Detecting past contraction in population size using haplotype homozygosity

C. Merle[1,2&3], J-M. Marin[1&3], F. Rousset[3&4] and R. Leblois[2&3]

1 - Institut de Montpeliérain 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)

Next generation genetic data

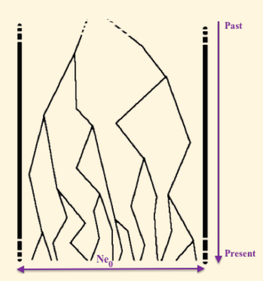
Classical inference methods of the demographic history (likelihood based methods (IS, MCMC), approximate bayesian methods and Site Frequency Spectrum), suitable for polymorphism data sets consisting in some loci and assuming the genealogies of different loci are independent, do not exploit the genetic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

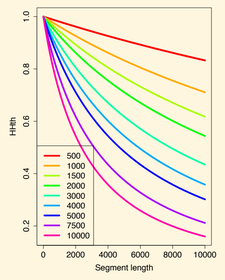
Take advantage of **genome wide data with known genome**
⇝ Consider the **dependency** of genealogies of **adjacent positions** in the genome.

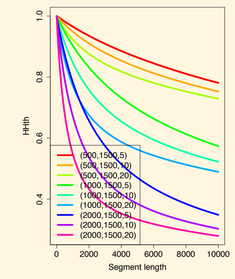
Pairwise alignment
GT CATGAGCCGA GT
GA CATGAGCCGA TT

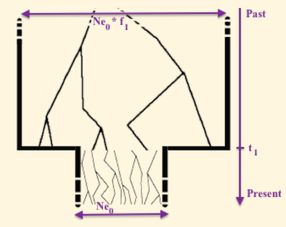
New inference approaches of demographic history based on the **conserved sequence lengths** in the pairwise alignment within a diploid genome : [3], [5], [4], [2], [1].

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.









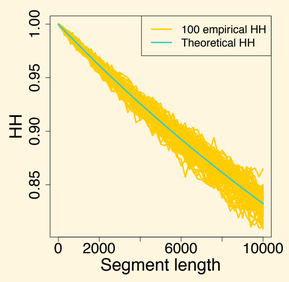
$\mathcal{M}_0$: Constant population size $\theta_0 = (Ne_0)$ $\mathcal{M}_1$: Contracting population size $\theta_1 = (Ne_0, t_1, f_1)$

Demographic history inference from genome wide sequence data

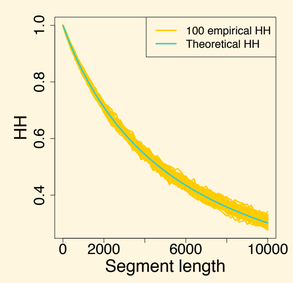
Haplotype Homozygosity: $HH(i)$: probability for i adjacent markers drown at random in the whole genome sequence to be homozygote.

Theoretical HH of [3], denoted HH_{th} : coalescent based computation, assuming the mutation and recombination rate known and constant along the genome.

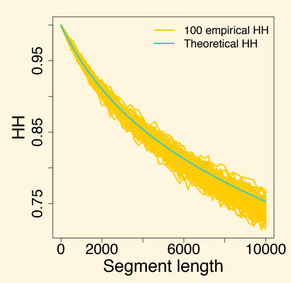
Empirical HH, denoted $\widehat{HH}$ computed from the observed data, is the segment proportion of at least i homozygotes adjacent markers.



$\mathcal{M}_0$: $\theta_0 = 500$



$\mathcal{M}_0$: $\theta_0 = 5000$



$\mathcal{M}_1$: $\theta_1 = (500, 1500, 10)$

We estimate θ_0 and θ_1 by :

$$\widehat{\theta_0} \in \arg\min_{\theta_0} \sum_{i \in \mathcal{I}} \left(\frac{HH_{\text{th}}(\theta_0, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

$$\widehat{\theta_1} \in \arg\min_{\theta_1} \sum_{i \in \mathcal{I}} \left(\frac{HH_{\text{th}}(\theta_1, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

Evaluation of $i \mapsto HH_{\text{th}}(\theta_1, i)$ is time consuming ⇝ optimization with the *blackbox* R package.

