Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors

Iona M. MacLeod,*[1,2] Denis M. Larkin,[3,4] Harris A. Lewin,[5] Ben J. Hayes,[2,6] and Mike E. Goddard[1,2]

[1]Department of Agriculture and Food Systems, Melbourne School of Land and Environment, University of Melbourne, Victoria, Australia

[2]AgriBio, Centre for AgriBioscience, BioSciences Research Division, Department of Environment and Primary Industries, Melbourne, Victoria, Australia

[3]Institute of Biological, Environmental and Rural Sciences, University of Aberystwyth, Aberystwyth, United Kingdom

[4]Department of Animal Sciences, University of Illinois at Urbana-Champaign

[5]Department of Evolution and Ecology, University of California, Davis

[6]AgriBio, Centre for AgriBioscience, La Trobe University, Melbourne, Victoria, Australia

***Corresponding author:** E-mail: macleodi@unimelb.edu.au.

Associate editor: Rasmus Nielsen

The sequences reported in this paper have been deposited in the National Center for Biotechnology Information (NCBI) dbSNP Short Genetic Variations database: http://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1055441.

Abstract

Whole-genome sequence is potentially the richest source of genetic data for inferring ancestral demography. However, full sequence also presents significant challenges to fully utilize such large data sets and to ensure that sequencing errors do not introduce bias into the inferred demography. Using whole-genome sequence data from two Holstein cattle, we demonstrate a new method to correct for bias caused by hidden errors and then infer stepwise changes in ancestral demography up to present. There was a strong upward bias in estimates of recent effective population size (N_e) if the correction method was not applied to the data, both for our method and the Li and Durbin (Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496) pairwise sequentially Markovian coalescent method. To infer demography, we use an analytical predictor of multiloci linkage disequilibrium (LD) based on a simple coalescent model that allows for changes in N_e. The LD statistic summarizes the distribution of runs of homozygosity for any given demography. We infer a best fit demography as one that predicts a match with the observed distribution of runs of homozygosity in the corrected sequence data. We use multiloci LD because it potentially holds more information about ancestral demography than pairwise LD. The inferred demography indicates a strong reduction in the N_e around 170,000 years ago, possibly related to the divergence of African and European *Bos taurus* cattle. This is followed by a further reduction coinciding with the period of cattle domestication, with N_e of between 3,500 and 6,000. The most recent reduction of N_e to approximately 100 in the Holstein breed agrees well with estimates from pedigrees. Our approach can be applied to whole-genome sequence from any diploid species and can be scaled up to use sequence from multiple individuals.

Key words: **haplotype homozygosity, next generation sequencing, linkage disequilibrium, effective population size, PSMC.**

Article

Introduction

In diploid populations, the strength and patterns of linkage disequilibrium (LD) between loci are strongly influenced by effective population size (Hill 1975). Consequently, the extent of LD in a population can be used to estimate past effective population size (N_e) (Hill 1981). Although knowledge of the ancestral demography is of interest in itself, it is also of importance, for example, when studying patterns of LD for evidence of selection (Grossman et al. 2010). The null hypothesis of no selection requires an accurate demographic model because variation in N_e can result in LD patterns that mimic selection (Pritchard and Przeworski 2001).

Although LD measured between pairs of loci, such as r^2, has been used to infer complex demography (Schaffner et al. 2005; Voight et al. 2005), multiloci measures of LD potentially capture more population genetic information and therefore have also been used to infer ancestral demography (Hayes et al. 2003; Meuwissen and Goddard 2007; Lohmueller et al. 2009; MacLeod et al. 2009). LD arises as a result of individuals in a finite population sharing chromosome segments inherited identical by descent (IBD) from a common ancestor. Longer segments arise as a result of more recent coalescent events while very short IBD segments are more likely to date back to very distant coalescent events. Therefore, the pattern of multiloci LD can be described by the distribution of the lengths of chromosome segments that are IBD, which in turn can be used to infer demography (Hayes et al. 2003).

