



HAL
open science

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 Evolutionary 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.

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 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)

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.

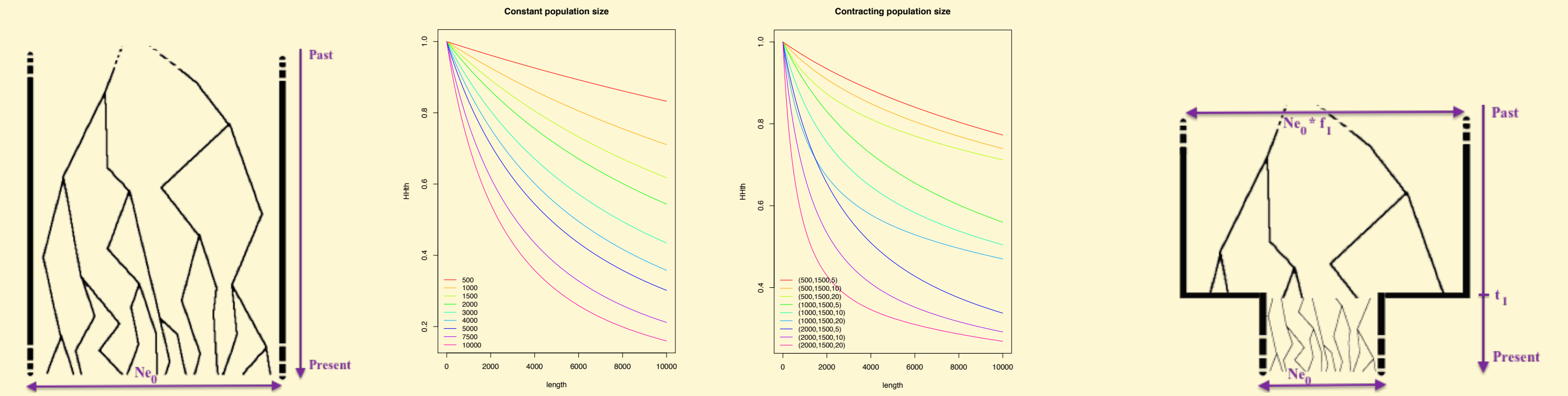
Take advantage of **genome wide data with known genome**

→ Consider the **dependency** of genealogies of **adjacent positions** in the genome.

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].

Pairwise alignment
GT CATGAGCCGA GT
GA CATGAGCCGA TT

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.



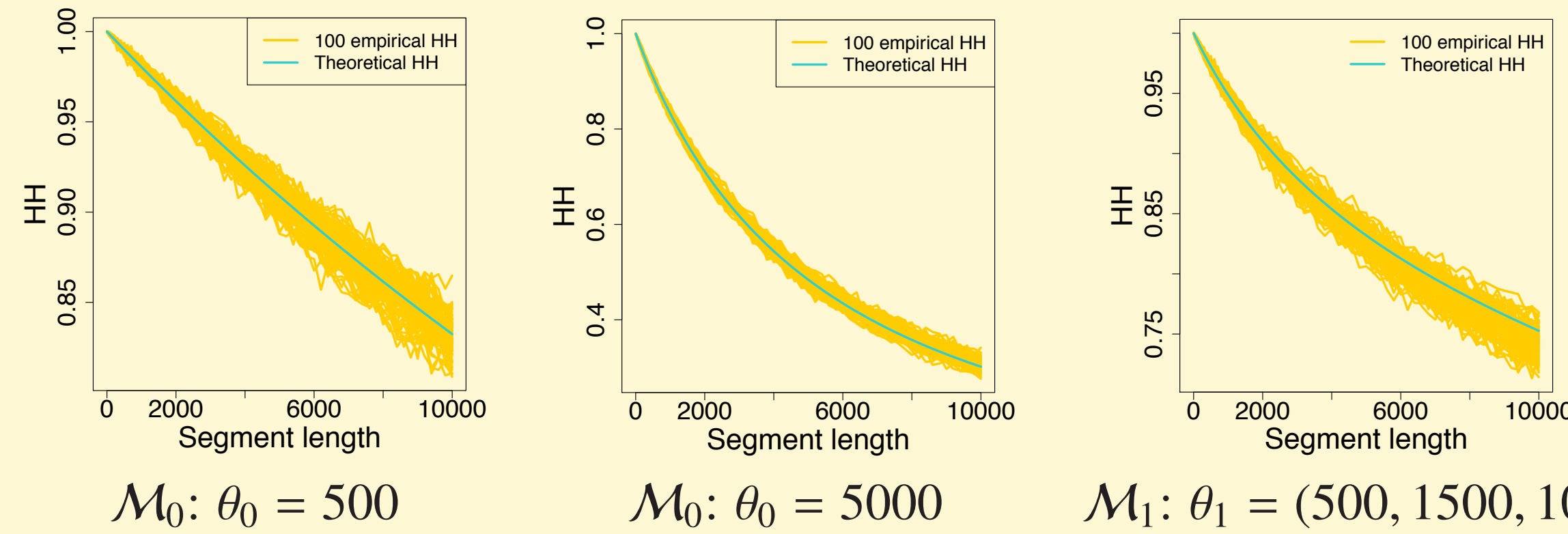
M_0 : Constant population size $\theta_0 = (Ne_0)$ M_1 : Contracting population size $\theta_1 = (Ne_0, t_1, f_1)$

DEMOGRAPHIC HISTORY INFERENCE FROM GENOME WIDE SEQUENCE DATA

Haplotype Homozygosity: $HH(i)$: probability for i adjacent markers drawn 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.



