A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors

I. M. MacLEOD[1,2]*, T. H. E. MEUWISSEN[3], B. J. HAYES[2] AND M. E. GODDARD[1,2]
[1] *Melbourne School of Land and Environment, University of Melbourne, VIC 3010, Australia*
[2] *BioSciences Research Division, Department of Primary Industries, VIC 3083, Australia*
[3] *Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Aas, Norway*

(*Received 16 August 2009 and in revised form 16 November 2009*)

Summary

The patterns of linkage disequilibrium (LD) between dense polymorphic markers are shaped by the ancestral population history. It is therefore possible to use multilocus predictors of LD to infer past population history and to infer sharing of identical alleles in quantitative trait locus (QTL) studies. We develop a multilocus predictor of LD for pairs of haplotypes, which we term haplotype homozygosity (HH_n): the probability that any two haplotypes share a given number of n adjacent identical markers or 'runs of homozygosity'. Our method, based on simplified coalescence theory, accounts for recombination and mutation. We compare our HH_n predictions, with HH_n in simulated populations and with two published predictors of HH_n. Our method performs consistently better across a range of population parameters, including populations with a severe bottleneck followed by expansion, compared to two published methods. We demonstrate that we can predict the pattern of HH_n observed in dense single nucleotide polymorphisms (SNPs) genotyped in a cattle population, given appropriate historical changes in population size. Our method is practical for use with very large numbers of individuals and dense genome wide polymorphic DNA data. It has potential applications in inferring ancestral population history and QTL mapping studies.

1. Introduction

High density DNA markers contain information about the genetic history of a population. For instance, they can be used to infer the effective population size, recombination rate (e.g. Hudson & Kaplan, 1985; Nielsen, 2000) or the existence of recent strong selection (e.g. Sabeti *et al.*, 2002). High density markers are also used in a variety of multimarker models to estimate the probability that individuals carry the same alleles at a putative quantitative trait loci (QTLs) by exploiting the presence of linkage disequilibrium (LD) between markers and the causal mutation (e.g. Meuwissen & Goddard, 2001; Durrant *et al.*, 2004; Zollner & Pritchard, 2005; Minichiello & Durbin, 2006; Lencz *et al.*, 2007).

Some historical information can be gained by considering the markers independently (e.g. single marker homozygosity), but further information resides in the LD between markers (Nordborg & Tavare, 2002). A variety of methods exist to measure pairwise LD based on allele frequencies and frequencies of two loci haplotypes (Zhao *et al.*, 2007). However, it has been pointed out that these measures are very diverse and likely not as informative for inferring population history or for QTL mapping compared to using data from multiple markers along a segment (Nordborg & Tavare, 2002).

Although we cannot observe chromosome segment identity by descent (IBD) status directly, we can observe whether or not two haplotypes contain identical marker alleles along a particular segment. That is, they are observed to be identical by state (IBS) and this run of homozygous markers may occur through recombination. We will refer to an unbroken run of homozygous markers as 'haplotype homozygosity' (HH_n). The pairs of observed haplotypes may be in

* Corresponding author. Agriculture and Food Systems, Melbourne School of Land and Environment, University of Melbourne, VIC 3010, Australia. Tel: 03_8344 7224. e-mail: macleodi@unimelb.edu.au

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

the same or different individuals in a randomly breeding diploid population, and HH_n is predicted for observed runs of 2, 3, … to n markers, at any specified recombinant distances. Sabatti & Risch (2002) developed a multilocus measure of haplotype homozygosity in diploids and demonstrated its use to measure LD. However, their method is based on observed allele frequencies in the population sample, rather than on historical population parameters, and is computationally demanding for more than two loci. It would not be possible, therefore, to use their method to infer ancestral population history.

To infer information about population parameters from multiple marker haplotype homozygosity data or to use HH_n for QTL analysis, we need theory that adequately predicts HH_n from historical population parameters. Having developed an analytical HH_n predictor, we can then use it for example to:

1. work backwards using observed HH_n sampled from a population to predict the population parameters (e.g. Meuwissen & Goddard, 2007)
2. predict the likely sharing of alleles at a putative QTL position given the observed surrounding HH_n, using either estimated population parameters from 1 above (e.g. Meuwissen & Goddard, 2007) or assumed parameters.

A full coalescent analysis of the data would be desirable but is not practical for large numbers of markers in many individuals with recombination (e.g. Zollner & Pritchard, 2005). A practical alternative is to consider the HH_n of marker data on pairs of haplotypes which can be summarized for all possible pairs of chromosome segments which are genotyped. Hill & Weir (2007) present an algorithm for exact forward-in-time predictors of 'multilocus IBD' (their terminology for our HH_n) between two haplotypes from a randomly breeding population, based on exact two loci methodology (Weir & Cockerham, 1974). The authors point out, however, that their methodology is only applicable to haploid populations and quickly becomes cumbersome for more than four loci. An approximate method was then developed for an extended number of loci, which can be more practically applied to both haploid and diploid populations (Hill & Hernandez-Sanchez, 2007). Although this approximate method was generally a good predictor of observed HH_n in simulated populations, it becomes less reliable when mutation is included in the model (Hill & Hernandez-Sanchez, 2007).

In contrast to the forward-in-time prediction of Hill & Hernandez-Sanchez (2007), Meuwissen & Goddard (2007) published an approximate coalescence approach for the prediction of pairwise multi-marker HH_n, which they extend to estimate effective population size and to predict allele sharing status of a putative QTL position. Their methodology is an extension of a previously published method (Meuwissen & Goddard, 2001) which has proven effective for QTL mapping in farm livestock (e.g. Farnir *et al*., 2002; Olsen *et al*., 2005; Gautier *et al*., 2006). The extended methodology includes the effect of mutation at the markers, which was not previously accounted for (Meuwissen & Goddard, 2007). There are, however, still some approximations in their method and it was not developed for ancestral changes in population size (see this paper).

Effective population sizes (N) are likely to change over time; for example, some human populations are thought to have undergone a bottleneck (i.e. sharp reduction in size) followed by expansion (e.g. Pluzhnikov *et al*., 2002; Tenesa *et al*., 2007). This is likely to affect the observed homozygosity patterns, so the theory needs to predict HH_n under these conditions. In fact, it has been demonstrated with simulations that estimates of segment homozygosity for a wide range of chromosome segment lengths can be used to estimate N at multiple times in the past (Hayes *et al*., 2003).

We develop a new method to predict HH_n that aims to avoid some of the approximations of the above two methods of Hill & Hernandez-Sanchez (2007) and Meuwissen & Goddard (2007), and accommodates recombination and mutation as well as changing population size. It is based on coalescence theory, but considers multilocus data in pairs of haplotypes, summarized across all possible pairs by HH_n. We compare results from our methodology with the above two published predictors of HH_n, as well as our modified version of the Meuwissen & Goddard (2007) method. We compare these results with HH_n data from simulated populations, including comparisons where population sizes vary over time.

Our method is consistently equal to or more accurate than the above methods, given a wide range of parameters tested. Using simulated populations with and without bottlenecks, we demonstrate that depending on the density of markers, HH_n patterns can be considerably altered by ancestral changes in N. McQuillan *et al*. (2008) found that the proportion of longer runs of homozygous single nucleotide polymorphism (SNP) markers within individuals in European populations (300 000 SNP panel) is clearly related to variations in recent population history. We demonstrate that our new method (herein referred to as MHH) predicts a pattern of HH_n that matches the pattern observed in a genotyped cattle population, given appropriate estimates of changing past population size.