In practice, if a pair of chromosome segments carry the same alleles at all positions we observe a pairwise "run of

homozygosity" (RoH), which is identical by state (IBS) but not always entirely IBD from a single common ancestor because recombination can be masked. Multiloci LD can therefore be summarized by the observed distribution of lengths of pairwise RoH separated by heterozygous sites. MacLeod et al. (2009) developed an analytical method to calculate the probability of observing pairwise RoH of n or more loci using simplified coalescent theory, which accounts for stepwise changes in N_e. They call this summary statistic "haplotype homozygosity" or HH_n and demonstrated that the analytical method could be exploited to infer a demography that was consistent with the observed distribution of RoH in empirical single-nucleotide polymorphism (SNP) array data. This method avoids the computational burden of Approximate Bayesian Computing approaches that simulate many replicates of genetic data (Beaumont et al. 2002). This study is the first application of the MacLeod et al. (2009) method to whole-genome sequence data.

The genome sequence of even a single individual in an outbred population provides vast numbers of pairwise RoH (up to millions), each with its own coalescent history, from which we can summarize the genome wide pattern of multiloci LD. Until recently, demographic inference studies have used either subsets of genome wide loci known to be polymorphic (such as SNP arrays) or polymorphic loci in samples of relatively short resequenced genome segments (Gronau et al. 2011). One previous study used individual whole-genome sequences from several humans to reconstruct demography, although they first condense nonoverlapping windows of 100 bp into a single "locus" defined as homozygous or heterozygous (Li and Durbin 2011). These authors apply a pairwise sequentially Markovian coalescent (PSMC) model that relies on the distribution of heterozygous sites within an individual sequence to infer historical N_e (Li and Durbin 2011). Their method appears to work well for human ancestral demographic inference, although they found estimates of human N_e in recent times were not reliable (from ~800 generations ago). We compare our inferred demography with that from the PSMC model (Li and Durbin 2011).

Importantly, in this study we introduce a new method to first correct for false-positive heterozygous errors in the sequence prior to inferring demography. These errors have a disproportionate impact on the longer RoH and even after careful quality control, a low level of false positives remains in sequence data. We demonstrate that even a low error rate can cause a serious bias in estimates of more recent N_e because the longer RoH that inform these N_e estimates are those most likely to be broken up by the false positives. Our results indicate that introducing the error correction method for false positives also considerably improves the accuracy of the PSMC estimates of recent N_e in cattle.

Using whole-genome sequence from two Holstein bulls, we applied stringent filters to remove common sequencing errors and then applied our correction method to account for residual false-positive heterozygous errors. We then inferred a stepwise pattern of changing ancestral N_e, which predicts a distribution of RoH matching that observed in the corrected sequence data. The sequenced bulls represent two key ancestors in the Holstein breed (Larkin et al. 2012): Walkway Chief Mark ("Mark") and Pawnee Farm Arlinda Chief ("Chief," the sire of Mark). We inferred the stepwise pattern of ancestral N_e using Mark's sequence, and then cross-validated the demographic model in Chief's sequence data. We also used simulated sequence data to further test the methodology (supplementary information, Supplementary Material online).

Our method would be useful for a range of outbred diploid species and can be readily scaled up to use sequence data from multiple individuals, for example, to estimate the change in N_e over time for wild animal populations.

Results

The results from the data analysis (fig. 1) are presented in three sections: error estimation and correction, observed summary statistics (HH_n) calculated from RoH, and demographic inference.

Sequence Error Correction

The sequence of both bulls was independently subjected to stringent filtering to reduce common sequencing errors, referred to henceforth as "filtered sequence." Residual error rates in the filtered sequence were then calculated for each bull: that is, remaining false-positive heterozygous SNP and false-negative missed heterozygous SNP. The numbers of heterozygous SNP detected in the filtered sequence (table 1) provided more than 1 million RoH from which to estimate HH_n in Mark. The proportion of missed heterozygous SNP positions (false negatives), estimated by comparing sequence positions matching independent 50,000 SNP array genotypes (SNP50), is higher in Chief's sequence compared with Mark's (table 1) due to the difference in average read depth (~7× for Chief and ~13× for Mark). The false-negative rate was relatively high because we used stringent filters to minimize false-positive heterozygous SNP that can cause a serious bias in the distribution of longer RoH. It was important to estimate the false-negative rate because this also affects the observed distribution of RoH, and this was corrected for by scaling the mutation rate (Materials and Methods).

