Accounting for Linkage Disequilibrium in genome scans for selection without individual genotypes: the local score approach

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

[1] : Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay
[2] : **Génétique, Physiologie et Systèmes d'Elevage (GenPhySE), INRA Toulouse**
[3] : Institut de Mathématiques de Toulouse (IMT), Université de Toulouse
[4] : Dynamiques et écologie des paysages agriforestiers (Dynafor), INRA Toulouse

Models in Ecology and Evolution, May 30-31, 2017

Outline

1. Motivations
2. The FLK & hapFLK approaches
3. The local score approach
4. Simulation results
5. Examples
6. Conclusions

Outline

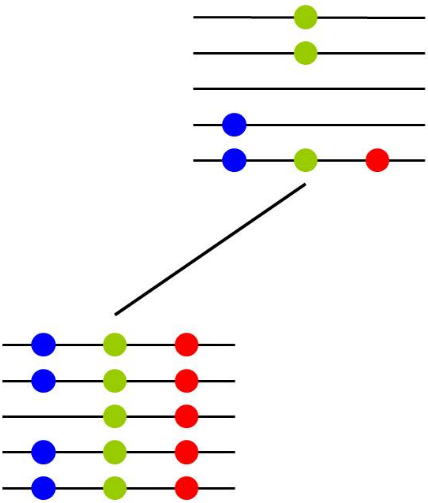
1. Motivations
2. The FLK & hapFLK approaches
3. The local score approach
4. Simulation results
5. Examples
6. Conclusions

Genome scans for selection

- Most genomic regions are neutral, but some of them are (or have been) under selection (natural or artificial).
- Detecting the regions under selection is important for theory (evolution) and applications (medicine, agronomy).
- Genome wide scans for selection now possible from dense genotyping (SNP chips) or sequencing (NGS) data.
- Focus on positive (adaptative) selection.

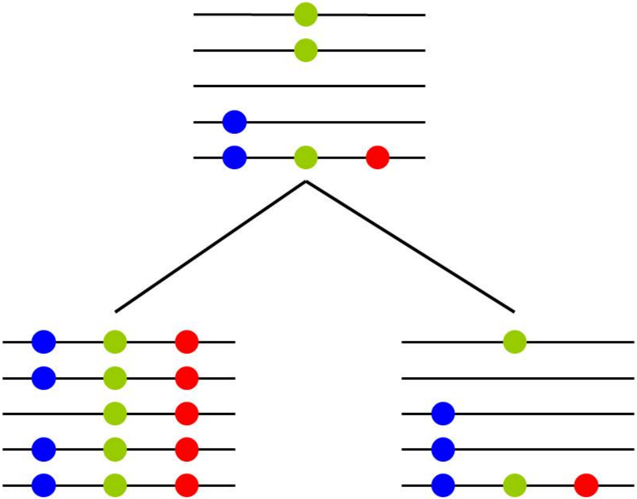
Population differentiation approach

Look for markers with contrasted allele frequencies between populations.

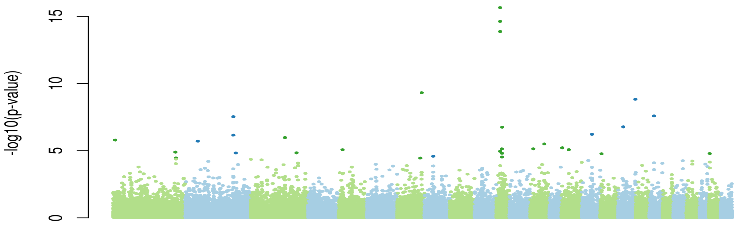


Population differentiation approach

Look for markers with contrasted allele frequencies between populations.



Linkage Disequilibrium (LD) helps!



- **Single-marker statistics** have a **large variance**, high values can be reached just by chance due to drift.
- Due to LD, markers in the **neighborhood of a selected locus** also show **elevated differentiation** between populations.