2. Methods

Notation used in this section is detailed in Table 1. Populations are assumed to be panmictic, diploid,

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

Table 1. *General notation presented in this section*

Symbol	Definition
T	Total number of generations of breeding in a finite constant size population
t	Any given generation counted backwards in time from the most recent generation of a population (i.e. the present day generation is $t=0$)
r	Recombinant distance between two adjacent marker loci measured in Morgans (M)
c	Recombinant length of a chromosome segment with n markers (Morgans)
N	Effective population size
m	Mutation rate per loci per generation
IBD	Identity-by-descent
R	$4Nr$
U	$4N\mu$

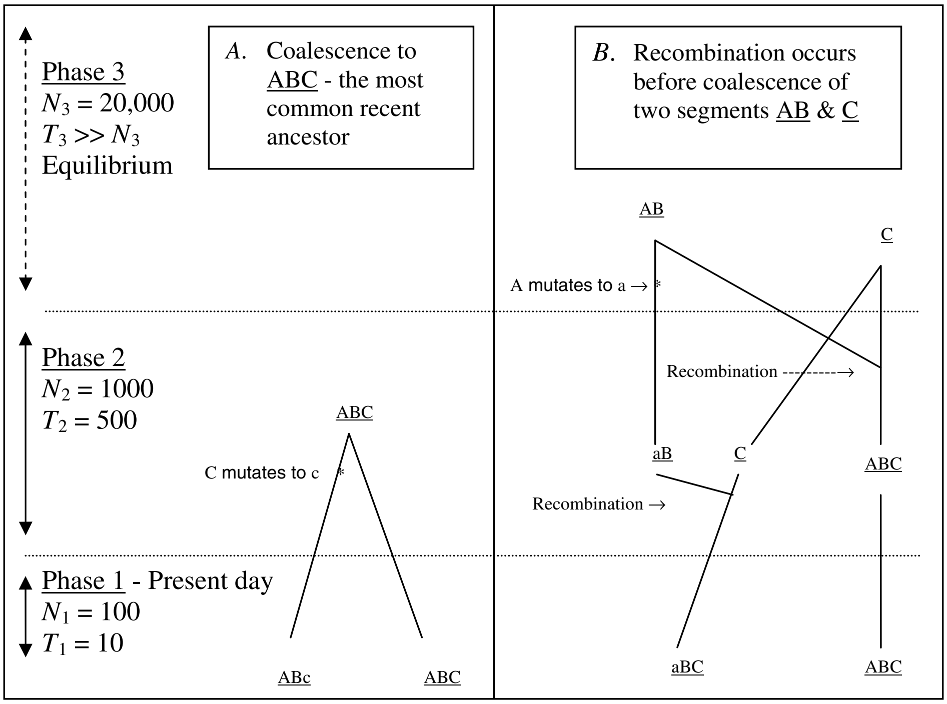

Fig. 1. Two possible coalescence pathways for a pair of chromosome segments with three markers, in a population which changes in size ancestrally (divided into three 'phases'). Tracing two DNA segments back in time, in coalescence A there is no recombination or coalescence in phase 1, and coalescence occurs in phase 2 with a mutation event on one segment. In coalescence B, both gametes recombine in phase 2 and in phase three loci '*a*' mutates, followed by coalesce of segments AB and C.

with no selfing and no selection or migration. Mutation rate is assumed equal across all alleles. The recombination rate can vary between allele pairs for all analytical methods tested, and marker haplotypes are assumed known without error. A range of parameters were tested to both represent realistic marker densities now available for QTL and population genetics studies in humans and animals, but also to test the robustness of the different methodologies to more extreme parameters similar to full genome sequence data.

HH_n can be modelled in an 'equilibrium population' with constant size and subject to the balancing forces of drift, recombination and mutation. Additionally, HH_n predictions can be extended to a dynamic population which varies in size over time and may not be in equilibrium.

(i) *New method of predicting multilocus* $\mathrm{HH_n}$ – *MHH*

MHH predicts the probability of multilocus HH_n for a pair of randomly sampled haplotypes with n markers, using a simplified coalescence approach. Multiple adjacent markers along only one pair of chromosome segments are traced backwards in time. Figure 1 outlines some possible events when tracing a pair of chromosome segments back in time, until coalescence with their most recent common ancestor, with ancestral changes in N expressed as 'phases'.

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

Assuming first a constant-sized population, the joint probability of homozygosity for all n markers (HH_n) can be split into two periods, where n is the total number of markers on a segment of recombinant distance c Morgans:

1. No coalescence, no recombination and no mutation (i.e. 'no event') for t generations. Let the probability of no event in one generation be written as

$$\alpha=\left(1-\frac{1}{2N}\right)(1-c)^2(1-\mu)^{2n}. \tag{1}$$

2. At generation $t+1$ an event takes place; either coalescence or recombination, and no mutation. The probability of coalescence at generation $t+1$ (no recombination and no mutation) is

$$\beta=\frac{1}{2N}(1-c)^2(1-\mu)^{2n}. \tag{2}$$

If the event is recombination, we calculate the probability of recombination in each possible interval between two adjacent markers k and $k+1$ and include the joint probability that all markers on both recombined segments will then coalesce or recombine without mutation. For more than two markers on a segment, we have to sum all possible recombination probabilities for each interval between adjacent markers k and $k+1$. Additionally, we trace back the $HH_{1\text{ to }k}$ and $HH_{k+1\text{ to }n}$ probabilities of the recombined segments, therefore sequentially computing haplotype homozygosity probabilities first for all two loci segments, then three loci, to four, etc. The probability of recombination in any given marker interval, including the joint HH of the recombined segments, is

$$\tau\approx\sum_{k=1}^{n-1}\left\{(1-(1-r_{k,k+1})^2)(1-\mu)^{2n}\left(1-\frac{1}{2N}\right)\right.$$
$$\left.[(HH_{1\text{ to }k})(HH_{k+1\text{ to }n})]\right\}, \tag{3}$$

where $r_{k,k+1j}$ is the recombinant distance between marker k and $k+1$. We ignore the possibility of more than one recombination per segment within one generation.

Combining the two periods above gives the total probability of marker haplotype homozygosity as

$$HH_n=\left[\sum_{t=0}^{T-1}(\alpha)^t\right](\beta+\tau). \tag{4}$$

To accommodate changing effective population sizes going back in time, we model the probabilities of no event followed by an event within each of the different historical population sizes, which we refer to as 'Phases'. Figure 1 shows a population, where effective size (N) has reduced from very large ancestrally to small in the present day, and the changes are attributed to three major historical phases (1, 2 and 3 going back in time). We calculate the probability of no event for t generations, followed by the event of coalescence or recombination occurring within any given ith phase of variable population size, and then sum the probabilities across all phases ($HH_{\text{All Phases}}$ now dropping the 'n' subscript for simplicity).

The probability of no event occurring for any given number of t_1 generations followed by an event within phase 1 is as before:

$$HH_{\text{Phase }1}=\left[\sum_{t=0}^{T_1-1}(\alpha_1)^t\right](\beta_1+\tau_1)=\left(\frac{1-\alpha_1^{T_1}}{1-\alpha_1}\right)(\beta_1+\tau_1), \tag{5}$$

where subscript '1' refers to phase 1 (present day) population parameters.

For each ith phase going back in time (before present), we calculate the probability of no event for t generations in phase i, and the probability that no event had occurred in any of the more recent phases:

$$HH_{\text{All Phases}}=\{HH_{\text{Phase }1}\}+$$
$$\left[\sum_{i=2}^{\text{most ancestral Phase}}\left[\prod_{h=1}^{i-1}(\alpha_h)^{T_h}\right]\left(\frac{1-\alpha_i^{T_i}}{1-\alpha_i}\right)(\beta_i+\tau_i)\right], \tag{6}$$

where subscript i refers to a particular phase counting from the present day, with a stable population size N_i, h designates phases more recent to phase i, and T_i or T_h is the total number of generations in phase i or h. The joint HH of the two recombined segments ($HH_{1,k}{*}HH_{k+1,n}$) in 'τ_i' above, should be traced back in time from the generation in which recombination took place within a given phase. To simplify this computationally, we approximate by assuming recombination always takes place in the first generation of a given phase. Any associated error can be minimized by splitting a long phase into a number of shorter phases (see section 3). To avoid this approximation, each phase can be reduced to a single generation, although computing time may ultimately enforce practical limitations.