The false-positive error rate in filtered sequence estimated from comparison with the SNP50 genotypes is relatively low for both bulls (table 1) and not very different to the potential error rate of 0.1% in SNP50 data (http://res.illumina.com/documents/products/datasheets/datasheet_bovine_snp50.pdf, last accessed July 23, 2013). Therefore, these estimates of false-positive rate (table 1) may be inflated by the assumption that SNP50 data is error free, and may also be imprecise because of the relatively low number of loci available for validation with SNP50 genotypes. We therefore computed a more robust estimate of the false-positive error rate using observed average sequence heterozygosity in regions matching long RoH in the SNP50 data (>10 Mb) (i.e., long runs of adjacent homozygous genotypes that were likely to be IBD regions). In filtered sequence, we found an average of one heterozygous sequence SNP per 55,118 bp in Mark and one per 87,464 bp in Chief in regions corresponding to long SNP50 RoH. This was

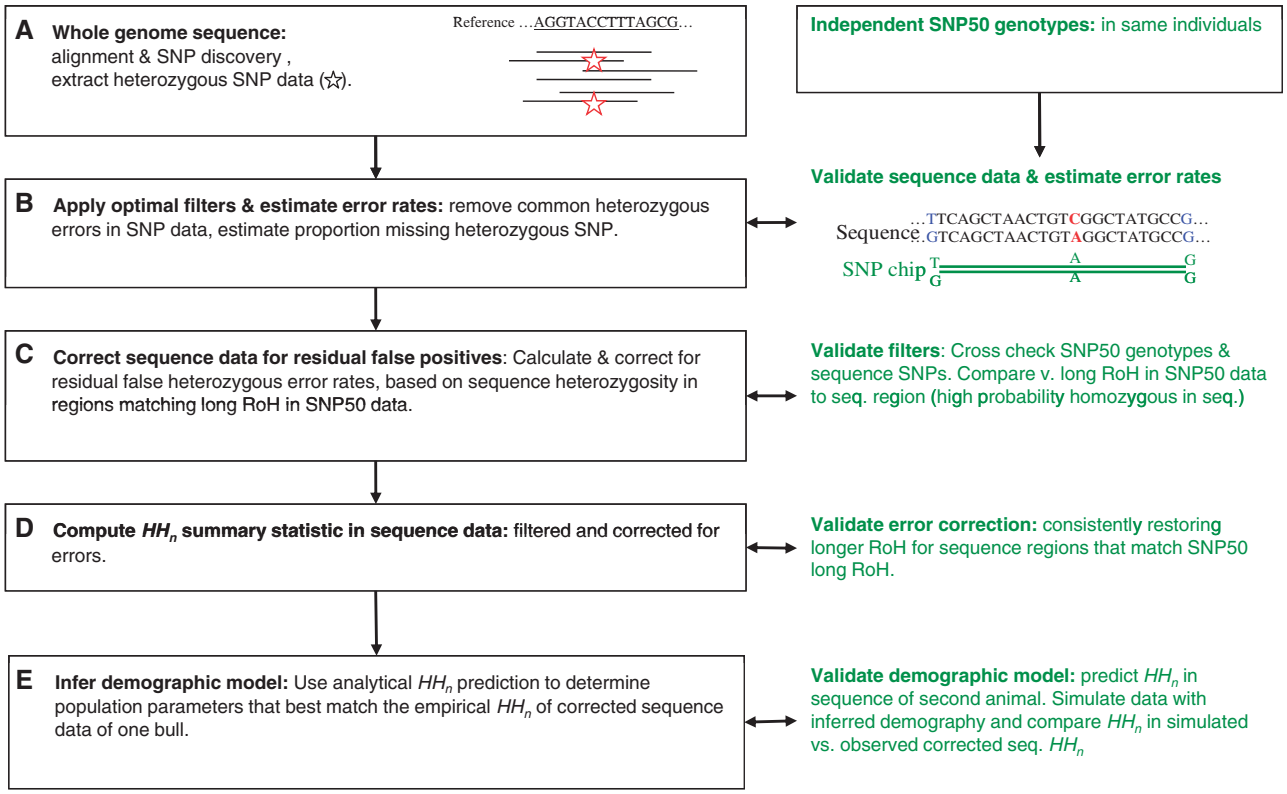

Fig. 1. The computational workflow first identified heterozygous positions within each genome sequence of two bulls (*A*). Heterozygous positions were then validated across independent SNP50 genotypes (*B*). After filtering to remove heterozygous errors, the residual false-positive rate in sequence was estimated and corrected for (*C*), and the summary LD statistic (HH_n) was calculated (*D*). The ancestral demography was inferred using an analytical model (*E*) and validated using the sequence of the second bull.

Table 1. SNP Discovered in Sequence Data and Independent SNP50 Genotypes for Mark and Chief after Stringent Filtering, and Estimation of Sequence Errors by Comparison with SNP50.

SNP Description	Mark	Chief
Number of heterozygous SNP recovered in filtered sequence data	1,243,113	757,266
