Detecting past contraction in population size using haplotype homozygosity
C Merle, Jean-Michel Marin, F. Rousset, Raphaël Leblois

To cite this version:
C Merle, Jean-Michel Marin, F. Rousset, Raphaël Leblois. Detecting past contraction in population size using haplotype homozygosity. Mathematical and Computational Evolutionnary Biology 2016, Jun 2016, Montpellier, France. hal-02932275

HAL Id: hal-02932275
https://hal.inrae.fr/hal-02932275
Submitted on 7 Sep 2020

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.
NEXT generation genetic data

Classical inference methods of the demographic history (likelihood based methods, IS, MCMC), approximate bayesian methods and Site Frequency Spectrum, suitable for polymorphism data sets consisting in some loci and assuming the genealogies of different loci are independent, do not exploit the recombination.

Take advantage of genome wide data with known genome

Pairwise alignment

GT CATGACCGGA

GA CATGACCGGA

New inference approaches of demographic history based on the conserved sequence lengths in the pairwise alignment within a diploid genome : [3], [5], [4], [2], [1]

The patterns of linkage disequilibrium (LD) between polymorphic markers are shaped by the ancestral population history as we can see on haplotype homozygosity curves which measure the LD.

DEMOGRAPHIC HISTORY INFERENCE FROM GENOME WIDE SEQUENCE DATA

Haplotype Homozygosity: $HH(i):$ probability for $i$ adjacent markers drawn at random in the whole genome sequence to be homozygotes.

Theoretical $HH$ of [3], denoted $HH_0:$ coalescent based computation, assuming the mutation and recombination rate known and constant along the genome.

Empirical $HH$, denoted $HH$ computed from the observed data, is the segment proportion of at least $i$ homozygotes adjacent markers.

We estimate $\theta_0$ and $\theta_1$ by:

$\hat{\theta}_0 \in \arg \min_{\theta_0} \sum_i w_i(\hat{HH}_0(\theta_0,i) - \hat{HH}(i))^2$,

$\hat{\theta}_1 \in \arg \min_{\theta_1} \sum_i w_i(\hat{HH}_1(\theta_1,i) - \hat{HH}(i))^2$.

Evaluation of $i \rightarrow HH_0(\theta_0,i)$ is time consuming

$\sim$ optimization with the Markovian R package.

Model choice criterion: embedded models

Penalized mean square criterion

$\min_{\theta_0, \theta_1} \sum_i w_i(\hat{HH}_0(\theta_0,i) - \hat{HH}(i))^2$.

$w_i(\theta) = \frac{\theta^2}{\theta^2 + \phi^2}$,

$\phi$ = estimate of Sobol’s sensitivity index of order one of a given parameter $\phi$ computed for $i$ adjacent markers under the more complex model (sensitivity R package).

On this simulated data set with constant population size $\theta_0 = 500$, the non penalized mean square criterion would choose $M_0$, whereas the penalized mean square criterion choose $M_1$.

Detected a contraction

Avoid choosing the more complex model

Detecting a contraction

Numerical results

Data sets simulated under $M_0$, $\theta_0 = N_0 = 500$

<table>
<thead>
<tr>
<th>$n_{pen}$</th>
<th>0000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
<td>2000</td>
<td>20000</td>
</tr>
</tbody>
</table>
| 100 | Adjustment | 0.74 | 0.59 | 0
| Penalized Adj. | 0.94 | 0.98 | 1 |

$\theta_0$, $N_0$

$\begin{cases} 494 & 494 \\ [478 - 510] & . \end{cases}$

$\theta_1$, $N_0$

$\begin{cases} 502 & 501 \\ [496 - 509] & . \end{cases}$

Data sets simulated under $M_1$, $\theta_1 = (N_1, h, f_1) = (500, 1500, 10)$

<table>
<thead>
<tr>
<th>$n_{pen}$</th>
<th>0000</th>
<th>15000</th>
<th>20000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
<td>2000</td>
<td>20000</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>Penalized Adj.</td>
<td>0.81</td>
<td>0.92</td>
<td>0.95</td>
</tr>
</tbody>
</table>

$\overline{N_1}$

$\begin{cases} 634 & 633 \\ [617 - 650] & - \end{cases}$

$\begin{cases} 561 & 529 \\ [547 - 576] & - \end{cases}$

$\begin{cases} 534 & 492 \\ [520 - 548] & - \end{cases}$

Data sets simulated under $M_0$, $\theta_0 = N_0 = 500$

<table>
<thead>
<tr>
<th>$n_{pen}$</th>
<th>0000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
<td>2000</td>
<td>20000</td>
</tr>
</tbody>
</table>
| 100 | Holstein data sets

$\overline{N_0}$

$\begin{cases} 6690 & 6924 \\ [6533 - 6847] & - \end{cases}$

$\overline{\theta_1}$

$\begin{cases} 6565 & 6532 \\ [5599 - 6372] & - \end{cases}$

$\overline{f_1}$

$\begin{cases} 4.23 & 4.0 \\ [3.90 - 4.57] & - \end{cases}$

C. Merle1,3,5, J-M. Marin1,3, F. Roussel3 and R. Leblois2,3

1 - Institut de Montpellier Alexander Grothendieck (IMAG, UM); 2 - Centre de Biologie pour la Gestion des Populations (CBGP, INRA); 3 - Institut de biologie computationnelle (IBC); 4 - Institut des Sciences de l’Evolution (ISEM, CNRS)

REFERENCES


ACKNOWLEDGEMENTS