Séminaire MIAT, 23 juin 2017

De la sélection des cailles au score local

Ou pourquoi faire compliqué quand on peut faire simple …

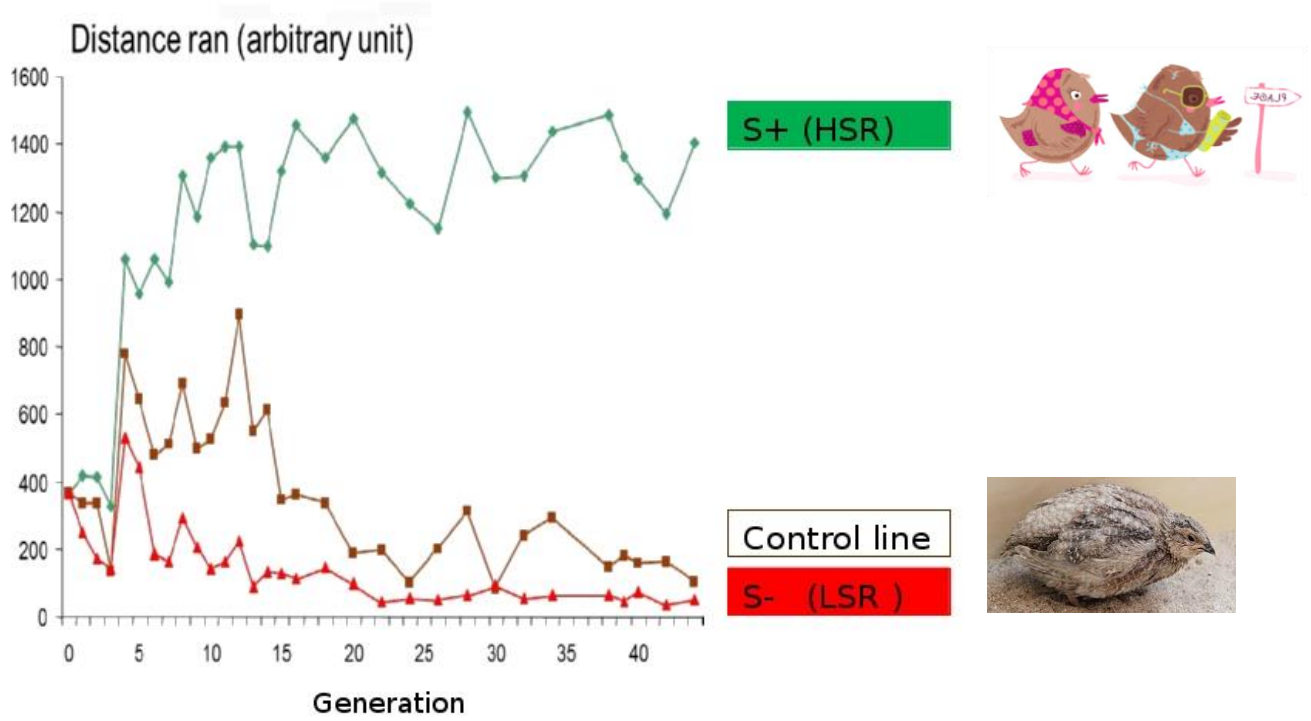
Magali San Cristobal
Maria-Ines Fariello, Simon Boitard, Sabine Mercier, David Robelin
(et beaucoup d'autres co-auteurs)