→ Account for LD in selection scans by:

1. using haplotype tests
2. looking for clusters of markers with high differentiation

Windowing approaches

- **Cut the genome into fixed windows** and computes a summary of the single-marker statistics within each window.
- **Summarize each window** using:
  - the average of single-marker statistics (Weir *et al*, 2005).
  - the number of markers exceeding a given threshold (Myles *et al*, 2008).
  - the number of markers differentially fixed between populations (Johansson *et al*, 2010).
- **Individual genotypes not required** (pooled sequencing).
- Limitations:
  - How to choose **window size**? the **single-marker threshold**?
  - How to decide that a **window** is **under selection?**

→ **Overcome these issues** using the statistical **local score** theory.

Outline

1 Motivations

2 The FLK & hapFLK approaches

3 The local score approach

4 Simulation results

5 Examples

6 Conclusions

F_{ST} based tests

$p = (p_1, \dots, p_i, \dots, p_n)$: allele frequencies at one SNP in several populations.

$\overline{p}$ and s_p^2: observed mean and variance of p.

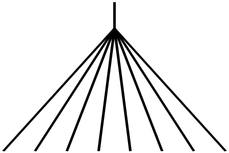
$$F_{ST} = \frac{s_p^2}{\overline{p}(1-\overline{p})}$$

- H_0 : "neutral evolution" (genetic drift) vs H_1 : "positive selection in one (or more) population ".
- H_0 rejected if F_{ST} too large.

Lewontin et Krakauer (LK) test (1973)

$$LK = \frac{n-1}{\bar{F}_{ST}} F_{ST}$$

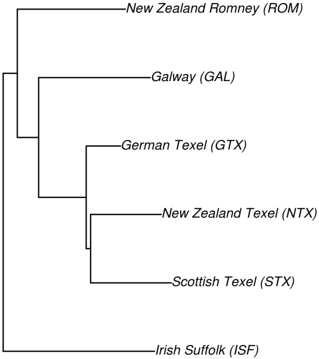
- LK **distribution under** H_0 **is** χ^2 with $n-1$ degrees of freedom.
- But, only true if populations have a **star like phylogeny with equal population sizes**.



FLK test (Bonhomme *et al*, 2010)

Extension of LK accounting for

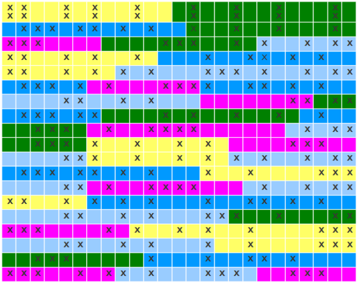
- differences in effective size between populations.
- differences in correlations between population pairs.



(first estimated from genome wide data)

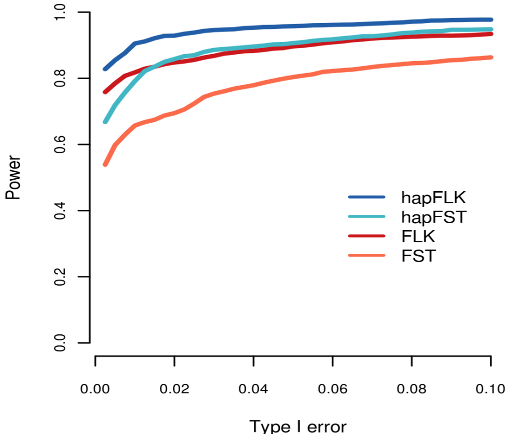
hapFLK test (Fariello *et al*, 2013)

- **Define local haplotypes** around each SNP position using the model of Scheet and Stephens (2006).



- **Compute haplotype frequencies** in each population.
- Apply FLK, considering haplotypes as alleles.

Detection power



4 populations with hierarchical structure, 1 under selection.