Number of heterozygous SNP in SNP50 data after strict quality control	12,627	12,509
Number of homozygous SNP in the SNP50 data after strict quality control	26,299	26,493
Number of matching heterozygous SNP positions: heterozygous in SNP50 and also in filtered sequence	6,344 (50.2%)	3,839 (30.7%)
Number of false-negative errors: heterozygous in SNP50 but homozygous in filtered sequence	6,283 (49.8%)	8,670 (69.3%)
Number of false-positive errors: homozygous in SNP50 but heterozygous in filtered sequence	30 (7.7×10^{-4} of all 38,926 positions validated in SNP50)	24 (6.2×10^{-4} of all 39,002 positions validated in SNP50)

in stark contrast with the remainder of the genome which had an average of one heterozygous SNP per 1,957 bp for Mark and one per 3,214 bp for Chief. Therefore, matching the long SNP50 RoH regions to filtered sequence provided estimates of average residual false-positive error rate of 1.8×10^{-5} in Mark and 1.1×10^{-5} in Chief.

Our method of stochastically correcting for residual false-positive errors remaining in filtered sequence (see Materials and Methods) had a significant impact on restoring the longer RoH and therefore moving the distribution of RoH closer to the true distribution. Figure 2 illustrates how the long RoH are much more likely to be disrupted by errors than very short RoH, by comparing "corrected sequence" and filtered sequence in a chromosome region with a long RoH (~33 Mb) in Mark's SNP50 data. In the corrected sequence, several very long RoH are evident (fig. 2*A*) compared with filtered sequence (fig. 2*B*), which displays >100 shorter homozygous segments in this same region. In figure 2*C*, it is clear that correction of the sequence data without controlling for uniform distribution of false positives (random deletion of SNP without using nonoverlapping windows) does not uncover the long RoH. The correction method is not

