



**HAL**  
open science

## **Integrating single marker tests in genome scans for selection: the local score approach**

María Inés Fariello Rico, Simon Boitard, Sabine Mercier, Magali San Cristobal

► **To cite this version:**

María Inés Fariello Rico, Simon Boitard, Sabine Mercier, Magali San Cristobal. Integrating single marker tests in genome scans for selection: the local score approach. *Mathematical and Computational Evolutionary Biology*, Jun 2017, Porquerolles, France. pp. 1-1, 2017. hal-01798027

**HAL Id: hal-01798027**

**<https://hal.science/hal-01798027>**

Submitted on 23 May 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : <http://oatao.univ-toulouse.fr/>  
Eprints ID : 19674

**To cite this version** : Fariello Rico, María Inés and Boitard, Simon<sup>✉</sup> and Mercier, Sabine and San Cristobal, Magali<sup>✉</sup> *Integrating single marker tests in genome scans for selection : the local score approach.* (2017) In: Mathematical and Computational Evolutionary Biology, 12 June 2017 - 16 June 2017 (Porquerolles, France). (Unpublished)

Any correspondence concerning this service should be sent to the repository administrator: [staff-oatao@listes-diff.inp-toulouse.fr](mailto:staff-oatao@listes-diff.inp-toulouse.fr)

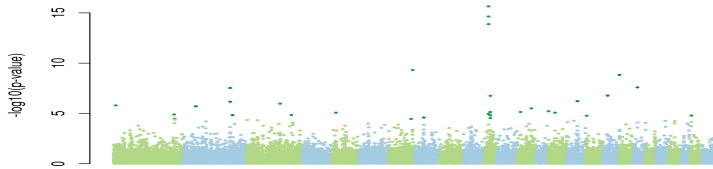
# Integrating single marker tests in genome scans for selection : the local score approach

María Inés Fariello<sup>1</sup>, Simon Boitard<sup>2</sup>, Sabine Mercier<sup>3</sup>, Magali San Cristobal<sup>4</sup>

(1) Univ. de la República, Montevideo, Uruguay. (2) GenPhySE, INRA Toulouse, France. (3) IMT, Univ. Toulouse, France. (4) Dynafor, INRA Toulouse, France.

## MOTIVATION

- Detect **genomic regions with high genetic differentiation between populations**, signatures of **adaptive selection**.
- **Single-marker statistics** have a **large variance** and **ignore LD** (Linkage Disequilibrium).

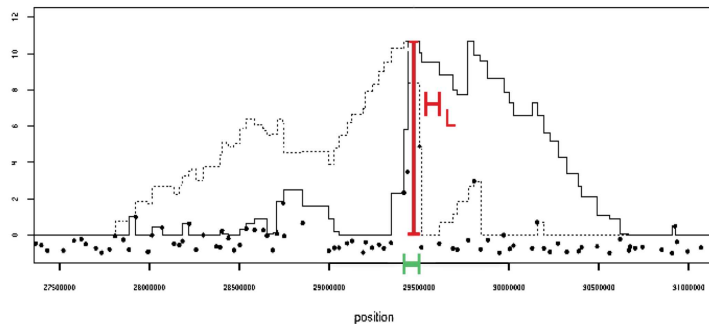


- **Haplotype-based tests** require **individual data**, not available when sequencing DNA pools (Pool-seq).
- **Window-based tests**: how to choose **window size**? **Statistical significance** of a window?
- **Local score**: detects regions **statistically enriched in markers with high genetic differentiation, without defining fixed windows**.

## DEFINITION

- For each marker  $m$ , **score**  $X_m = -\log_{10}(p_m) - \xi$  (black points),  $p_m$  p-value of a test for selection (i.e. rejecting neutrality).
- **Cumulate scores** using the **Lindley process** (solid line)