Outline

1 Motivations

2 The FLK & hapFLK approaches

3 The local score approach

4 Simulation results

5 Examples

6 Conclusions

Definition

- For each marker m, define the **score**:

$$X_m = -log10(p_m) - \xi$$

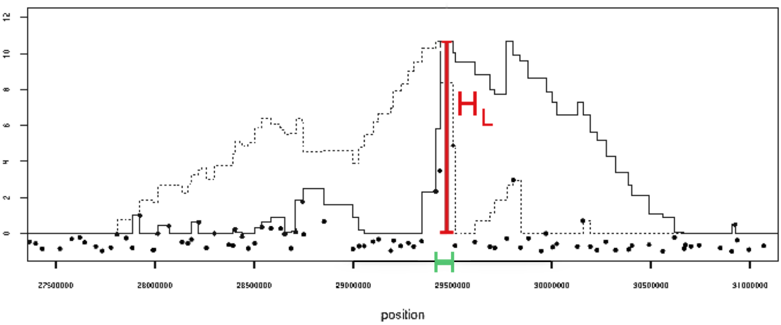
  p_m p-value of a test for selection, ϵ fixed threshold.

- Low p-value $=$ H_0 (neutral evolution) unlikely $=$ high score.
- **Cumulate scores** using the so-called **Lindley process**:

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

- Look for **local maxima of the Lindley process**, which are asociated to **genomic regions** that are **enriched in high scores / low p-values**.
- Here p_m is the p-value of FLK.

Example



- The Lindley process (black line) has several **excursions above 0** (local maxima).
- The **global maximum** (H_L) is called the **local score**.
- Each excurion is associated to an interval enriched in high scores (in green).

choosing ξ

- **p-value threshold** in log10 scale.
- Ex: $\xi = 2$ cumulates p-values below 10^{-2}.
- For **high** ξ, only **most significant markers** contribute:
  $\rightarrow$ similar to single point approach.
  $\rightarrow$ **strong selection**.
- For **low** ξ, more markers contribute:
  $\rightarrow$ **longer** intervals.
  $\rightarrow$ **recent selection**.

Statistical evidence for selection

- How likely is a given excursion under neutrality?
- Depends on:
  - the number of markers in the sequence (M).
  - the correlation between scores (ρ).
- We provided **two approaches** allowing to **compute significance thresholds for excursions** :
  1. **analytical formula:** valid if single-marker p-values are unifrom under neutrality.
  2. **re-sampling approach:** valid for all datasets, but requires some computing time.