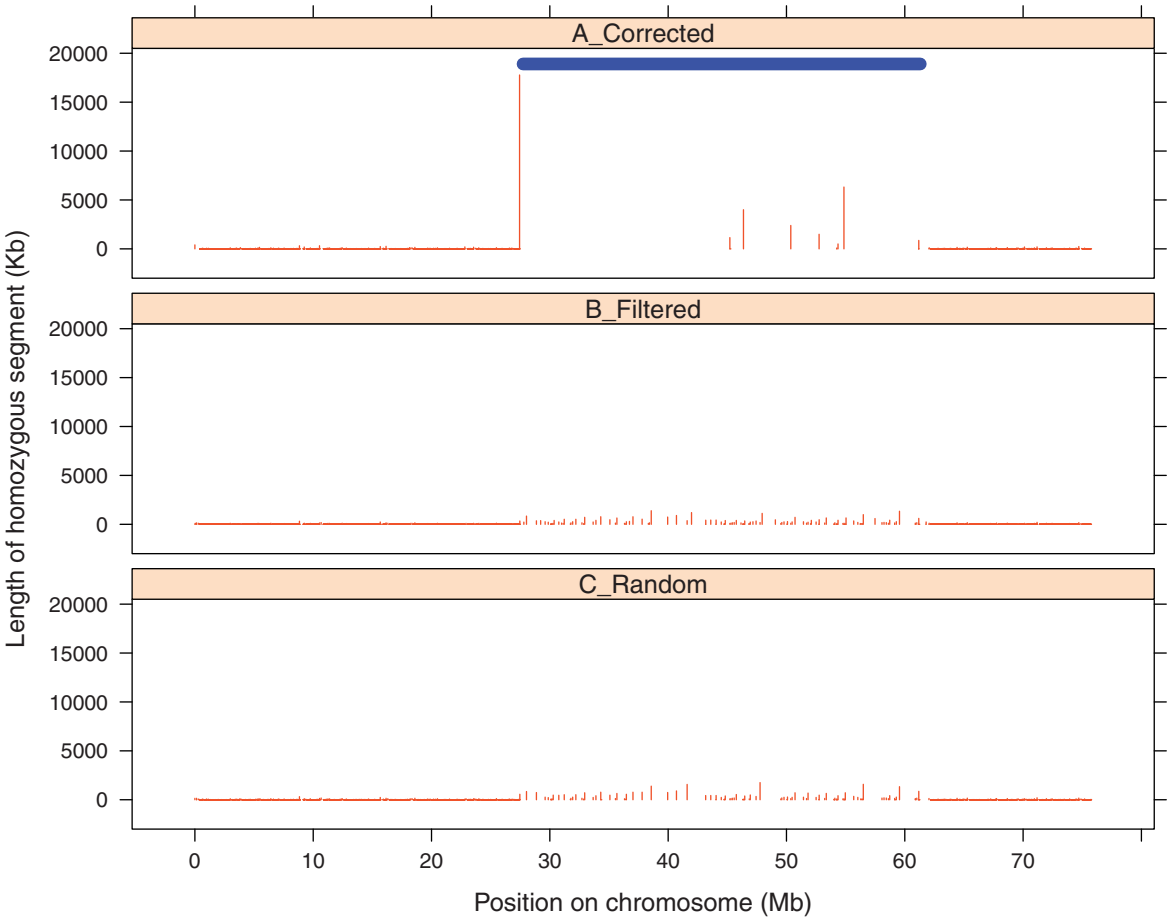

Fig. 2. The blue bar indicates the position of a long RoH in Mark's independent SNP50 genotype data. Position and length of runs of homozygosity (RoH) in Mark's stochastically corrected sequence data (*A*) and filtered only data (*B*). In (*C*), the same proportion of heterozygous errors was randomly deleted from the filtered data as for corrected data (*A*), but without enforcing uniform deletion from nonoverlapping windows. This demonstrates that our correction method effectively unmasks long RoH in the sequence data (*A*). Very much shorter RoH are observed in the regions flanking the SNP50 RoH even in the corrected data, because these are rarely affected by the low level of residual errors.

therefore randomly creating long homozygous runs, but rather is unmasking those that were previously hidden by a low level of false-positive errors. In the regions flanking the SNP50 RoH, there is little discernable difference in corrected compared with filtered sequence because these very much shorter RoH are rarely affected by the low error rates in filtered sequence (fig. 2). Similar patterns as seen in figure 2 were observed across the genome for both bulls for all regions with long RoH in SNP50 data (Mark's chromosomes 1 to 10 shown in supplementary figs. S5, S6, and S7, Supplementary Material online). After correction for false-positive errors there remained 1,198,677 RoH for Mark and 728,059 for Chief, referred to henceforth as "corrected sequence."