M_0 : $\theta_0 = 500$ M_0 : $\theta_0 = 5000$ M_1 : $\theta_1 = (500, 1500, 10)$

We estimate θ_0 and θ_1 by :

$$\widehat{\theta}_0 \in \arg \min_{\theta_0} \sum_{i \in \mathcal{I}} \left(\frac{HH_{th}(\theta_0, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

$$\widehat{\theta}_1 \in \arg \min_{\theta_1} \sum_{i \in \mathcal{I}} \left(\frac{HH_{th}(\theta_1, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

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

MODEL CHOICE CRITERION: EMBEDDED MODELS

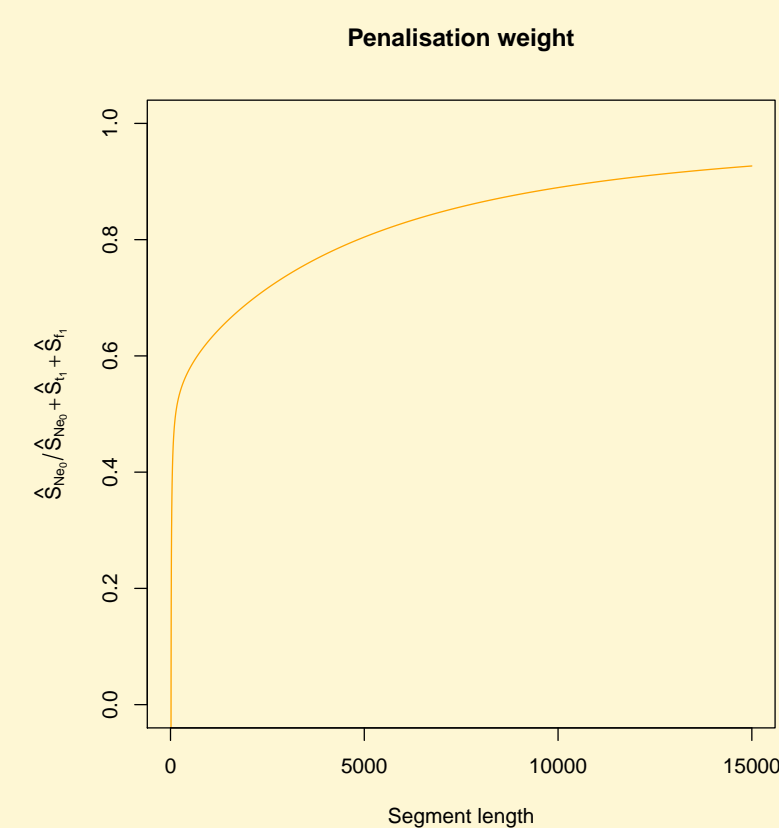
Penalized mean square criterion

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

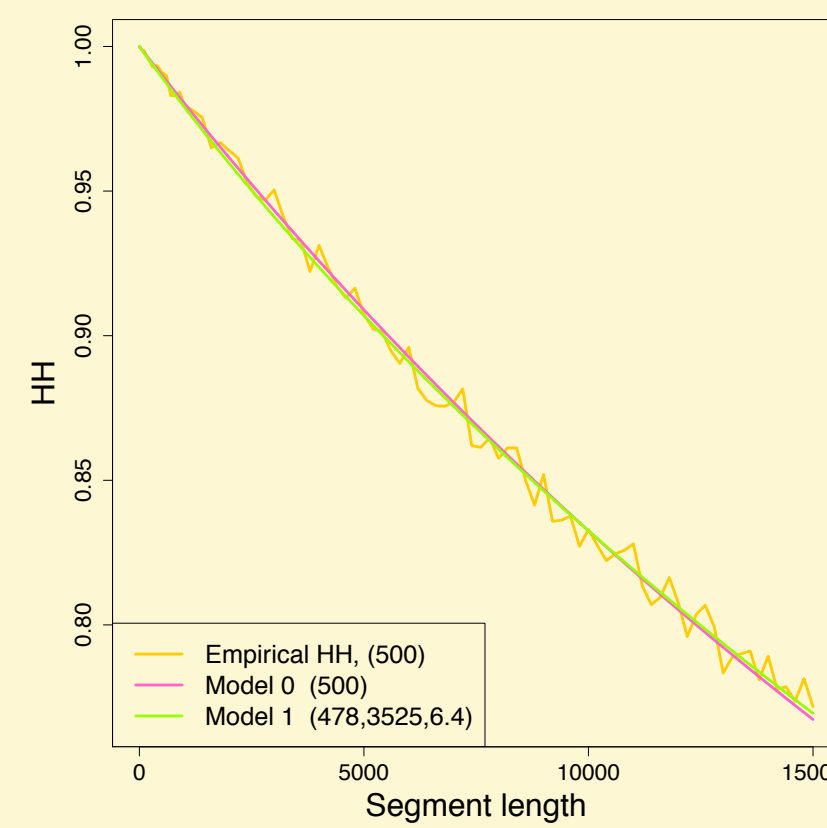
$$w_0(i) = \widehat{S}_{Ne_0}^{(1)}(i) / (\widehat{S}_{Ne_0}^{(1)}(i) + \widehat{S}_{t_1}^{(1)}(i) + \widehat{S}_{f_1}^{(1)}(i)),$$

