Local Score Based Method Applied On Pool-Sequenced Behavior-Divergent Lines Precisely Detected Selection Signatures Related To Autism In Quail.

M.I. Fariello,[1,2] S. Boitard,[3,4] S. Mercier,[5,6] D. Robelin,[2] T. Faraut,[2] C. Arnould,[7] E. Lebihan,[8] J. Recoquillay,[8] G. Salin,[2,9] P. Dehais,[10] F. Pitel,[1] C. Leterrier,[7] M. SanCristobal[2,6,11]

[1]Institut Pasteur and Universidad de la República, Montevideo, Uruguay, [2]INRA, Université de Toulouse, INPT, ENSAT, UMR1388 GenPhySE, France, [3]INRA, AgroParisTech, GABI, Jouy-en-Josas, France, [4]MNHN / CNRS / EPHE / UPMC, Paris, France, [5]Université Toulouse Le Mirail, France, [6]Institut de Mathématiques de Toulouse, France, [7]INRA – CNRS – Université de Tours, France, [8]INRA, UR83 Recherches Avicoles, Nouzilly, France, [9]INRA, GeT, Castanet-Tolosan, France, [10]INRA Toulouse, SIGENAE, France, [11]INSA Toulouse, GMM, France

ABSTRACT: Detecting genomic footprints of selection is an important step in the understanding of evolution. Accounting for haplotype information in genome scans for selection allows increasing the detection power, but haplotype-based methods require individual genotypes and are not applicable when only allele frequencies are available. We propose here to take advantage of the local score approach to accumulate (possibly small) signals from single markers over a genomic segment, to clearly pinpoint a selection signal. This method gave results similar to haplotype-based methods on benchmark data sets with individual genotypes. Results obtained for a divergent selection experiment on behavior in quail, where two lines were sequenced in pools, are precise and biologically coherent. This local score approach is general and can be applied to other genome-wide analyzes such as GWAS or genome scans for selection.

Keywords: footprint of selection; local score; quail; NGS; pool sequencing.

Introduction

Detecting genomic regions that have evolved under selection has received much interest these last years. Linkage disequilibrium (LD) leads to the persistence of a footprint of selection around positively selected mutations, so the selection signature is not limited to a single causal mutation, but generally extends to a wider genomic interval including this mutation. Consequently, detection power is expected to increase when searching for such intervals rather than considering markers independently of each other. Haplotypic methods (e.g. Voight et al. (2006), Sabeti et al. (2007), Fariello et al. (2013)) require at least individual genotype data and are rather computationally demanding. On the other hand, single marker statistics have a lot of variability. Sliding window approaches do not require individual genotypes and run faster than haplotype-based methods. However, they are also less powerful and imply to choose a window size, which is usually done arbitrarily. To overcome this problem, alternative approaches to find clusters of high Fst values were proposed by Myles et al. (2008) and Johansson et al. (2010).

The objective of this study is to present a new strategy for detecting footprints of selection, which runs very fast and is suited to the case where only allele frequencies are available. We show that this novel approach performs well by detecting genes associated to social reinstatement behavior in quail, using unpublished pooled NGS data from two quail lines that have been divergently selected for this trait (Mills and Faure (1991)). This strategy aims at accumulating selection signals via small p-values of single-marker tests (or equivalently large values of -log10(p-value), further defined as a score) in an automatic manner, using the statistical theory of local scores (e.g. Mercier et al. (2003)). Usually the local score approach plays an important role in bio-informatics, where it is used for computing sequence alignment scores (Karlin and Altschul (1990)). In addition it can be used to extract sequence segments of unknown length, which are exceptionally interesting due to physico-chemical characteristics (e.g. exceptional hydrophobic segments for transmembrane proteins).

Materials and Methods

HapMap Data. This data set was used as a benchmark. We tested a 4Mb region (134-138 Mb) on Human chromosome 2 containing the LCT gene, because a known causal mutation for the lactase persistent phenotype in the CEU population is located in chromosome 2 at position 136,325,116. Data was taken from the HapMap Phase III dataset and consisted in the genotypes of 370 founder individuals from the CEU, TSI, CHB and JPT populations. Only 25 % of the available SNPs (that is 497 SNPs) were included in the analysis.

Quail Data. Two divergent lines produced and maintained at the INRA experimental unit 1295 (UE PEAT, F-37380 Nouzilly, France) were used in the experiment. These lines show high (HSR) and low (LSR) social reinstatement behavior (e.g. Mills and Faure (1991), Schweitzer et al. (2011)). Ten individuals from generation 50 of each quail line were pooled and sequenced (paired-ends, 100 bp) on a HiSeq 2000 sequencer (Illumina). In the absence of an available genome sequence for the quail, the reads of the two divergent lines were mapped to the chicken genome assembly (GallusWU2.58). Allele frequencies and then F-LK values (i.e. Fst corrected for population structure, Bonhomme et al. (2010)) were finally computed at all SNPs.