Fariello et al (2017) Mol Ecol

Plan : du particulier au général

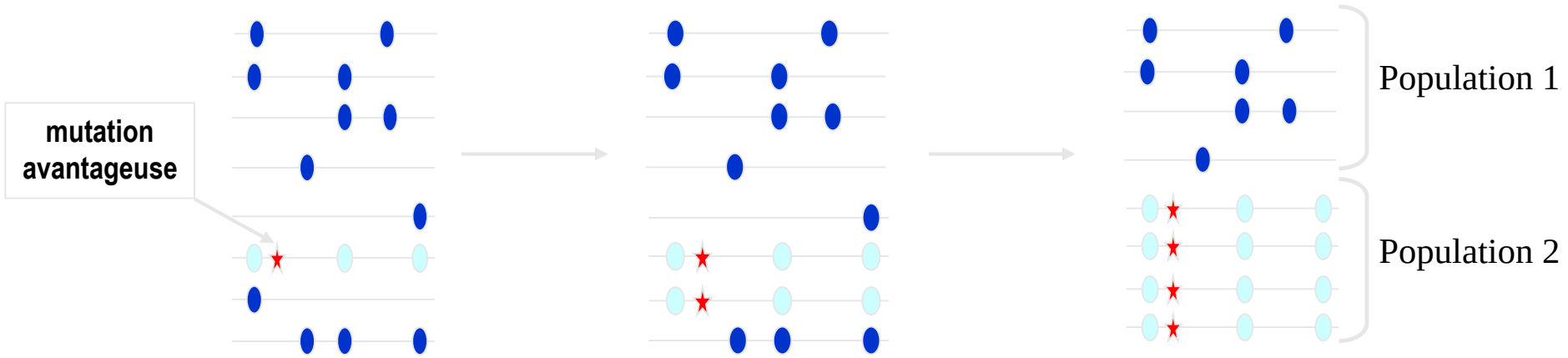
- On regarde des cailles courir
- On se demande ce qu'il y a là-derrière
- Flute, les méthodes usuelles ne marchent pas
- On propose autre chose : nom de code SL
- Ce quelque chose de nouveau et d'ancien est applicable dans bien d'autres domaines

A l'INRA on s'amuse à sélectionner des cailles bougonnes et des cailles copines



Séquençage d'un pool de 10 animaux dans chaque lignée

Qu’est-ce qui se passe ?



On détecte la sélection en mesurant localement la diversité génétique intra et entre populations

Au niveau de la mutation sélectionnée (★) et des sites neutres en déséquilibre de liaison :

✓ réduction de diversité génétique dans la population sélectionnée → Hét.

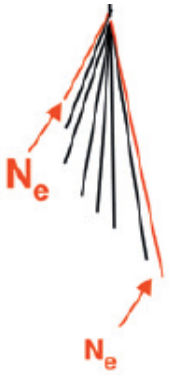
✓ augmentation de la différenciation entre populations → F_{st}

d’après P Garnier Géré

Nous, on a déjà travaillé sur ce sujet, mais pas que nous

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

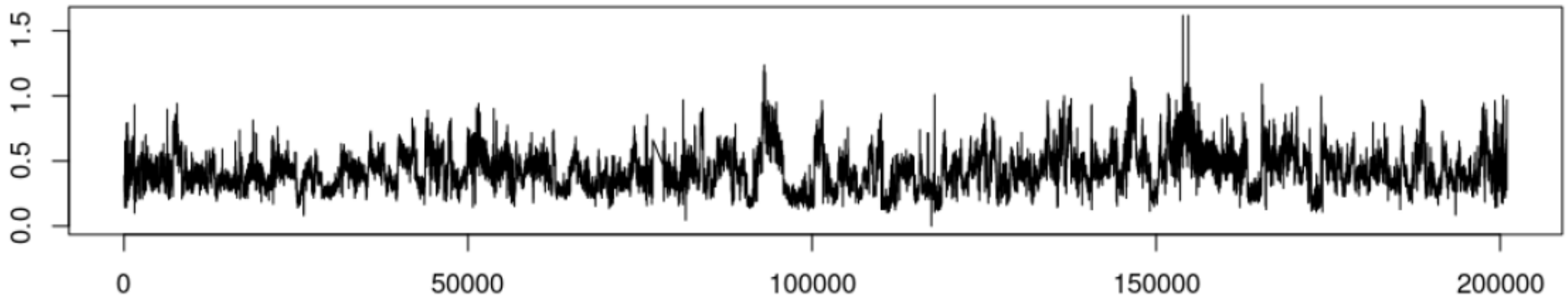
- FLK (Bonhomme et al 2010)



- hapFLK (Fariello et al 2013)

