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

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

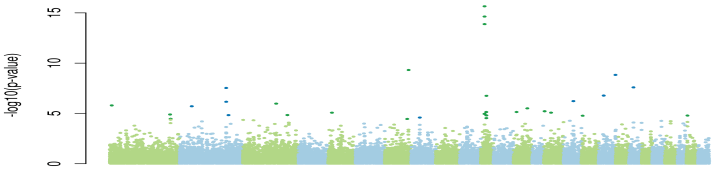
Integrating single marker tests in genome scans for selection : the local score approach

María Inés Fariello[1], **Simon Boitard**[2], Sabine Mercier[3], Magali San Cristobal[4]

(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 Disequilinrium).

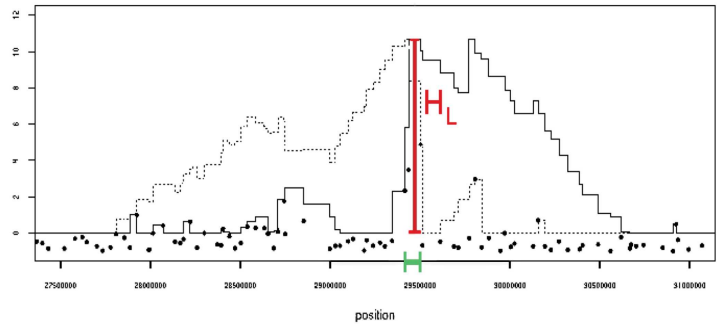


- **Haplotype-based tests** require **individual data**, not availale when sequencing DNA pools (Pool-seq).
- **Window-based tests**: how to choose **window size**? **Statistical significance** of a window?
- **Local score:** detects regions **statistically enriched** in markers with **high genetic differentiation**,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

Definition

- For each marker m, **score** $X_m = -log10(p_m) - \xi$ (black points), p_m p-value of a test for selection (i.e. rejecting neutrality).
- **Cumulate scores** using the **Lindley process** (solid line)

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



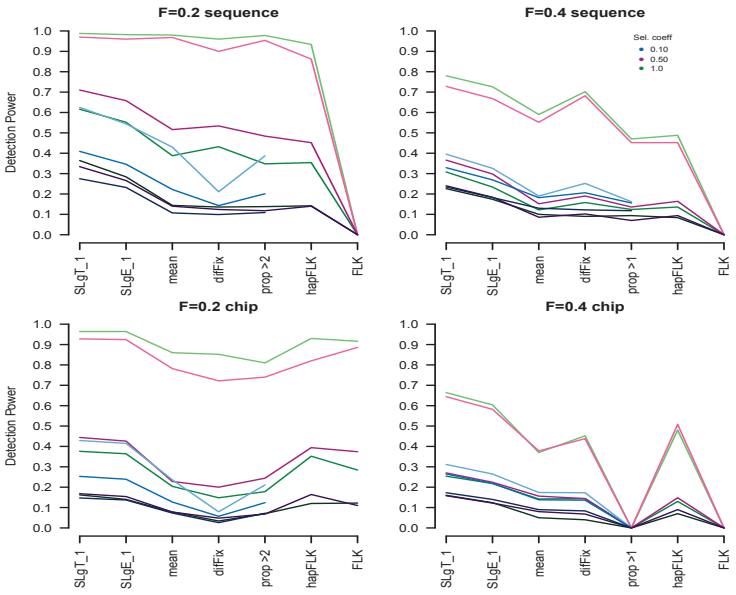
- **Excursions above 0** of the Lindley process, indicate **genomic regions enriched in high scores / low p-values** (green interval).
- Here p_m is the p-value of the FLK test (Bonhomme *et al*, 2010).

Important features

- One single **tunning parameter**: ξ, 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 (ρ).
- **Two new approaches** to compute it, as a function of M and ρ:
  → **analytical formula:** valid if single-marker p-values are uniform under neutrality.
  → **re-sampling approach:** valid for all datasets.

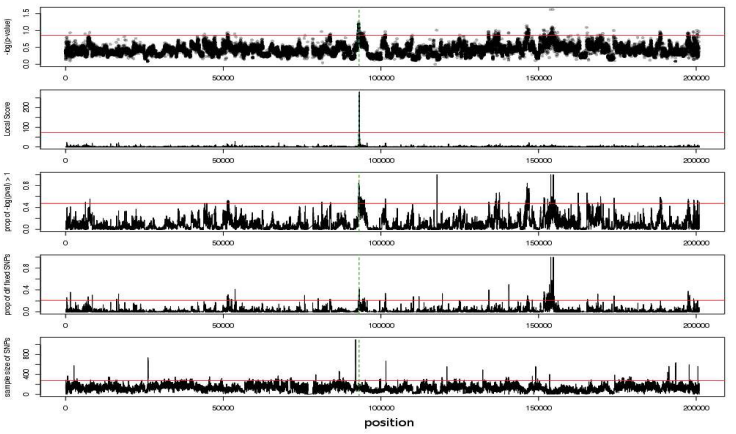
Simulation results

- **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 behavorial traits in Humans.



Conclusions

- The **local score** accounts for **LD whithout 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**.