(ii) *Meuwissen & Goddard method – MG*

The method developed by Meuwissen & Goddard (2007) (hereafter called 'MG') predicts the probability of observed markers on a pair of haplotypes being homozygous. Their model is also a simplified coalescence approach that assumes that all haplotype pairs eventually coalesce or recombine if traced an infinite number of generations back in time. Their probability of observing marker homozygosity depends on the probability of the unobserved underlying IBD pattern at and between the markers. It was

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

developed for constant-sized equilibrium populations only.

(iii) *Modifications to Meuwissen & Goddard – MGmod*

We develop a modified version of MG (MGmod) that allows for variable ancestral population sizes and includes two other minor improvements. Implementation for MGmod is given in Appendix A.

(iv) *Hill and Hernandez–Sanchez method – HHF*

The approximate HH_n predictor (tracing events forward in time) developed by Hill & Hernandez-Sanchez (2007), can be used for either equilibrium or dynamic populations with variable N. The method (hereafter called HHF) initially predicts non-IBD probability which is then used to calculate HH_n (or 'multilocus IBD'), for any number of markers on the segment (Hill & Hernandez-Sanchez, 2007). Their method predicts IBD of alleles with respect to a base population in which all alleles are considered unique, while our coalescence method considers observed IBS. The Hill & Hernandez-Sanchez (2007) definition of 'multiloci IBD' is that each pair of homozygous loci trace back to a common ancestor, but adjacent homozygous loci may coalesce in different ancestors if there has been intervening recombination. The method is based on an approximation for multilocus segments which requires all adjacent two loci non-IBD probabilities are calculated by the exact transition matrix method of Weir & Cockerham (1974). In this paper, we consider only diploid populations and include mutation, recombination and variable ancestral population size in the transition matrix calculations.

(v) *Simulated genotype data and observed patterns of* HH_n

We forward simulated Wright–Fisher randomly breeding populations of diploid individuals with no selfing. To save computational time, we simulate a given number of pre-determined marker positions on chromosome segments, with an equal probability of recombination between each adjacent marker. All alleles in the founder population are unique and simulations were designed to reach equilibrium before varying population sizes in more recent generations. Alleles have equal probability of mutation and each mutation event is recorded uniquely, giving an infinite alleles model. Variable input parameters include: effective population size, number of generations, as well as mutation and recombination rate. Simulations accommodate changing population parameters for any given number of generations, i.e. 'i phases' with different N_i, r_i and μ_i for T_i generations.

The average single and multiple marker HH_n can be calculated over any given number of replicated populations (calculating a mean standard error of HH_n estimated across replicates). Results from simulations are averaged over 1000–5000 replications, having found this number to give standard errors of less than 1% of the mean. To improve computational efficiency in simulations where population size and number of generations were very large ($N \geqslant 1000$), we scaled down the N and T, with the corresponding r and μ scaled up proportionately to maintain original R and U values, (as implemented by Hayes & Goddard, 2003 and Hoggart *et al*., 2007). This has almost no effect on the final HH_n predictions because it is the U and R values that control the coalescence of linked markers (Hudson, 1991). The HH_n results from the simulations were used to benchmark the different analytical predictions of HH_n.

(vi) *Observed* HH_n *in a panel of dense SNPs genotyped in dairy cattle*

We use a panel of dense bovine SNP genotypes to check whether or not we can use our method to predict the observed HH_n patterns in real data, given appropriate estimates of the population parameters. Individual SNP genotypes were obtained for 798, Australian Holstein–Friesian bulls with an outbred pedigree, using the Illumina® Infinium BovineSNP50 BeadChip with approximately 56 000 SNPs. Samples were screened for the proportion of missing genotypes, and animals with greater than 10% missing genotypes were removed. The SNPs were included only if they met the following criteria; call rate $>90\%$ and minimum allele frequency (MAF) $>0{\cdot}05$. SNPs with no recorded heterozygotes were excluded and any animals with genotypes incompatible with pedigree were removed. After screening, 730 out of the 798 animals were retained for the analysis. There were 38 259 SNPs that satisfied all selection criteria and only autosomal SNPs were used.

The SNPs were ordered by chromosome position using Bovine Genome Build 4·0 (http://www.ncbi.nlm.nih.gov/projects/genome/guide/cow/). For animals with $<10\%$ of missing genotypes, we imputed alleles using fastPHASE (Scheet & Stephens, 2006), having first confirmed the likely high accuracy of this by deleting and imputing a small proportion of known genotypes in our data set. According to the physical positions of the SNPs, the average spacing of the markers retained for this study across the genome is 0·066 megabase pairs (Mb).

HH_n was measured within animals to avoid inaccuracies in haplotyping or bias due to some sampled animals possibly being more highly related than the average of the population. Genotypes were checked for unbroken runs of homozygous markers,

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

and we recorded the number of adjacent homozygous SNPs, as well as the physical length of each homozygous segment. We do not have data for the recombinant distance between SNPs so we assume 1 Mb is approximately equal to the recombinant distance of 1 centi-Morgan (cM) (Arias *et al.*, 2009 estimated an average of 1 Mb = 1·25 cM in cattle).

We define HH_n in this data as the probability of n adjacent SNPs being homozygous. We calculated HH_n assuming equal recombinant distance between markers genome wide:

$$HH_n = \sum_{i=1}^{S} x_i \Big/ (SA),$$

where n is a given number of markers on a homozygous haplotype for which HH_n probability is being calculated, i is every possible overlapping segment with n markers along the entire genome, numbered 1 to S, x_i is the total number of individual animals homozygous for each segment i and A is the total number of animals genotyped.

(vii) *Estimating effective population size from observed cattle* $\mathrm{HH_n}$

We explored the feasibility of using our HH_n methodology to work backwards to estimate population size given observed HH_n in the above data. We exploit the theory that LD over shorter distances reflects more ancestral population parameters than LD at larger distances. LD on a segment size of c Morgans is most affected by the population size approximately $1/(2c)$ generations in the past, assuming a linearly changing population size (Hayes *et al.*, 2003). We used the following steps with iterations until a close fit of predicted HH_n to the observed HH_n was achieved:

1. Assume prior knowledge of the most ancestral population size.
2. Assume the recombination distances are known without error.
3. Adjust mutation rate to match single marker homozygosity observed in the data.
4. Assess the fit of predicted HH_n to observed HH_n across each segment length with 2, 3, 4, etc. markers, using the chosen parameters and reject N_i. if:

   $$|(HH_{\text{Predicted}} - HH_{\text{Observed}})| / HH_{\text{Observed}} \geqslant \delta,$$

   where δ is a predetermined threshold (we used 0·07).
5. If the data does not fit the observed HH_n pattern: start at the poor fit position with the smallest segments (c Morgans), and estimate the approximate number of generations back in time (T_i) for which the estimate of population size is incorrect ($\sim 1/(2c)$). Then either:
   (a) Introduce a new 'phase' of changed population size from this generation forward in time if HH_n is not accurately predicted for all segment sizes larger than that being considered.
   (b) Or, if only a section of the curve fits poorly then adjust population size in an existing or new phase, only over the generations for which HH_n is not well predicted.

Repeat steps 3, 4 and 5 until the predicted HH_n fits the observed HH_n pattern across all segment lengths.

3. Results

We show limited results for MGmod because this method was equal to or less accurate than MHH for all comparisons.

(i) *Equilibrium constant sized population*

The analytical methods were first compared under the assumption of a neutral Wright–Fisher constant sized equilibrium population. Studies of physical and linkage maps in cattle and humans found that on average 1·25 and 1·27 cM are equivalent to 1 Mb (Kong *et al.*, 2004; Arias *et al.*, 2009), implying that intersite recombination rates are approximately $1{\cdot}3 \times 10^{-8}$. Given that single nucleotide mutation rates are generally considered to be of a similar order (Nachman & Crowell, 2000), we start with values of r similar to μ. Figure 2*A* and *B* compare methods with identical population parameters $U = 4N\mu = 0{\cdot}4$, except for a change in the recombination rate between markers (r); $R = 4Nr = 2$ in Fig. 2*A* and $R = 0{\cdot}4$ in Fig. 2*B*. In Fig. 2*A*, all methods give results which are close to those of simulated populations with identical population parameters. In contrast, the parameters in Fig. 2*B* show that HHF considerably overestimates the probability of HH_n compared to the simulated results, particularly as the number of markers on the segment increases. In Fig. 2*A* and *B*, MHH and MG, agree closely with the simulated HH_n, across a range of segment lengths. Computationally, MHH is fast and can easily be implemented with multilocus predictions for hundreds of markers.