Outline

1 Motivations

2 The FLK & hapFLK approaches

3 The local score approach

4 Simulation results

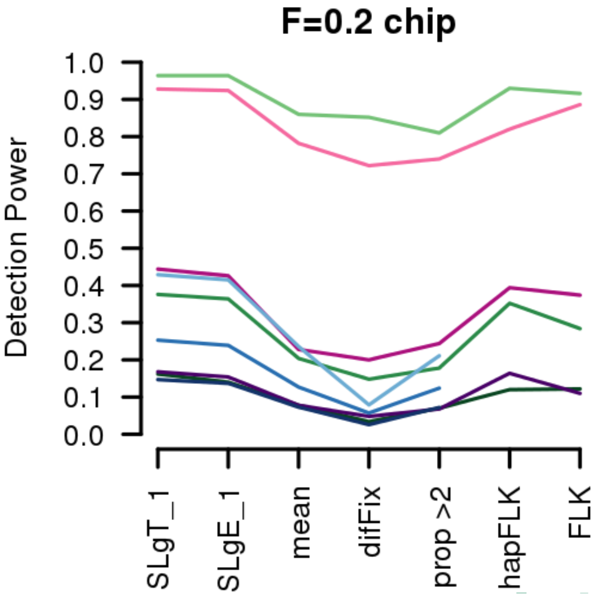
5 Examples

6 Conclusions

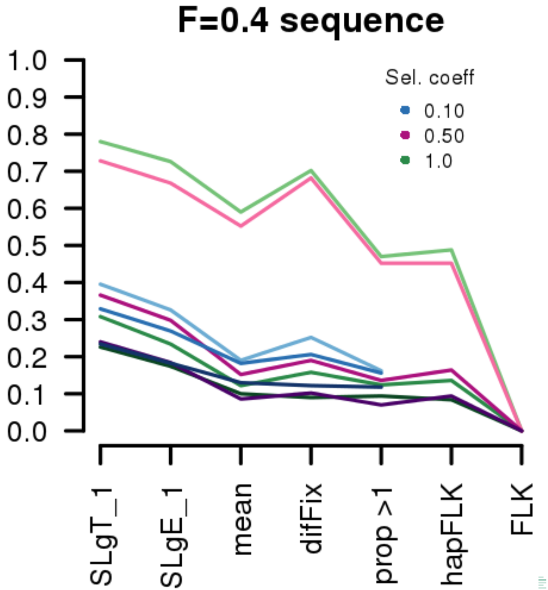
Simulation procedure

- **Two populations** with same effective size, one neutral and **one under selection**.
- Genomic region of **10Mb** with **one selected site**.
- Several statistics compared, in different scenarios.
- **Detection threshold** of each statistic such that selection is detected in 5% of the **neutral samples** (type I error 5%).
- For the local score, also computed using our re-sampling approach
  → observed type I error 6%.
- Tunning parameters (window size, ξ ...) chosen to optimize detection power.

Results


F=0.2 chip
Detection Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
SLgT_1
SLgE_1
mean
difFix
prop >2
hapFLK
FLK

Results


F=0.4 sequence
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Sel. coeff
0.10
0.50
1.0
SLgT_1
SLgE_1
mean
difFix
prop >1
hapFLK
FLK

Outline

1 Motivations

2 The FLK & hapFLK approaches

3 The local score approach

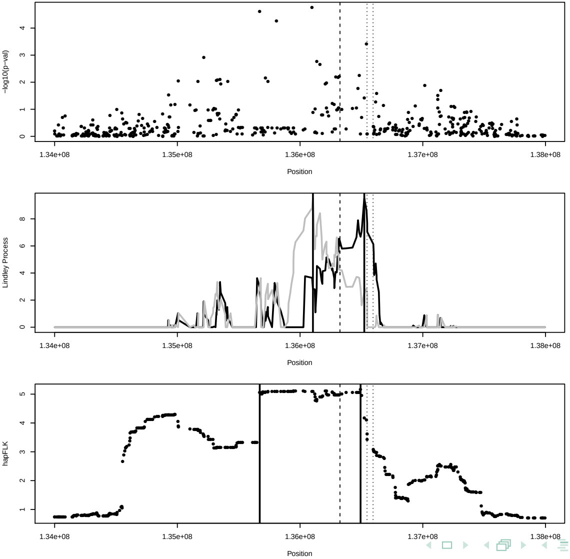
4 Simulation results

5 Examples

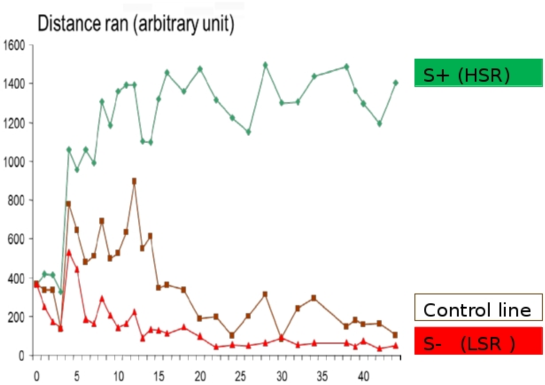
6 Conclusions

Lactase region in Humans

Test of selection based on HapMap genotypes (Europea and Asia).