$$w_1(i) = 1.$$

$\widehat{S}_{\phi}^{(1)}(i)$: 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).

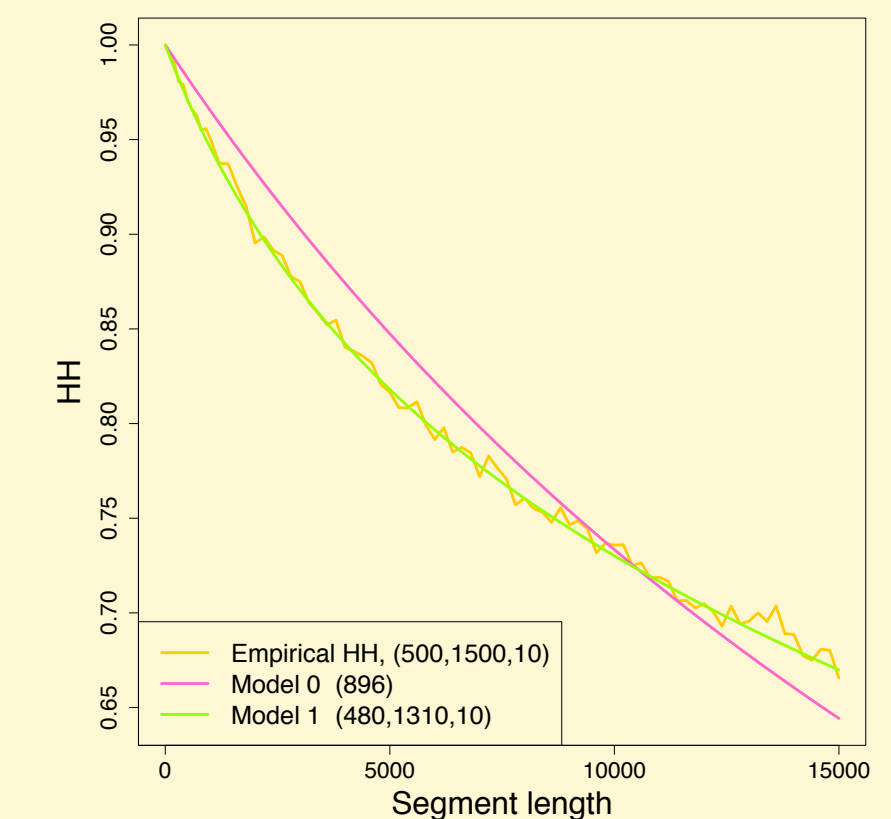


Avoid choosing the more complex model



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

Detecting a contraction



On this simulated data set with contracting population size $\theta_1 = (500, 1500, 10)$, the penalized mean square criterion choose M_1 .

NUMERICAL RESULTS

Data sets simulated under $M_0, \theta_0 = Ne_0 = 500$

n_{pos}	5000		10000	
Locus	2000	200000	2000	200000
Data sets	100	1	100	1
Adjustment	0.74	1	0.59	0
Penalized Adj.	0.94	1	0.98	1
$\widehat{\theta}_0$	494	494	502	501
$\widehat{Ne_0}$	[478 - 510]	.	[496 - 509]	-

Data sets simulated under $M_0, \theta_0 = Ne_0 = 5000$

n_{pos}	5000		10000	
Locus	2000	200000	2000	200000
Data sets	100	1	100	1
Adjustment	0.41	0	0.54	0
Penalized Adj.	0.86	1	0.79	0
$\widehat{\theta}_0$	4998	5044	4915	4880
$\widehat{Ne_0}$	[4960 - 5036]	-	[4881 - 4949]	-

Data sets simulated under $M_1, \theta_1 = (Ne_0, t_1, f_1) = (500, 1500, 10)$

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

Holstein data sets

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

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