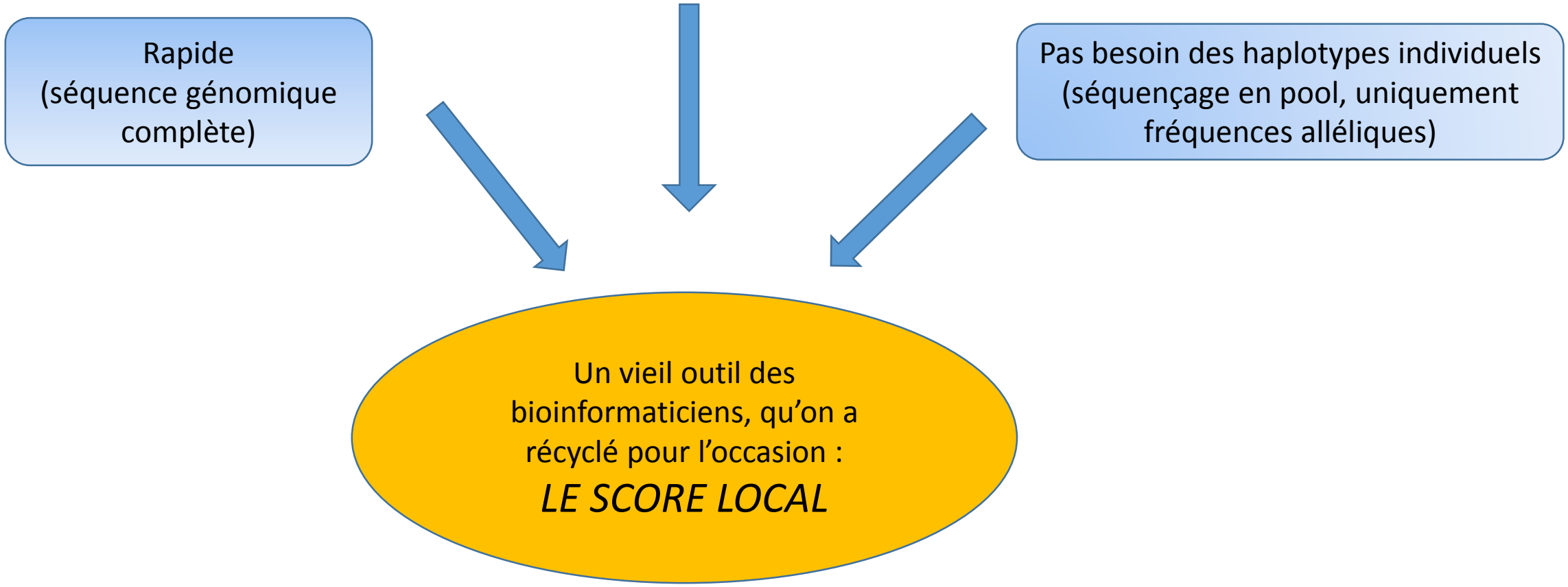
- Fenêtres glissantes, etc ….

Eh ben ça donne rien sur les cailles !



Car beaucoup de bruit (dérive génétique) masque le signal (sélection)