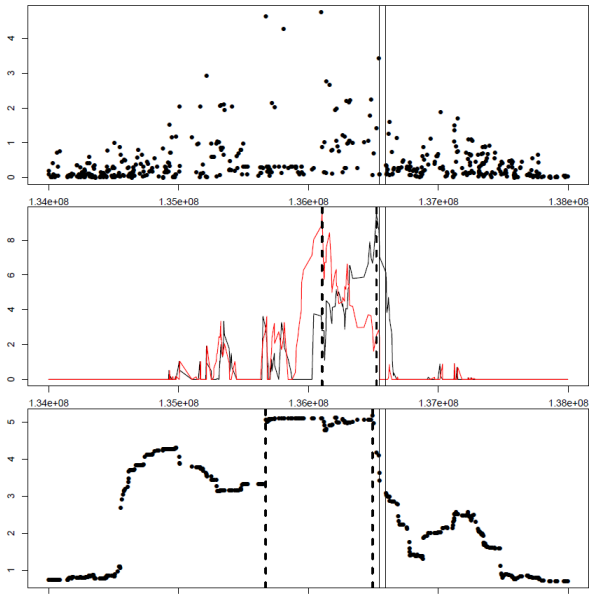

Figure 1: Selection footprints for HapMap data: focus on Lactase region. The lactase gene is located within the 2 vertical solid lines (in the 3 plots). Top graph: -log10(pvalue(F-LK)) of the single marker F-LK test. Middle graph: Lindley process based on the score function -log10(pvalue(F-LK))-1, starting from the left (black curve) and from the right (red curve). Bottom graph: hapFLK values. The detected intervals range between the dotted vertical lines.

Local score and score definition. We propose to start from results of single marker tests along the genome and to highlight segments of adjacent loci with small p-values. Our strategy is to highlight segments of unknown length of high accumulated "scores" X.

Given a score X (here X = - log10(p-value) (up to an additive constant)), the Lindley process is defined as and $h_0 = 0.$. Then, the local score corresponds to the maximum of the Lindley process. The significant region goes from the last 0 before the local score, till the local score.

Results and Discussion

On the HapMap data, the intervals given by the haplotypic method hapFLK (Fariello et al. (2013)) and the local score were close to the Lactase gene, but the local score provided a smaller interval (Figure 1). The local score was based on the score function -log10(pvalue(F-LK))-1, where the p-value was computed from the single marker F-LK test (Bonhomme et al. (2010)). This means that p-values of single marker tests that were lower than 10^{-1} were accumulated to find an interval achieving the local score. On this benchmark region, the local score clearly highlighted a well-known target of selection in Human, thus performing as well as the hapFLK test.

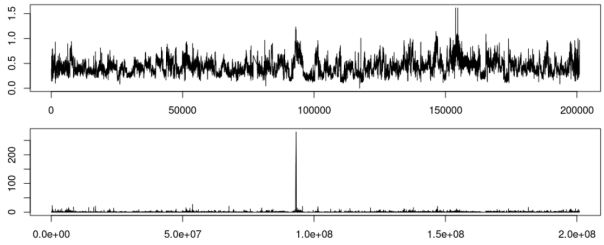

Figure 2: Selection footprints for Quail data on GGA1. Top: windowed Fst (sliding window of 10kb), bottom: local score (Lindley process) of score function -log10(pvalue(Fst))-1.

On the quail data, all methods gave unclear patterns, except the local score that clearly pinpointed 4 short genomic regions (see Figure 2 for GGA1). One or two genes lay in each candidate region. Some of them are known to be related to behavior traits in Human, in particular related to autism, suggesting that the detected regions are not false positives.

Conclusion

The local score approach is general and can be applied to other genome-wide analyzes such as GWAS (e.g. Guedj et al. (2006)) or genome scans for selection. It was particularly suited for the quail data (pool sequencing), because it is computationally very fast and does not require individual genotypic data. It gave clear and coherent results while competing methods lacked power in our data on quails.

Literature Cited

Bonhomme, M., Chevalet, C., Servin, B., et al. 2010. Genetics, 186:241-262.

Fariello, M.I., Boitard, S., Naya, H. et al. 2013. Genetics, 193:929-41.

Guedj, M., Robelin, D., Hoebeke, M., et al. 2006. Stat. Appl Genet. Mol. Biol., 5: Article22.

Johansson, A.M., Pettersson, M.E., Siegel, P.B., Carlborg, O. 2010. PLoS Genet., 6: e1001188.

Karlin, S., Altschul, S.F. 1990. PNAS, 87:2264-8.

Mercier, S., Cellier, D., Charlot, D. 2003. Journal of Applied Probability.

Mills, A., Faure, J. 1991. J. Comparative Psychology, 105:25-38.

Myles, S., Tang, K., Somel, M., et al. 2008. Ann. Hum. Genet., 72:99-110.

Sabeti, P.C., Varilly, P., Fry, B. et al. 2007. Nature, 449: 913-918.

Schweitzer, C., Levy, F., Arnould, C. 2011. Animal Behaviour, 81:535-542.

Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K. 2006. PLoS Biol., 4: e72.