Table 2 gives estimates of single base heterozygosity for each bull. The observed single base heterozygosity rates are very different because Chief was sequenced at about half the read depth of Mark resulting in divergent false-negative error rates. However, the estimates of true single base heterozygosity were very similar for both bulls. This is important because it lends credibility to the independently estimated error rates and data correction for each bull. Estimated true heterozygosity indicates that on average we would expect to find one heterozygous base pair every 1,070 bp: that is, approximately 2.47 million heterozygous SNP across all autosomes assuming a total length of 2,545,896,661 bp based on the Btau 4.0 reference genome (http://www.ncbi.nlm.nih.gov/assembly/GCF_000003205.2/, last accessed July 23, 2013).

Table 2. Observed Single Base Heterozygosity in Filtered Sequence, Sequence Corrected for Residual False-Positive Errors, and Predicted Estimates of True Single Base Heterozygosity for Sequence in Both Bulls (i.e., with No False Negatives).

	Mark	Chief
Filtered sequence heterozygosity	4.883×10^{-4}	2.974×10^{-4}
Corrected sequence heterozygosity	4.708×10^{-4}	2.860×10^{-4}
Estimated true heterozygosity[a]	9.371×10^{-4}	9.319×10^{-4}

[a]True heterozygosity was estimated from observed heterozygosity, given an autosome length of 2,545,896,661 bp and detection rate of true heterozygous SNP by validation with the SNP50 genotypes (table 1).

Distribution of RoH–HH_n

The HH_n summary statistic (MacLeod et al. 2009) is defined here as the probability that any pair of homologous

chromosomes are observed IBS for at least n base pairs to the right of a site chosen uniformly at random. The HH_n summary statistic was independently estimated in each bull from the observed RoH in filtered sequence and also in corrected sequence. To test the precision of our stochastic correction method, used to restore the distribution of RoH closer to the true distribution, we replicated the data correction 25 times for each bull sequence and calculated HH_n in each replicate. The coefficient of variation of HH_n across the 25 replicated data sets increased slightly with the size of segment (n), but was never greater than 0.3% in Chief, 0.2% in Mark. Also, we used our goodness of fit measure, Q (adapted from eq. 1 in Materials and Methods) to estimate the pairwise deviation between Mark's corrected replicates:

$$Q = \frac{\left(HH_{n_{\text{replicate i}}} - HH_{n_{\text{replicate j}}}\right)^2}{HH_{n_{\text{replicate i}}}}$$

The maximum value of Q for between replicate HH_n in corrected sequence did not exceed 1×10^{-4}. This was important because we also use the Q value for demographic inference to measure of the goodness of fit between observed HH_n and an analytical prediction of HH_n for a specified demography. Our experience with real and simulated data indicated that an appropriate threshold for Q was $\delta \leq 0.001$. The low variation between replicates of stochastically corrected data demonstrated that the correction method is robust and indicated that a single stochastically corrected sequence would have been adequate for our estimate of HH_n. However, we used the averaged HH_n across replicates for our demographic inference. The observed HH_n for filtered and corrected sequence in figure 3 demonstrates that although there was only a very low level of residual false-positive heterozygous errors in filtered sequence, these errors still create a significant bias. False positives particularly affect the probability of observing longer RoH (fig. 3) and therefore may bias the estimates of more recent N_e. For example, the probability of observing RoH of 200 kb or longer is approximately 10% in the corrected data for both bulls but is only around 5% for the filtered data. Lengths of more than 0.5 Mb are almost never observed in the filtered data whereas after correction, there is a 6% probability of observing them (fig. 3).

Demographic Inference

Our method of inferring stepwise N_e searches for a demographic model that analytically predicts an HH_n in close concordance with observed HH_n for a range of segment sizes up to 1 Mb, with "goodness of fit" Q (eq. 1, Materials and Methods) assessed by a threshold parameter ($\delta \leq 0.001$). The reliability of the HH_n analytical predictions was demonstrated by comparing these with empirical HH_n in sequence data simulated using our inferred demography (supplementary fig. S2, Supplementary Material online). Importantly, the goodness of fit test (Q) between predicted and empirical HH_n in each of the simulated data sets also met our threshold parameter ($\delta \leq 0.001$).