Une idée : cumuler les faibles signaux proches



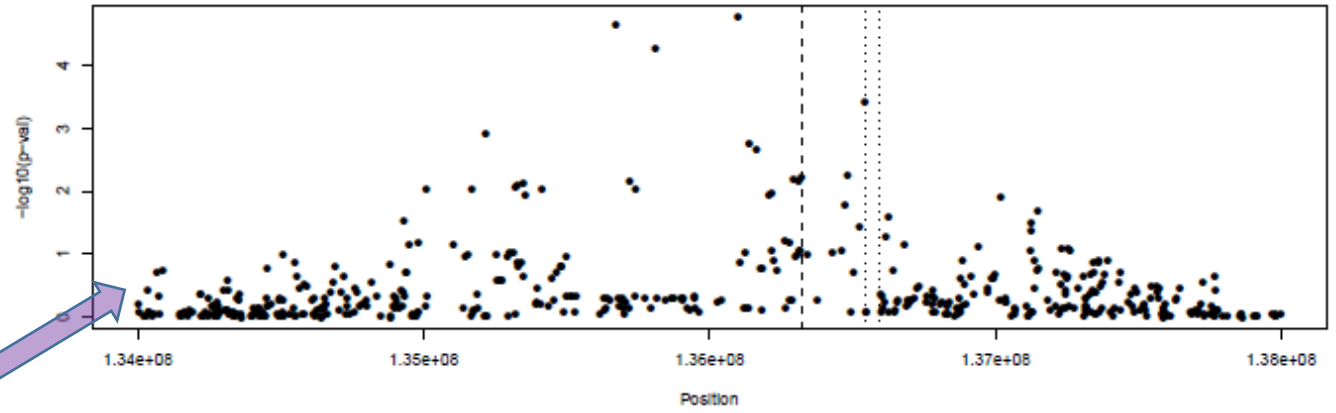
LE score local

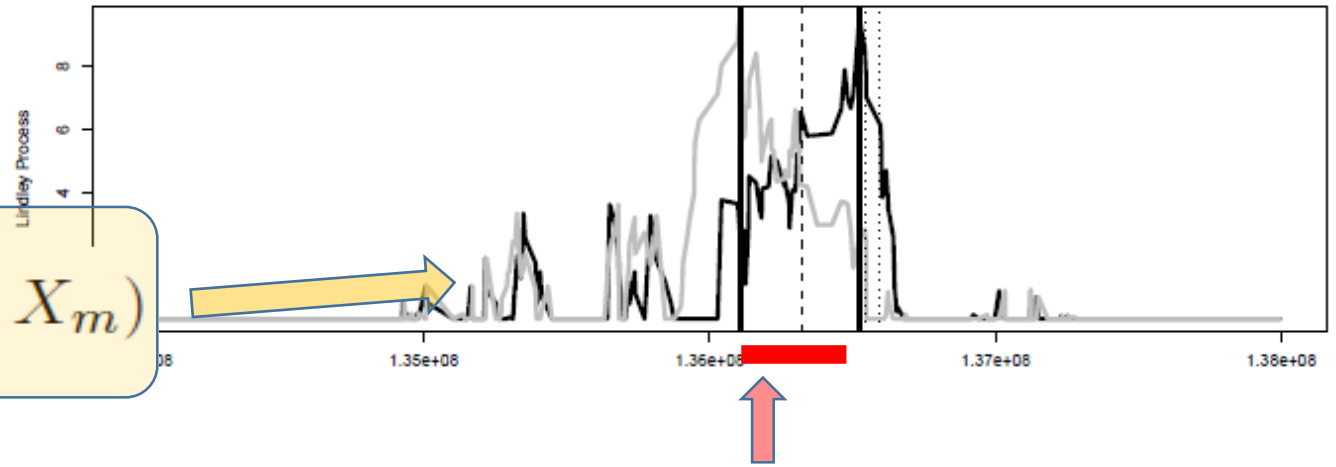
Des tests marqueur par marqueur :
p-valeurs p_m

Des scores :

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

tq $E(X_m) < 0$





Le processus de Lindley :

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

$$H_M(X) = \max_{1 \le m \le M} h_m.$$

Le score local :