A further comparison of predicted and observed HH_n was made with U constant at 0·6 ($N = 15\,000$, $\mu = 1 \times 10^{-5}$), while R is 36 (i.e. $r = 60\mu$) or $R = 1{\cdot}8$. Given this N, the marker spacing with $R = 36$ is similar to that for 60 000 SNPs evenly spaced on the bovine or human genome assuming approximate genome size of 35 M (Kong *et al.*, 2004; Arias *et al.*, 2009), while $R = 1{\cdot}8$ is approximately equivalent to 1·2 million SNPs. With $R = 36$ and $r = 60\mu$, all methods agree almost exactly with the simulated

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

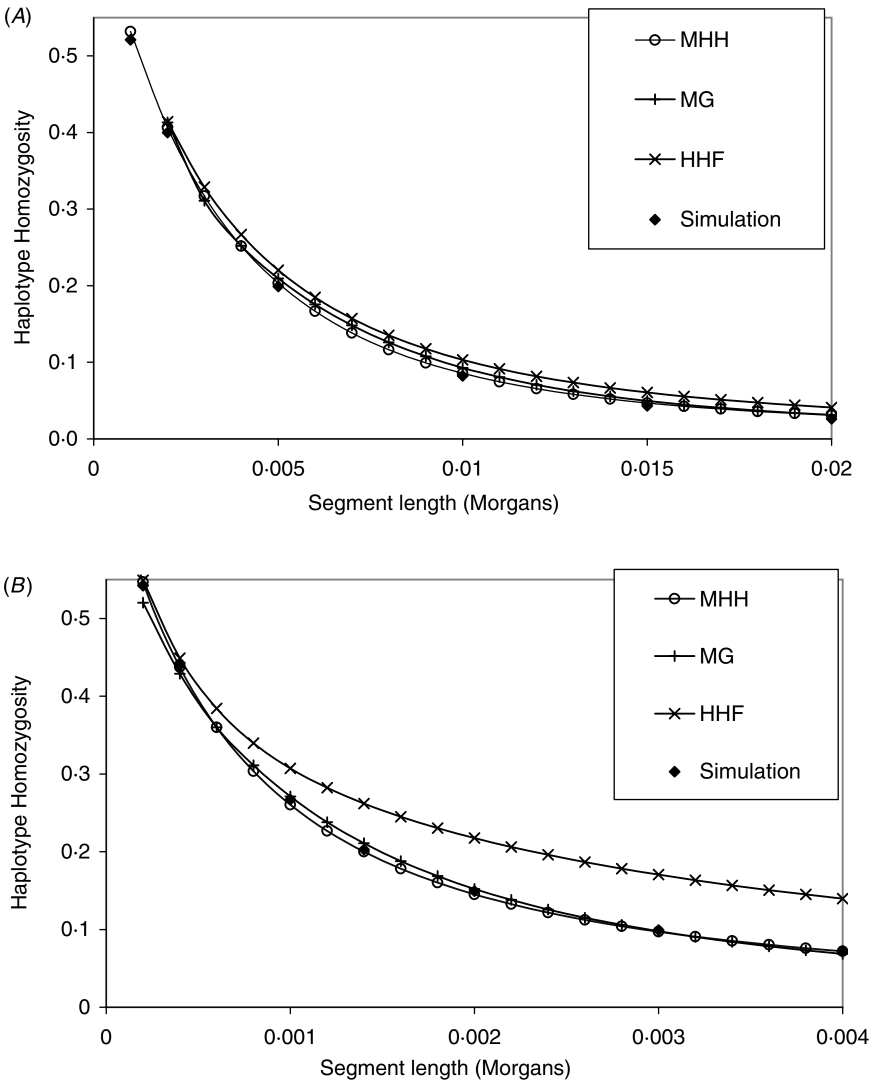

Fig. 2. (*A*, *B*) Graphs show HH_n for chromosome segments with 2, 3, 4, etc. markers, that are evenly spaced. All populations have reached a drift–recombination–mutation equilibrium, assuming $N=500$, $T=8000$, $\mu=0{\cdot}0002$. In (*A*) marker intervals (r) are 0·001 M compared to (*B*), where $r=0{\cdot}0002$. Comparisons are made between HH_n predictions using three analytical methods and observed HH_n in simulated populations (average of 2000 replicates).

populations (results not shown), but with $R=1{\cdot}8$, HHF over predicts HH_n compared to the simulation values (Table 2). HHF is sensitive to the rate of mutation relative to recombination rates, performing well when $\mu \ll r$ but less well as they approach similar values.

(ii) *Dynamic population – large ancestral size to small present day*

All methods were tested for scenarios where the population size varies over time, excluding the MG method because it was not developed for changing population sizes. For this example, parameters reflect estimated population dynamics in domestic cattle and

Table 2. *Observed (simulated data) and predicted* $\mathrm{HH_n}$ *(three methods) with population parameters;* $N=15\,000$*,* $T=300\,000$*,* $\mu=0{\cdot}00001$ *and* $r=0{\cdot}00003$

Segment length in Morgans (number of markers in brackets)	Simulation (2500 reps with S.E.M. in brackets)	MHH	MG	HHF
0·00003 (2)	0·4235 (0·00377)	0·4258	0·4125	0·4296
0·00006 (3)	0·3010 (0·00296)	0·3059	0·3087	0·3165
0·00015 (6)	0·1437 (0·00157)	0·1410	0·1476	0·1634
0·0003 (11)	0·0623 (0·00056)	0·0614	0·0637	0·0823
0·00045 (16)	0·0361 (0·00026)	0·0367	0·0365	0·0523
0·0006 (21)	0·0249 (0·00015)	0·0258	0·0250	0·0375

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

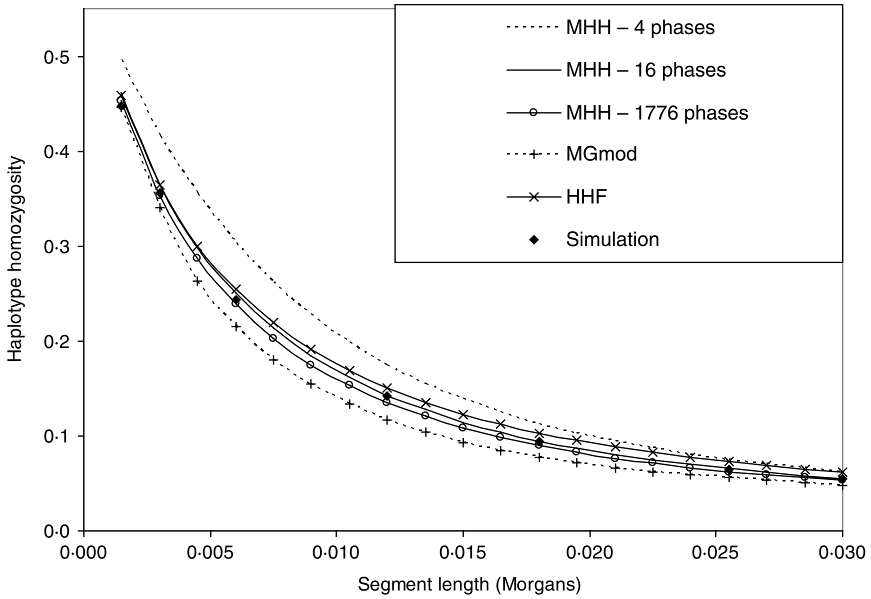

Fig. 3. Predictions of HH_n using three analytical methods compared to observed values in simulated populations (average of 5000 replicates). Population size is changing over time from very large ancestral $N=100\,000$, and gradually decreasing to present day $N=100$. ($r=0{\cdot}0015$ M and $\mu=10^{-5}$). There are four phases of differing population size. HH_n predicted by MHH is shown for these four phases, but also calculated with non-equilibrium phases split into a number of shorter phases (16 or 1776 phases total), to reduce approximation error.

their ancestors (Gautier *et al.*, 2007; de Roos *et al.*, 2008). The population is characterized by four distinct 'phases' decreasing from a very large ancestral size (in equilibrium) to very small:

Phase 4: ancestral size of 100 000 (600 000 generations) in equilibrium.
Phase 3: $N=2500$ for 1500 generations.
Phase 2: $N=300$ for 270 generations.
Phase 1: $N=100$ for five generations (from present day).