We used Mark's corrected sequence to infer demography and calculated the upper and lower range of N_e within a given

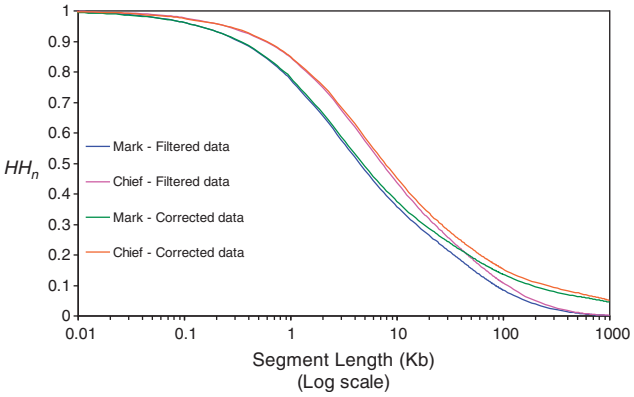

FIG. 3. The observed pattern of HH_n in the filtered and corrected sequence for Mark and Chief indicates that even when a low level of false-positive heterozygous errors remains in the sequence (filtered data), the observed pattern of HH_n is biased downwards, particularly in segments > 10 kb. The HH_n curves differ between bulls because Chief was sequenced at half the read depth of Mark, and the observed corrected data does not account for missed heterozygous SNP (false negatives).

Phase (time period) that met our goodness of fit threshold between predicted and observed HH_n (fig. 4A). The upper and lower limits for N_e were estimated while fixing the original N_e estimate in all other time periods. This was considered a reasonable approach because estimates of N_e in any given time period are most dependent on the distribution of particular lengths of RoH. In several very short stepwise changes, where $N_e \gg$ number of generations for that time period, it was only possible to give a lower limit but not an upper limit. This occurs because generally LD patterns are not affected by relatively brief surges in population size or bottlenecks that occur over very brief time periods $\ll N_e$ (Nordborg and Tavare 2002). On reaching the best fit model in figure 4A, we then adjusted two or more neighboring Phases of differing N_e to a single N_e value, provided HH_n remained within our threshold goodness of fit. We then re-estimated the predicted upper and lower N_e limits in this revised demography (fig. 4B). This illustrates that it is difficult to determine exact time boundaries for changes in N_e, although the overall demographic pattern remains similar in figure 4A and B.

The inferred demography found to closely predict the observed HH_n for Mark's corrected data (fig. 4A) reduces from the predefined large ancestral N_e (~62,000) to very small (~90) in present day ("present" being 1978, Mark's birth year). We have converted cattle generations into years before present using an average generation interval of 5 years: a reasonable estimate based on average generation intervals in modern cattle (Gutierrez et al. 2003; Mc Parland et al. 2007) as well as likely reproductive behavior in their wild ancestors. However, even if the estimated generation interval is increased to 7 years, this would not much affect our conclusions. Our results suggest that some 166,000 years ago (33,200 generations ago) there was a sharp reduction in N_e to ~17,000 and this remained stable until around 12,000 years ago. Over the following 3,000 years there was a further steep decline in N_e to ~3,500. From that time the reduction in N_e became