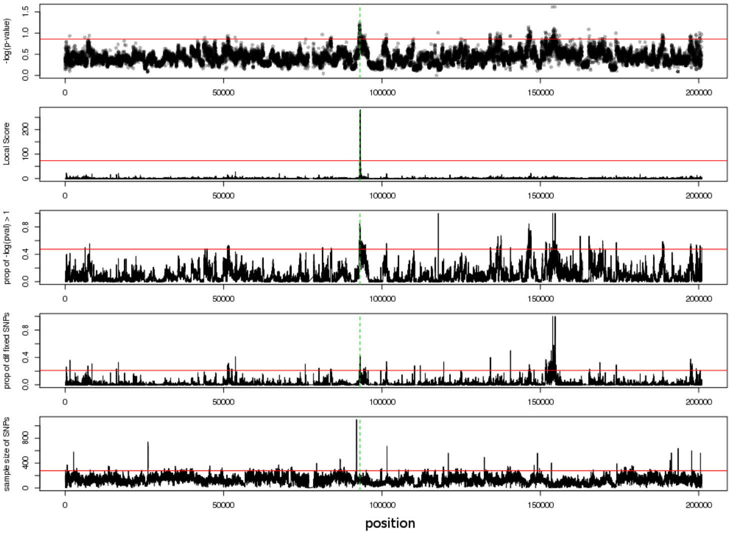


Divergent selection experiment on behaviour in Quail



- **Pooled DNA** from each line sequenced at generation 50
- **Strong drift** ($F = 0.4$).

Selection scan on chromosome 1



Significant regions genome-wide

Chr.	Position	L (kb)	Genes
1	92,963,481-93,182,440	219	NSUN3, **ARL13B**
2	1,584,033-1,688,400	104	VIPR1
3	61,586,217-61,604,464	19	ECHDC1, RNF146
3	75,088,250-75,170,494	82	MMS22L
4	11,412,372-11,452,609	40	GLOD5
4	90,953,044-91,008,245	56	**CTNNA2**
6	35,234,870-35,336,720	102	FOXI2, **PTPRE**
6	6,311,718-6,644,395	333	UBE2D1, CISD1, **IPMK**
10	17,825,157-17,825,227	0.07	
25	1,296,647-1,296,706	0.059	

Genes **in bold** have been associated to **autistic disorders** or **behavorial traits** in Humans.

Outline

1 Motivations

2 The FLK & hapFLK approaches

3 The local score approach

4 Simulation results

5 Examples

6 Conclusions

Detecting selection using the local score

- Accounts for **LD whithout individual genotypes**.
- One single tunning parameter, ξ, with intuitive interpretation. $\xi = 1$ recommended for detection power.
- **Statistical significance** of candidate regions easy to compute.
- **Increased detection power** compared to single-marker, window-based or haplotype-based tests.
- Convincing results on 2 real datasets with different features.
- Can be applied to **any single-marker test providing p-values**, for selection scans or **any** other **context**.
- Ref: Fariello *et al*, Molecular Ecology 2017.

Acknowledgements

Quail husbandry and sampling:

- **Cécile Arnould & Christine Leterrier**, Unité de Physiologie de la Reproduction et des Comportements, INRA Tours
- **Julien Recoquillay**, Unité de Recherches Avicoles, INRA Tours
- **David Gourichon**, Pôle d'Expérimentation Avicole, INRA Tours

Computing Facilities:

- Genotoul bioinformatics platform Toulouse Midi-Pyrénées.

DNA preparation and sequencing:

- **Olivier Bouchez & Gérald Salin**, GeT-PlaGe Genotoul, INRA Toulouse
- **Sophie Leroux & Frédérique Pitel**, GenPhySE, INRA Toulouse

Bioinformatic and statistic analyses:

- **Patrice Dehais**, SIGENAE, INRA Toulouse
- **David Robelin & Thomas Faraut**, GenPhySE, INRA Toulouse

Advertising

- PhD position available at Toulouse, from september 2017.
- Supervised by Lounès Chikhi (Evolution et Diversité Biologique) and Olivier Mazet (INSA).
- Influence of population structure on past population size estimation (Mazet *et al*, Heredity 2017).