Phases 1–3 are not in equilibrium. The marker interval is 0·0015 M (i.e. approximately 20 000 markers on the bovine genome) and $\mu=1\times10^{-5}$. These parameters result in average single marker homozygosity of 0·62 in the simulations (equivalent to bovine SNP data analysed in this study).

Figure 3 shows that HHF and MGmod method lie reasonably close to the observed values of HH_n in simulated data. MHH, overestimates HH_n when calculated over four phases of changing N. This is expected due to the approximation of following the HH_n probability of recombined segments back in time from the most recent generation of each different phase of changed population size, when in fact the recombination may take place at any generation within each phase (see Methods section). We correct the overestimation by splitting longer non-equilibrium phases into an increased number of shorter phases and Fig. 3 shows two extra HH_n curves calculated with MHH using this correction (16 or 1776 phases).

The '16 phase' HH_n is calculated by splitting *Phase 2* into nine phases and *Phase 3* into five. The '1776 phase' is *Phases 1–3* split into single generation phases to eliminate approximation error. Phase 4 is not split because it is in equilibrium. Splitting up of longer phases results in HH_n prediction which is very close to that of the simulated data (Fig. 3). Comparisons with the above parameters show that when the ith phase is split into j shorter phases, once $T_j=0{\cdot}03N_i$, there was very little change in HH_n if phases were split further (Table 3). We therefore split long phases to $T_j\leqslant0{\cdot}03N_i$ for all further results with MHH.

(iii) *Large population with bottleneck and expansion*

We test whether or not there is sufficient difference in the HH_n pattern, to distinguish between populations that have undergone an ancestral bottleneck or no bottleneck, given they have the same single marker homozygosity. The population with no bottleneck is an equilibrium population; $N=50\,000$ and $T=507\,000$. The contrasting population undergoes a severe bottleneck, and then expands again to the original size:

Phase 3: $N=50\,000$ for 500 000 generations (ancestral equilibrium population).
Phase 2: $N=2500$ for 2000 generations – bottleneck.
Phase 1: $N=50\,000$ for 5000 generations (present day back in time).

To maintain single locus homozygosity of 0·62 in both populations, a mutation rate of 3×10^{-6} was used for

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

Table 3. *Observed* $\mathrm{HH_n}$ (*simulated population with 5000 replicates*) *compared with MHH predicted* $\mathrm{HH_n}$, *splitting each long phase of constant population size into an increasing number of shorter phases* (*population parameters as for Fig. 3*)

Ancestral changes in N	Simulation HH_n (S.E.M)	HH_n with long phases split into shorter phases (below)				
		4 phases	12 phases	62 phases	306 phases	1776 phases
$N_4=100\,000$, $T_4=600\,000$		1	1	1	1	1
$N_3=2500$, $T_3=1500$		1	5	30	30	1500
$N_2=300$, $T_2=270$		1	5	30	270	270
$N_1=100$, $T_1=5$		1	1	1	5	5
Segment length (M) (No. markers)						
0·0015 (2)	0·4630 (0·00261)	0·4971	0·4647	0·4549	0·4535	0·4529
0·0030 (3)	0·3698 (0·00221)	0·4170	0·3705	0·3561	0·3536	0·3531
0·0060 (5)	0·2568 (0·00170)	0·3043	0·2578	0·2421	0·2391	0·2388
0·012 (8)	0·1473 (0·00104)	0·1750	0·1478	0·1373	0·1350	0·1348
0·018 (13)	0·0974 (0·00062)	0·1129	0·0977	0·0916	0·0899	0·0899
0·0255 (18)	0·0667 (0·00039)	0·0748	0·0669	0·0637	0·0626	0·0626
0·030 (21)	0·0561 (0·00029)	0·0617	0·0562	0·0540	0·0531	0·0530

the equilibrium population, and a slightly higher rate of $5{\cdot}5\times10^{-6}$ for the bottleneck population.

Figure 4*A* contrasts observed and predicted HH_n in the bottleneck and non-bottleneck populations, with marker spacing of 0·0005 M (equivalent to approx. 70 000 SNP evenly spread across the human genome). The same contrasts are made in Fig. 4*B* except that $r=2{\cdot}5\times10^{-5}$ M (approx. 1·4 million SNP evenly spread on the human genome). The analytical predictions and observed HH_n in Fig. 4*A* are indistinguishable for both bottleneck and non-bottleneck populations. Therefore, given this marker spacing and no prior knowledge of the population ancestral histories, it would not be possible to use the analytical methods to detect the bottleneck.

In Fig. 4*B*, marker intervals are much smaller and the simulation HH_n displays a different pattern in the bottleneck compared to non-bottleneck population. In both populations, MHH predicts similar HH_n to simulated values. Although HHF is in good agreement for the non-bottleneck population, it is markedly less accurate in the bottleneck population where the value of μ is closer to r (Fig. 4*B*).

In Fig. 5, the above bottleneck population is compared to one with no large ancestral population, to test whether or not the bottleneck masks the ancestrally larger population effect on the HH_n pattern. This contrasting population had the following structure:

Phase 2: $N=2500$ for 52 000 generations – most ancestral phase – equilibrium.
Phase 1: $N=50\,000$ for 5000 generations – most recent phase.

As before, single marker homozygosity was maintained at 0·62 with $\mu=2{\cdot}64\times10^{-5}$ for the population with no large ancestral N. In both populations $r=2{\cdot}5\times10^{-5}$ M. The analytical methods predict different HH_n patterns for the two populations, but again HHF overestimates HH_n ($\mu\approx r$), while the MHH prediction is close to that observed in the simulated populations (Fig. 5).

(iv) *Prediction of observed bovine* $\mathrm{HH_n}$ *and estimation of population size*

Figure 6 shows the HH_n of observed dense SNP data from a population of dairy bulls genotyped for 38 259 SNPs, with a single marker homozygosity of 0·62. Marker spacing was relatively even, at an average of 0·066 Mb. We used MHH to predict the observed HH_n pattern, given the marker spacing and assuming 100 Mb is equivalent to one Morgan recombinant distance. It was not possible to accurately predict the observed cattle HH_n using one phase of constant population size, with single marker homozygosity of 0·62 (Fig. 6). With constant N, HH_n could be modelled moderately closely for few markers, but as the haplotype length increased to more than 12 markers on the haplotype, HH_n was considerably under predicted. To more closely model the observed HH_n, we allowed the population size to vary over time (see Method section). We assumed a starting ancestral population of 50 000 in equilibrium, because previous studies suggest a very large ancestral bovine population size (Hayes *et al.*, 2003; de Roos *et al.*, 2008).

