quad (1)$$

where $\bar{p}$ is the sample mean of the vector p. The test statistic is equal to $T_{LK} = (n - 1)F_{ST}/\bar{F}_{ST}$ where $\bar{F}_{ST}$ is the average of F_{ST} in (1) over the L loci.

The F-LK test: taking account of the tree-like short-term evolutionary history. Under genetic drift, the first 2 moments of p are $E(p) = p_0 1_n$ and

$$Var(p) = Fp_0(1 - p_0), \quad (2)$$

where p_0 is the founder allele frequency, 1_n the n-vector of 1's, and F is the kinship (or coancestry) $(n \times n)$ - matrix linking the n populations.

A natural extension of the LK test is to take account of the tree-like short-term evolution of the populations, via the F matrix in the variance of allele frequencies. Let $\hat{p}_0$ be the linear estimate of p_0 with minimum variance, and 1_n be the n-vector made of 1's. Then the test statistics is:

$$T_{F-LK} = (p - \hat{p}_0 1_n)'\hat{V}ar(p)^{-1}(p - \hat{p}_0 1_n). \quad (3)$$

It can be shown that T_{F-LK} follows approximately a χ^2_{n-1} distribution under genetic drift.

Simulations. We simulated haplotype samples of partially linked loci, under neutrality (H_0) and directional selection on one locus in one population (H_1). The populations originated from an equilibrium ancestral population of constant size. The generated haplotypes consisted of 1,000 SNPs (or biallelic segregating sites) equally distributed along a 100 Mb chromosome.

Selection was modelled as follow: selection occurs on a single locus (SNP) of the haplotype, on the less frequent allele of the SNP. The fitness of the ancestral (resp. derived) allele is proportional to 1 (resp. 1+s).

Data. A small dataset was tested for signature of selection: 34 SNPs located in candidate genes (Blott et al. (2003) and Blott et al. in preparation). Samples of 4 major European pig breeds were genotyped: the Landrace (LR), the Large White (LW), the Piétrain (PI) and the Duroc (DU). To estimate the F matrix for calculating the T_{F-LK} statistic, we made use of a phylogenetic approach with an Asian breed as outgroup, and 50 genome-wide distributed microsatellite markers previously studied on the same samples in the PigBioDiv project (http://www.projects.roslin.ac.uk/pigbiodiv/, (SanCristobal et al., 2006)).

Results and discussion

Computer simulations showed similar results for both tests in a star-like topology. The extended test T_{F-LK} clearly outperformed T_{LK} in a complex tree-like topology (Figure 1) when selection occurred in a large population, in terms of power and false discovery rate. When selection happened in a small population, T_{LK} was slightly more powerful than T_{F-LK}. By construction, T_{F-LK} gives more weight to the largest populations.

On the real data set, only the extended test detected outliers (Figure 2). This is in agreement with the simulation study and with the prior expectation that selection acted on some of the considered genes, since they are linked with important biological functions with contrasted patterns between breeds.

The application of the test on a genome wide 60K SNP data set in sheep (http://www.sheephapmap.org/) took a few seconds and revealed coherent genomic regions that have probably been the target of selection (not shown).

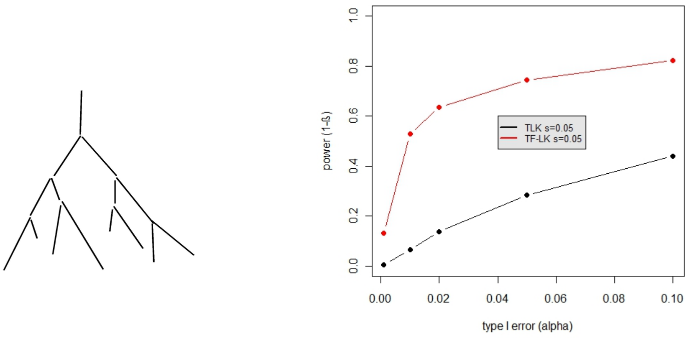

Figure 1: Example of a tree-like divergence between 8 populations (left) with selection occurring in a large population (small branch length). In this case, the extended test outperformed the LK test (right) in terms of power in a simulation study.

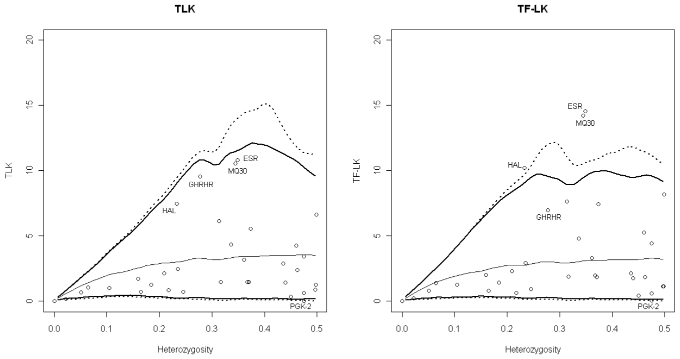

Figure 2: In a real pig data set, the LK test (left) detected no selection, while le F-LK test did (right). These tests are here conditioned on the level of heterozygosity. The 95% (solid line) and 98% (dotted line) envelopes are plotted.

Conclusion

The test developed here for detecting signatures of selection takes account of the evolutionary history of considered livestock populations (e.g. pig breeds), hence relying under a more realistic null hypothesis. This avoids false positives and increases power dramatically, when selection occurred in a large population. This approach is a step forward to more powerful tests suited for livestock breeds.

Acknowledgements. The authors would like to thank the DNA providers of the European project PigBioDiv (Bio4-CT-0188). JA acknowledges financial support from PigBioDiv2 (QLK5-CT-2002-01059) and MB from DéLiSus (ANR-07-GANI-001).

References

Beaumont, M. and Balding, D. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13(4):969–980.

Beaumont, M. and Nichols, R. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 263(1377):1619–1626.

Blott, S., Andersson, L., Groenen, M., San Cristobal, M., Chevalet, C., Cardellino, R., Li, N., Huang, L., Li, K., Plastow, G., and C., H. (2003). Characterisation of genetic variation in the pig breeds of china and europe - the pigbiodiv2 project. *Archivos de Zootecnia*, 52(198):207–217.

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, 103(4):285–298.

Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics*, 180(2):977–993.

Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74:175–195.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu Rev Genet*, 39:197–218.

SanCristobal, M., Chevalet, C., Haley, C., and et al. (2006). Genetic diversity within and between european pig breeds using microsatellite markers. *Animal Genetics*, 37:189–198.