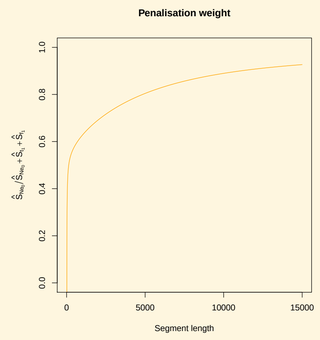
Model choice criterion: embedded models

Penalized mean square criterion

$$\arg\min_{j \in \{0,1\}} \sum_{i \in \mathcal{I}} w_j(i) \left(\frac{HH_{\text{th}}(\widehat{\theta_j}, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

$$w_0(i) = \widehat{S}^{(1)}_{Ne_0}(i) / \left(\widehat{S}^{(1)}_{Ne_0}(i) + \widehat{S}^{(1)}_{t_1}(i) + \widehat{S}^{(1)}_{f_1}(i) \right),$$

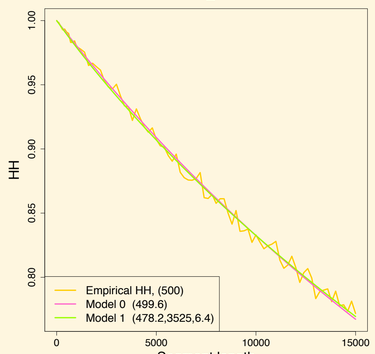
$$w_1(i) = 1.$$



$\widehat{S}^{(1)}_{\phi}(i)$: estimate of **Sobol's sensitivity index** of order one of a given parameter ϕ computed for i adjacent markers under the more complex model (*sensitivity* R package).

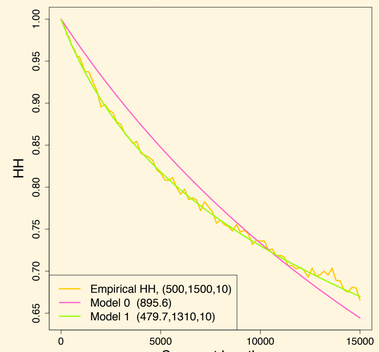
On a simulated data set with constant population size $\theta_0 = 500$, the non penalized mean square criterion would choose $\mathcal{M}_1$ whereas the penalized mean square criterion choose $\mathcal{M}_0$.
On a simulated data set with contracting population size $\theta_1 = (500, 1500, 10)$, the penalized mean square criterion choose $\mathcal{M}_1$.

Avoid choosing the more complex model



$\mathcal{M}_0$: $\theta_0 = 500$

Detecting a contraction



$\mathcal{M}_0$: $\theta_1 = (500, 1500, 10)$

Numerical results

Data sets simulated under $\mathcal{M}_0$, $\theta_0 = Ne_0 = 500$

	n_{pos}	5000		10000	
	Locus	2000	200000	2000	200000
	Data sets	100	1	100	1
TPR	Adjustment	0.74	1	0.59	0
	Penalized Adj.	0.94	1	0.98	1
	$\widehat{\overline{Ne_0}}$	494 [478 – 510]	494 –	502 [496 – 509]	501 –

Data sets simulated under $\mathcal{M}_0$, $\theta_0 = Ne_0 = 5000$

	n_{pos}	5000		10000	
	Locus	2000	200000	2000	200000
	Data sets	100	1	100	1
TPR	Adjustment	0.41	0	0.54	0
	Penalized Adj.	0.86	1	0.79	0
	$\widehat{\overline{Ne_0}}$	4998 [4960 – 5036]	5044 –	4915 [4881 – 4949]	4880 –

Data sets simulated under $\mathcal{M}_1$, $\theta_1 = (Ne_0, t_1, f_1) = (500, 1500, 10)$

	n_{pos}	10000		15000		20000	
	Locus	2000	200000	2000	200000	2000	200000
	Data sets	100	1	100	1	100	1
TPR	Penalized Adj.	0.81	1	0.92	1	0.95	1
	$\widehat{\overline{Ne_0}}$	634 [617 – 650]	633 –	561 [547 – 576]	529 –	534 [520 – 548]	492 –
	$\widehat{\overline{t_1}}$	2573 [2402 – 2744]	2445 –	2031 [1918 – 2144]	1592 –	1881 [1727 – 2036]	1446 –
	$\widehat{\overline{f_1}}$	12.4 [11.8 – 13.0]	11.5 –	12.3 [11.7 – 12.8]	9.1 –	12.2 [11.6 – 12.7]	9.49 –

Holstein data set

	n_{pos}	10000	
	Locus	2000	200000
	Data sets	100	1
CDR	Penalized Adj.	0.93	1
	$\widehat{\overline{Ne_0}}$	6690 [6533 – 6847]	6924 –
	$\widehat{\overline{t_1}}$	5965 [5559 – 6372]	6532 –
	$\widehat{\overline{f_1}}$	4.23 [3.90 – 4.57]	4.0 –

References

[1] Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97(3):404–418, 2015.
[2] Kelley Harris and Rasmus Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics*, 9(6):e1003521, 2013.
[3] I. MacLeod, T. Meuwissen, B. Hayes, and M. Goddard. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genetics research*, 91(6):413–426, 2009.
[4] I. MacLeod, D. Larkin, H. Lewin, B. Hayes, and M. Goddard. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. *Molecular Biology and Evolution*, 30(9):2209–2223, 2013.
[5] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Peer. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822, 2012.

Acknowledgements

IMAG INSTITUT MONTPELLIERAIN ALEXANDER GROTHENDIECK
Labex NUMEV
CeMEB
CBGP
INRA SCIENCE & IMPACT