We found that a decreasing population size (to present day) gave a much closer prediction of the observed data than a constant size (Fig. 6). In particular, modelling of HH_n was sensitive to changes in population size in most recent 760 generations (approximately $1/(2r)$). Repeated estimation of parameters,

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

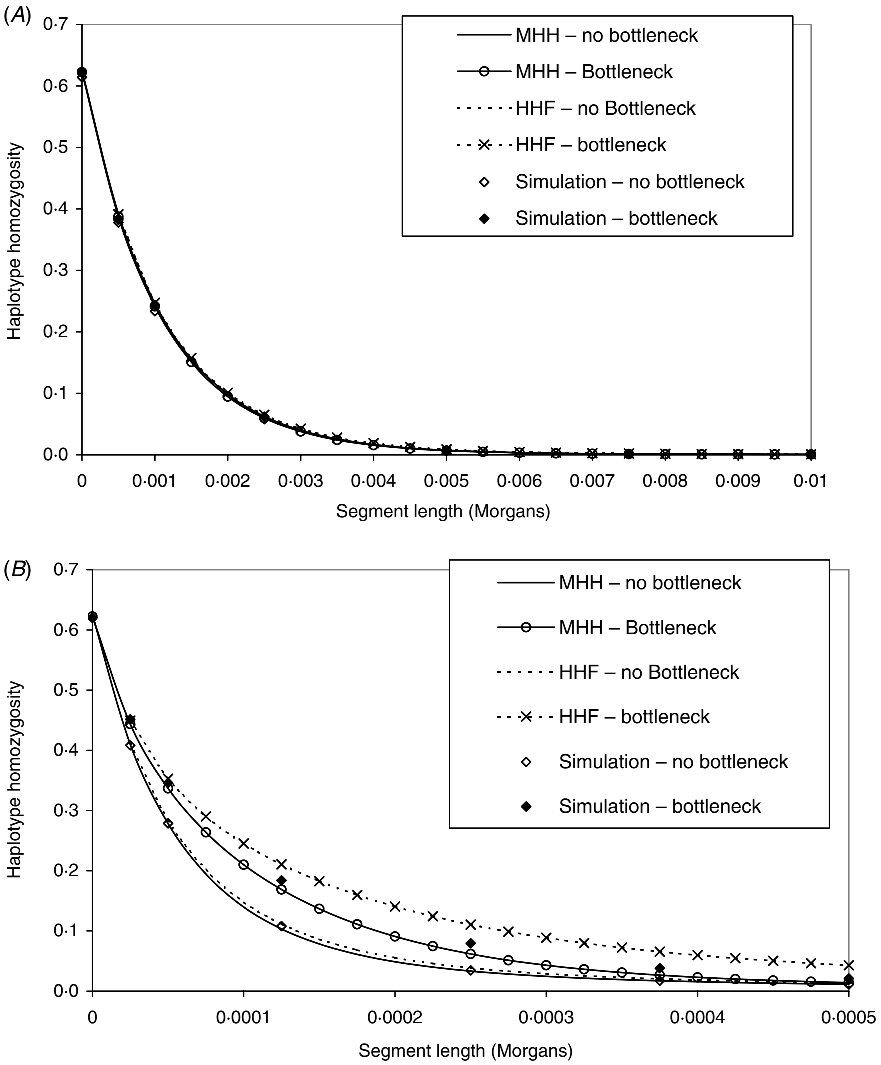

Fig. 4. (*A*, *B*) Both graphs show predictions of HH_n in two different populations, using two analytical methods compared to observed values in simulated populations (average of 1000 replicates). One population is in equilibrium with constant size ($N = 50\,000$) and a second is of the same ancestral and present size, but with a bottleneck ($N = 2500$, $T = 2000$), 5000 generations before present day. In (*A*), the marker intervals (r) are 0·0005 M, while in (*B*), $r = 2{\cdot}5 \times 10^{-5}$ M. The mutation rates in the two different populations have been adjusted to maintain the same single marker homozygosity.

which predict HH_n with a good fit to the observed bovine HH_n followed a similar pattern of decreasing population size. Estimated parameters which closely predict the observed HH_n (Fig. 6) are:

Phase 6: $N = 50\,000$, $T = 100\,000$ (equilibrium) – most ancestral population.
Phase 5: $N = 2500$, $T = 696$ (split into 12 phases of 58 generations each).
Phase 4: $N = 1000$, $T = 20$.
Phase 3: $N = 500$, $T = 18$ (split into two phases of nine generations each).
Phase 2: $N = 200$, $T = 18$ (split into three phases of six generations each).
Phase 1: $N = 80$, $T = 6$ (split into three phases of two generations each).

The mutation rate used to match the single marker homozygosity (0·62) in the observed data was $4{\cdot}7 \times 10^{-6}$. The predicted HH_n was not sensitive to the population size in the most ancestral phase; smaller or larger N in equilibrium (e.g. 10 000 or 100 000) with appropriately adjusted mutation rates gave very similar results. This is in keeping with the findings that chromosome segment homozygosity over short recombinant distances reflects effective population size more distant in the past than longer segments (Hayes *et al.*, 2003). Simplifying the above six phases to three, by averaging the most recent 40 generations to $N = 285$ and the next 718 ancestral generations to $N = 2460$, and the most ancestral to $N = 50\,000$, gives a much poorer prediction of the observed HH_n (Fig. 6).

4. Discussion

This study demonstrates that our new method (MHH) is consistently more accurate across a range

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

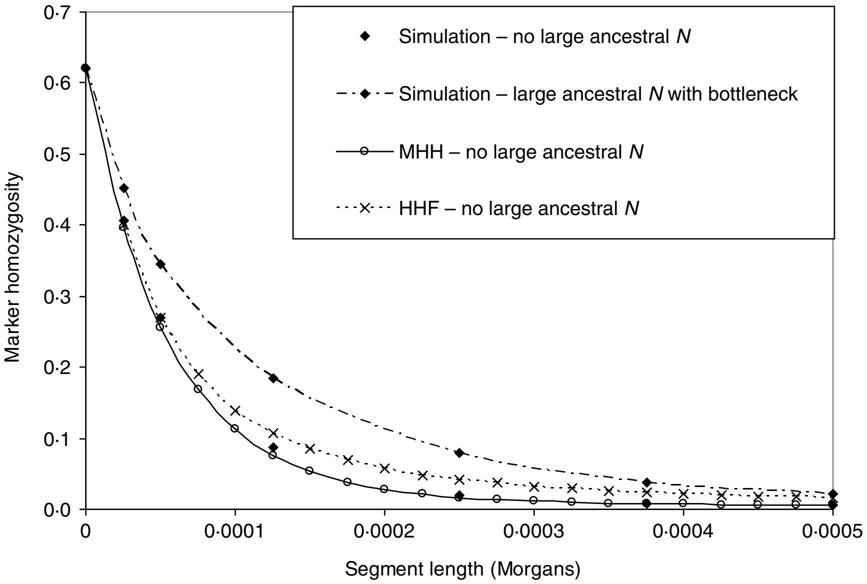

Fig. 5. HH_n predictions using two analytical methods, for an expanding population as for Fig. 4*B* ('bottleneck'), but with no large ancestral size: ancestral $N=2500$ for $T=52\,000$, followed by $N=50\,000$ for $T=5000$ to present day. Predictions are compared with observed HH_n in simulated populations (2000 replicates). Also plotted is the observed HH_n from the bottleneck population in Fig. 4*B* (i.e. with large ancestral size $=50\,000$ and bottleneck $N=2500$ for $T=2000$), to demonstrate that the bottleneck does not mask the more ancestral population size effect on HH_n.