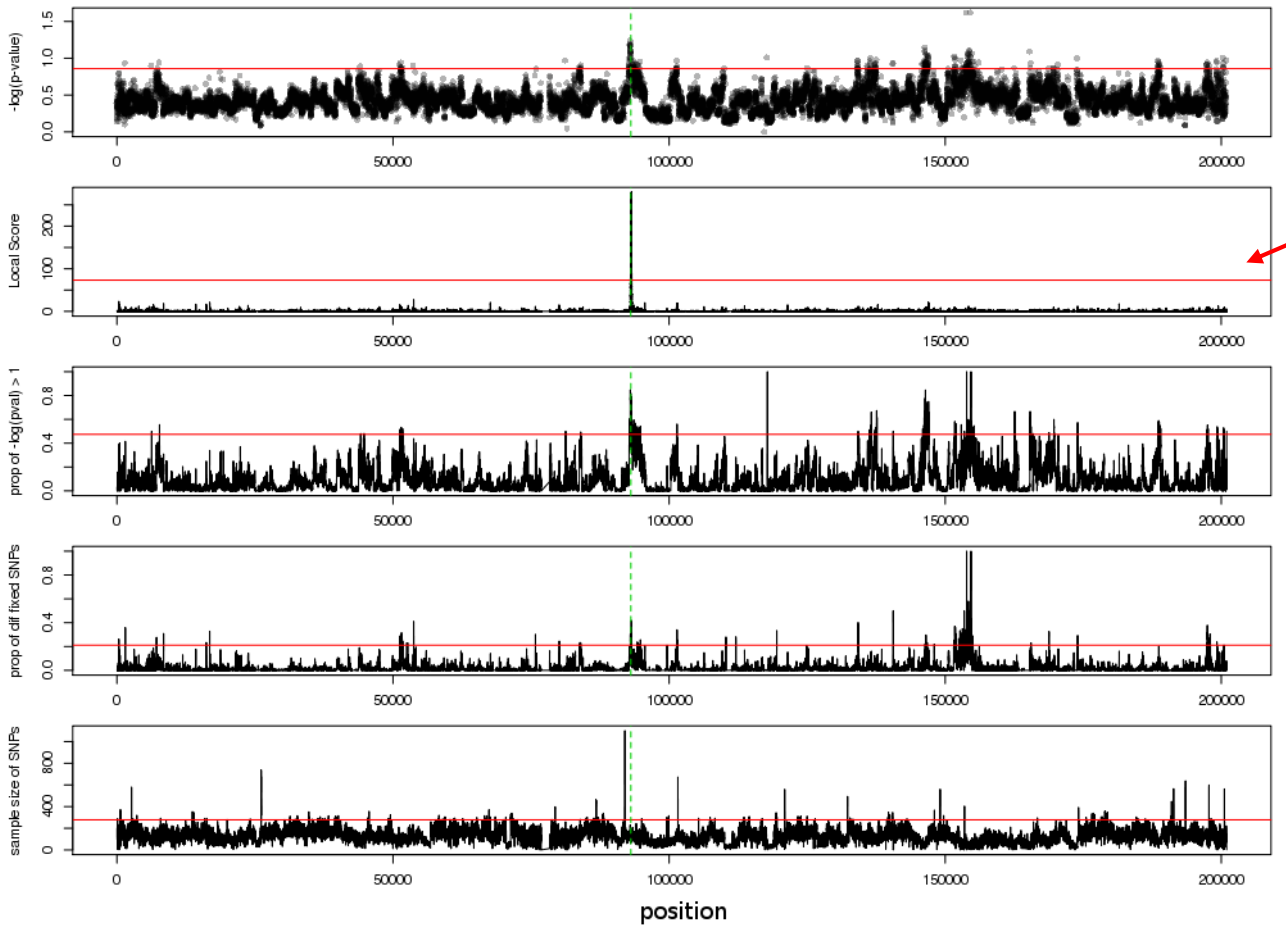
$$H_M(X) = \max_{1 \le i \le j \le M} \{X_i + \cdots + X_j\}$$

Et sur les cailles, ca marche du feu de Dieu !

Test marqueur par marqueur

LE score local !

Fenêtres glissantes



Ok, mais on coupe où ?

Une 10ne de régions génomiques courtes, de 1 à 3 gènes dans chaque, dont des gènes candidats (autisme chez l'homme)