$$h_0 = 0, \quad h_m = \max(0, h_{m-1} + X_m)$$



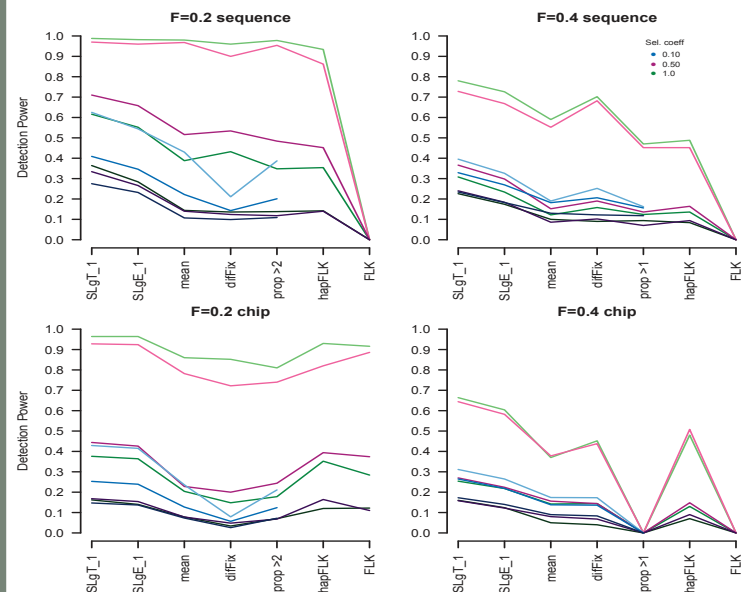
- **Excursions above 0** of the Lindley process, indicate **genomic regions enriched in high scores / low p-values** (green interval).
- Here  $p_m$  is the p-value of the FLK test (Bonhomme *et al*, 2010).

## IMPORTANT FEATURES

- One single **tuning parameter**:  $\xi$ , p-value threshold in log10 scale.
  - **High values** put emphasis on **high scores**: strong selection.
  - **Low values** put emphasis on **extended regions**: recent selection.
  - $\xi = 1$  recommended to optimize detection power.
- **Statistical significance of an excursion** depends on:
  - the number of markers in the sequence ( $M$ )
  - the auto-correlation of scores ( $\rho$ ).
- **Two new approaches** to compute it, as a function of  $M$  and  $\rho$ :
  - **analytical formula**: valid if single-marker p-values are uniform under neutrality.
  - **re-sampling approach**: valid for all datasets.

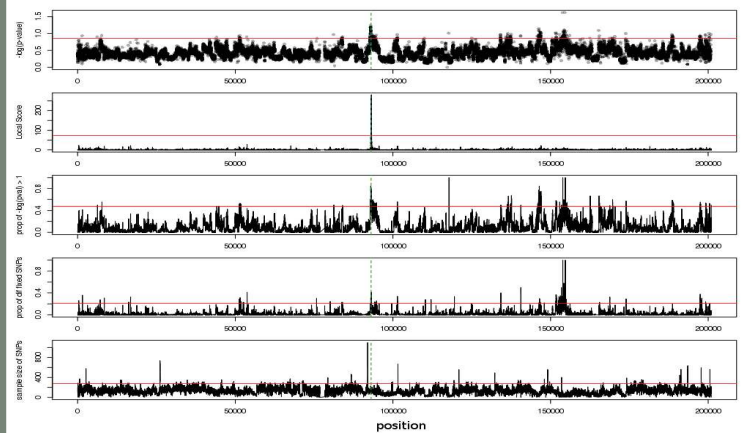
## SIMULATION RESULTS

- **Two divergent populations**, one neutral and one under selection.
- Methods: **single-marker test** (FLK), **haplotype-based test** (hapFLK, Fariello *et al*, 2013), **window-based FLK tests** (mean, diffx, prob >2), **local score**, detection threshold computed from our re-sampling approach (SLgT) or neutral simulations (SLgE).
- Observed type I error 6% for SLgT, 5% for other tests.



## APPLICATION TO DIVERGENT QUAIL LINES

- Divergent selection on behaviour, **pool-seq** from each line (G50).
- Chromosome 1: much clearer signal with the local score.
- Ten significant regions genome-wide, with **relevant candidate genes** related to autistic disorders or behavioral traits in Humans.



## CONCLUSIONS

- The **local score** accounts for **LD without individual genotypes**.
- **Statistical significance** of candidate regions easy to compute.
- **Increased detection power** compared to single-marker, window-based or haplotype-based tests.
- Can be applied to **any single-marker test providing p-values**.