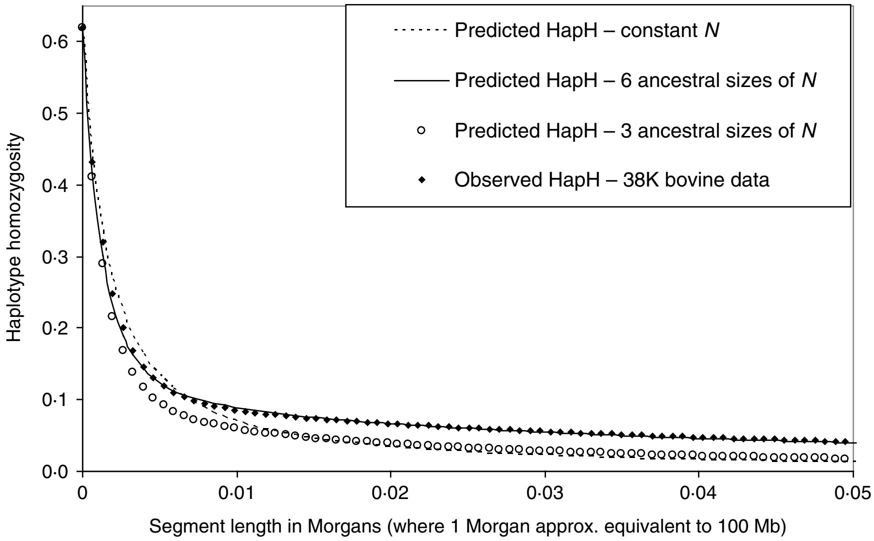

Fig. 6. Observed genome wide HH_n in cattle, genotyped for 38 259 SNPs (38 K data), compared with MHH predicted HH_n. Predictions are shown for constant population size ($N=120$), as well as sharply decreasing population size with three or six ancestral sizes (from $N=50\,000$ to $N<300$).

of input parameters (N, T, r and μ), than the other methods tested, particularly when recombination and mutation rates were of similar value. We were also able to predict the observed pattern of HH_n in a cattle population, given appropriate estimates of historical population sizes changing over time. When accommodating changes in population size, our method approximates the HH_n of recombinant segments from the most recent generation of each phase, causing some overestimation of HH_n. However, we demonstrate that this is overcome by splitting long non-equilibrium phases into a number of shorter phases. There are other ways to overcome this approximation, and all are likely to increase the computing time; ours resulted in a linear increase in user CPU time for every extra phase.

The theory for HH_n developed here applies directly to the situation where sites in the genome are defined without knowledge of whether or not they are polymorphic. This would be the case if the data were genome sequence data where all sites are recorded. In this case, the mutation rate and recombination rate

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

are both about 10^{-8} per base (Nachman & Crowell, 2000; Kong *et al.*, 2004; Arias *et al.*, 2009). Under coalescence theory for the standard neutral model, it is the *R* and *U* parameters that are critical in determining the coalescence patterns (e.g. Przeworski *et al.*, 2000). Therefore our results with the mutation rate and the recombination rate of 10^{-5} could apply to sequence data but with *N* increased by 1000-fold. However, our results would also apply approximately in the following situation. Consider a sequence of 1000 bases as a single 'locus' with a mutation rate 1000 times the per base mutation rate. Any mutation that occurs in any of these 1000 bases is recorded as a mutation in the locus. Now the mutation rate and the recombination rate will be approximately 10^{-5} per locus. In practice, the data often consist of genotypes at sites that are known to be polymorphic. This data will also approximate the data from the 1000 base locus described above because the ascertainment of polymorphisms could be described approximately as sequencing 1000 bases, for instance, and selecting one site that is polymorphic to be genotyped. In predicting HH_n at such known polymorphic SNPs, we set the mutation rate so that the observed single SNP homozygosity matches that predicted. In this way, the match of the predicted to the observed data depends on the pattern of LD and is not influenced by the match at single SNPs. The bias that use of the analytical HH_n method would introduce if the same higher mutation rates were assumed in order to predict homozygosity at a postulated QTL position given the SNP HH_n, needs further investigation.

The original MG methodology was not developed for changing population size and has three main approximations, two of which have been corrected in this paper, that work in opposing directions and therefore sometimes cancel each other out. However, both the modified and original methods still apply a third approximation; the assumption that, conditional on a recombination occurring on the chromosome segment underlying the markers, marker coalescence on the segments separated by recombination are calculated independently. This assumption ignores the fact that they share a joint coalescence up to the time the recombination takes place. The result of this approximation is particularly evident where the population size is decreasing rapidly to the present day (Fig. 3). In smaller populations, the method under-predicts the HH_n, because it does not account for the correlation in the time taken for the markers to coalesce given a recombination occurs. MHH models more accurately the multi-marker probability of coalescence before and after the segments recombine, which gives better performance under a range of parameters.

The main drawback of HHF is that marker HH_n probabilities are increasingly over predicted as the mutation rate approaches the recombination rate between markers (e.g. Fig. 2*B*). This is apparent also in Fig. 4*B* ($r=5\,\mu$), with marker density close to that already feasible in the human genome, and expanding *N* similar to human populations which are estimated to be large and expanding (Zhao *et al.*, 2006; Tenesa *et al.*, 2007). Hill & Hernandez-Sanchez (2007) give a derivation explaining why their approximate method performs less well in the presence of mutation. However, they point out that for livestock populations their method should perform well, because the influence of mutation is relatively minor due to small present day population sizes in livestock.

The parameters for the bottleneck populations were chosen to verify that the multi-marker HH_n patterns are strongly influenced by marker intervals only for a finite number of ancestral generations backwards in time from the present day. Chromosome segment homozygosity on a segment of '*c*' Morgans is influenced by the population size at approximately 1/(2*c*) generations ago, assuming a linear change in population size (Hayes *et al.*, 2003). Our results in Fig. 4*A* clearly demonstrate that no distinction was possible between the populations with and without a bottleneck 5000 generations ago when marker intervals were 0·0005 M. In this case, we would expect a two marker haplotype to be influenced by population size only up to 1000 generations ago. We did detect different patterns of HH_n in the bottleneck versus no bottleneck populations with much smaller marker intervals of $2{\cdot}5\times10^{-5}$ M (Fig. 4*B*).

Our results indicate that we should be able to predict changes in historical population sizes with an efficient search strategy to optimize the parameters given in the data. Clearly, it is not possible to predict the observed HH_n pattern in cattle when a constant population size is assumed (Fig. 6). The time period over which we can predict the changing population sizes will depend on the marker intervals. Our estimated cattle population parameters indicate a reduction in the population size from past to present (over approximately 760 generations), and follow a similar trend to estimates from studies which have used two marker LD to predict historic cattle population size (Gautier *et al.*, 2007; de Roos *et al.*, 2008). This is also in keeping with historical events which have resulted in decreasing effective population size in cattle: first due to domestication approximately 1500–2000 generations ago (assuming a generation interval of 5–7 years), then breed formation followed by breed society registration, and finally modern intense selection schemes with artificial breeding techniques (approximately 5–8 of most recent generations). There is also a possibility that recent intense selection of dairy cattle has resulted in an excess of intermediate length homozygous segments, which would result in a smaller estimate of present day *N*

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press

than the true N. However, the observed bovine SNP data used for this study covered the entire genome and is therefore less likely to be heavily biased by selection.

The consistently better performance of MHH compared to the others in this study for increasing population size and very dense markers, implies that it is a good method for use with human genotype data where effective population size has in more recent times been increasing rapidly (Zhao *et al.*, 2006; Tenesa *et al.*, 2007). With our new analytical method, it should be possible to develop formal search algorithms to predict changing population size given observed dense SNP data. We are also extending MHH to predict local recombination rates from dense SNP data, using a likelihood inference approach. A further application of our method is to predict relative probability of homozygosity at a putative QTL position, given the surrounding observed marker homozygosity (Meuwissen & Goddard, 2007), or similarly to impute missing marker information.

Appendix A

MGmod – Modifications to MG

Our main modification to MG is to allow for implementation with changes in ancestral population sizes, in a similar way as described for our method in this paper (i.e. considering multiple 'phases' of constant population size). Our modification of MG follows the same framework and notation as the Meuwissen & Goddard (2007) published method, where the probability of homozygosity at the markers is calculated by considering the conditional marker homozygosity given the underlying segment IBD pattern, and the prior probability of the IBD pattern:

$$P(y)=\sum_{\text{all}\,\pi} P(y|\pi)^{*}P(\pi). \tag{A1}$$

For $P(\pi)$, we calculate the probability of segment coalescence in each 'phase' of a given population size and finite number of generations, and sum the probabilities across all phases for the full probability of coalescence. For the following segment pattern, we would first calculate each term on the right-hand side, across all three phases. $P(\pi=[0\ 1\ 0])=P(\pi=[.\ 1\ .])-P(\pi=[1\ 1\ .])-P(\pi=[.\ 1\ 1])+P(\pi=[1\ 1\ 1])$. Likewise, we also sum $P(y=1|\pi)$, over each phase of changed population size, for each possible IBD pattern. These multi-phase conditional and prior probabilities are then substituted in the above equation (A1).