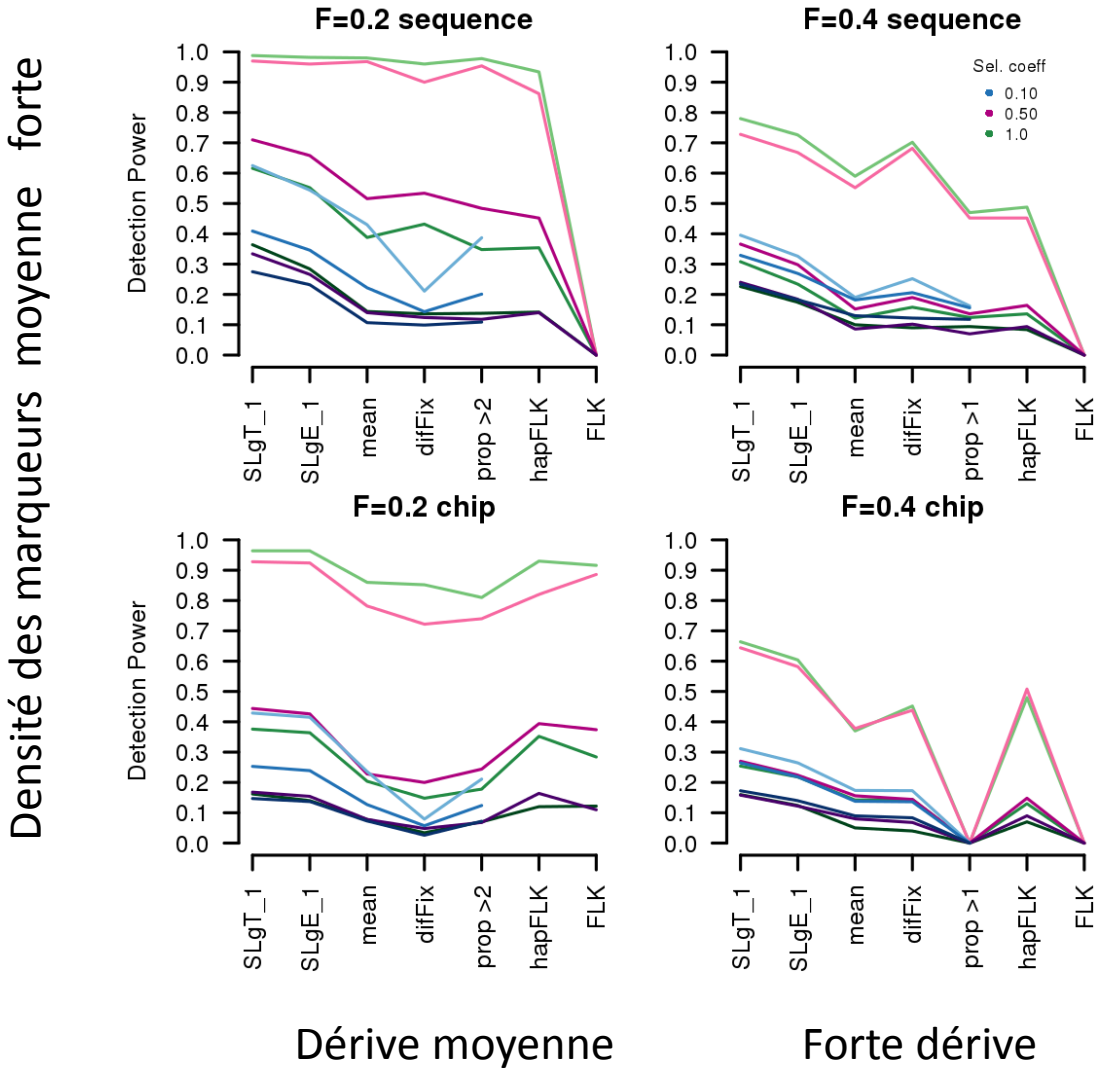
Le calcul du seuil : où en est-on ?

- Cas iid
  - Distribution asymptotique (Karlin et Dembo 1992)
  - Distribution exacte pour séquences courtes (Mercier et Daudin 2001)
- Cas markovien
  - Distribution asymptotique (Robelin 2005 ; ce travail) : loi de Gumbel

$$log(-log(F_{M,\rho}(y)) = a_{L,\rho} + b_{L,\rho}y + e$$

  - Distribution exacte mais temps de calcul trop long (Hassenforder et Mercier 2007)

Les simulations nous donnent raison



Le score local se maintient au niveau de notre « champion » hapFLK si dérive moyenne

Le score local est particulièrement adapté quand la dérive est forte

Vos « oui mais … »

- La molette à régler :
  - Plus intuitif que la taille des fenêtres
  - Seuil des p-valeurs : ξ = 2 : on cumule les p-valeurs supérieures à 10^{-2}
  - Même localisation avec 1 et 2
  - Petit : plus de marqueurs, intervalles plus longs et plus nombreux, sélection faible
  - Grand : moins de marqueurs, intervalles plus courts et moins nombreux, sélection forte
- 10 animaux par pool : ok (simulations)

C’était facile, pas cher, et ça a rapporté gros

- Le score local est idéal quand il y a beaucoup de bruit dans les données, et qu’on s’attend à ce que le signal s’étale sur un segment : traces de sélection, GWAS, etc …
- Mise en œuvre extrêmement simple et rapide
- On a avancé sur sa distribution dans le cas non iid
- Résultat miraculeux sur les cailles : des courts segments génomiques très bien identifiés et contenant des gènes candidats

Merci !

- Beaucoup de monde a travaillé :

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

- Merci pour votre attention!