Our second modification to the Meuwissen and Goddard method (Meuwissen & Goddard, 2007) is to replace the single conditional probability of a marker being homozygous (i.e. no mutation before coalescence) given it is located on an IBD segment, with three conditional probabilities: depending on whether or not the IBD segment is bounded by recombination on 0, 1 or 2 sides. If an IBD segment is known to be bounded by recombination, the probability of mutation on this segment is higher than if unbounded because shorter IBD segments indicate longer coalescence times and more time for mutation to occur.

The derivations are therefore written in full for the modifications and the three conditional marker probabilities are:

(a) Within or at the end of an IBD segment, which is unbounded by any known recombination; $P(y=1|\pi=[1_n])=P(y=1\ \&\ \pi=[1_n])/P(\pi=[1_n])$, where $\pi=[1_n]$ in n IBD adjacent intervals.

(b) Within or at the end of an IBD segment, which is bounded on one side by recombination; $P(y=1|\pi=[0\ \ 1])=P(y=1\ \ \&\ \ \pi=[0\ \ 1])/P(\pi=[0\ 1])$, where $P(y=1\ \&\ \pi=[0\ 1])=P(y=1\ \&\ \pi=[.\ 1])-P(y=1\ \&\ \pi=[1\ 1])$ and $P(\pi=[0\ 1])=P(\pi=[.\ 1])-P(\pi=[1\ 1])$.

(c) Within or at the end of an IBD segment, which is bounded on both sides by recombination; $P(y=1|\pi=[0\ \ 1\ \ 0])=P(y=1\ \ \&\ \ \pi=[0\ \ 1\ \ 0])/P(\pi=[0\ 1\ 0])$, where $P(\pi=[\ \ldots 0\ \ldots])$ is calculated in terms of IBD segments, thus: $P(\pi=[0\ 1\ 0])=P(\pi=[.\ \ 1\ \ .])-P(\pi=[1\ \ 1\ \ .])-P(\pi=[.\ \ 1\ \ 1])+P(\pi=[1\ 1\ 1])$ and likewise, $P(y=1\ \&\ \pi=[0\ 1\ 0])=P(y=1\ \ \&\ \ \pi=[.\ \ 1\ \ .])-P(y=1\ \ \&\ \ \pi=[1\ \ 1\ \ .])-P(y=1\ \&\ \pi=[.\ 1\ 1])+P(y=1\ \&\ \pi=[1\ 1\ 1])$.

Our third modification is to calculate the joint probability of no mutation for all markers on one IBD segment. In the original method, the probability of mutation when more than one marker is found on an IBD segment is considered independent for each marker. However, this probability is correlated because the markers occur on the same IBD segment. We have altered the method to calculate the joint probability of no mutation for all markers occurring on the same IBD segment:

$$P(y=(1)^{n}|\pi=(1)^{n-1})$$
$$=\frac{1}{2N}(1-c)^{2}*\sum_{t=0}^{T-1}\left[\left(1-\frac{1}{2N}\right)(1-c)^{2}(1-\mu)^{2n}\right]^{t},$$

where n is the number of markers on the IBD segment, t is any given generation, T is the total number of generations and c is the total recombinant distance of the IBD segment (Morgans).

I.M. gratefully acknowledges scholarship support from Melbourne University, Meat and Livestock Australia and Department of Primary Industries Nancy Millis award. M.G. acknowledges support from Australian Research Council project DP0770096. The authors are grateful to C.P. Van Tassell and T.S. Sonstegard for provision of the bovine SNP data through a collaborative agreement with the Bovine Functional Genomics Laboratory, USDA-ARS.

References

Arias, J., Keehan, M., Fisher, P., Coppieters, W. & Spelman, R. (2009). A high density linkage map of the bovine genome. *BMC Genetics* **10**, 18.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J. & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics* **179**, 1503–1512.

Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P. & Morris, A. P. (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *The American Journal of Human Genetics* **75**, 35–43.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M., Moisio, S., Simon, P., Wagenaar, D., Vilkki, J. & Georges, M. (2002). Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* **161**, 275–287.

Gautier, M., Barcelona, R. R., Fritz, S., Grohs, C., Druet, T., Boichard, D., Eggen, A. & Meuwissen, T. H. (2006). Fine mapping and physical characterization of two linked quantitative trait loci affecting milk fat yield in dairy cattle on BTA26. *Genetics* **172**, 425–436.

Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Foglio, M., Grohs, C., Boland, A., Garnier, J.-G., Boichard, D., Lathrop, G. M., Gut, I. G. & Eggen, A. (2007). Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* **177**, 1059–1070.

Hayes, B. J. & Goddard, M. E. (2003). Evaluation of marker assisted selection in pig enterprises. *Livestock Production Science* **81**, 197–211.

Hayes, B. J., Visscher, P. M., McPartlan, H. C. & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* **13**, 635–643.

Hill, W. G. & Hernandez-Sanchez, J. (2007). Prediction of multilocus identity-by-descent. *Genetics* **176**, 2307–2315.

Hill, W. G. & Weir, B. S. (2007). Prediction of multi-locus inbreeding coefficients and relation to linkage disequilibrium in random mating populations. *Theoretical Population Biology* **72**, 179–185.

Hoggart, C. J., Chadeau-Hyam, M., Clark, T. G., Lampariello, R., Whittaker, J. C., De Iorio, M. & Balding, D. J. (2007). Sequence-level population simulations over large genomic regions. *Genetics* **177**, 1725–1731.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (ed. D. Futuyma & J. Antonovics), pp. 1–44. New York: Oxford University Press.

Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Kong, X., Murphy, K., Raj, T., He, C., White, P. S. & Matise, T. C. (2004). A combined linkage–physical map of the human genome. *The American Journal of Human Genetics* **75**, 1143–1148.

Lencz, T., Lambert, C., Derosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M., Kucherlapati, R. & Malhotra, A. K. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences of the USA* **104**, 19942–19947.

Mcquillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., Macleod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., Campbell, H. & Wilson, J. F. (2008). Runs of homozygosity in European populations. *The American Journal of Human Genetics* **83**, 359–372.

Meuwissen, T. H. & Goddard, M. E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genetics, Selection, Evolution* **33**, 605–634.

Meuwissen, T. H. E. & Goddard, M. (2007). Multipoint IBD prediction using dense markers to map QTL and estimate effective population size. *Genetics* **176**, 2551–2560.

Minichiello, M. J. & Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics* **79**, 910–922.

Nachman, M. W. & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942.

Nordborg, M. & Tavare, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**, 90.

Olsen, H. G., Lien, S., Gautier, M., Nilsen, H., Roseth, A., Berg, P. R., Sundsaasen, K. K., Svendsen, M. & Meuwissen, T. H. (2005). Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6. *Genetics* **169**, 275–283.

Pluzhnikov, A., Di Rienzo, A. & Hudson, R. R. (2002). Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* **161**, 1209–1218.

Przeworski, M., Hudson, R. R. & Di Rienzo, A. (2000). Adjusting the focus on human variation. *Trends in Genetics* **16**, 296–302.

Sabatti, C. & Risch, N. (2002). Homozygosity and linkage disequilibrium. *Genetics* **160**, 1707–1719.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., Mcdonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.

Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics* **78**, 629–644.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**, 520–526.

Weir, B. S. & Cockerham, C. C. (1974). Behavior of pairs of loci in finite monoecious populations. *Theoretical Population Biology* **6**, 354.

Zhao, H., Nettleton, D. & Dekkers, J. C. M. (2007). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. *Genetical Research* **89**, 1–6.

Zhao, Z., Yu, N., Fu, Y.-X. & Li, W.-H. (2006). Nucleotide variation and haplotype diversity in a 10-kb noncoding region in three continental human populations. *Genetics* **174**, 399–409.

Zollner, S. & Pritchard, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092.

https://doi.org/10.1017/S0016672309990358 Published online by Cambridge University Press