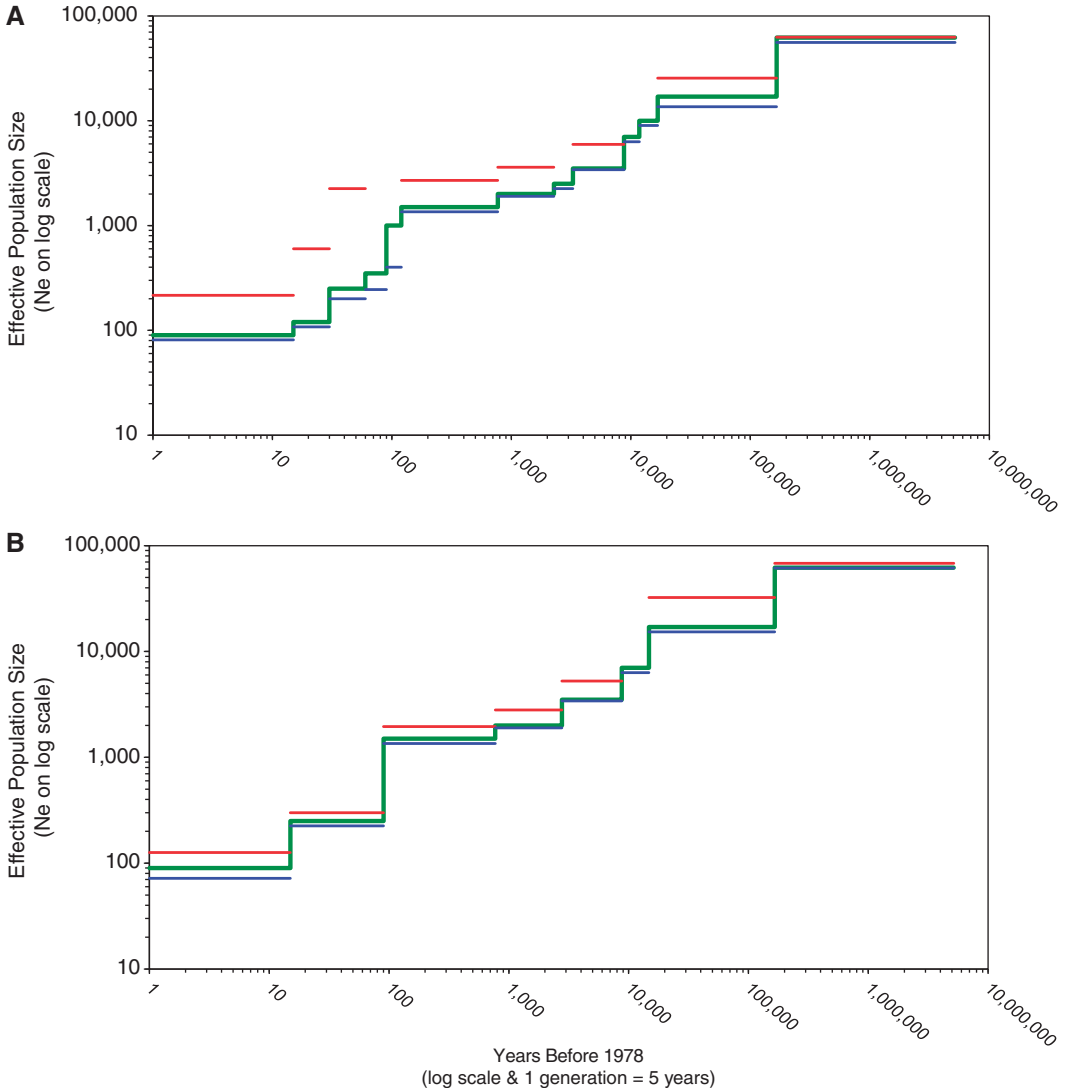

Fig. 4. (*A*, *B*) Inferred ancestral demography from Mark's corrected sequence (green), showing upper (red) and lower (blue) limits for predicted N_e within each stepwise estimate. The ranges shown for N_e estimates are those for which the predicted HH_n summary statistic meets our goodness of fit threshold ($\delta \leq 0.001$). (A) is the initially inferred demography, while in (B) the adjacent time Phases of N_e were adjusted to a single N_e value where possible, while still ensuring the demography met the goodness of fit criteria. Note that in short time periods where N_e considerably exceeds the number of generations for that time period, it is not possible to define a clear upper limit because sharp surges in N_e over few generations leave no signature on the distribution of RoH.

more gradual but some 120 years ago the N_e dropped rapidly again from around 1,500 to the current estimate of around 90.

In figure 5, we contrast the inferred demography using corrected sequence with that inferred from filtered sequence (i.e., not corrected for residual false positives), and also compare our results with those using the PSMC demographic inference method of Li and Durbin (2011). Parameter settings used for the PSMC method are given in the supplementary information, section 7, Supplementary Material online. Clearly the residual false positives in the filtered data have a major effect on estimates of N_e in recent time for both methods: in contrast to the sharp reduction in recent N_e, inferred from corrected data, both models infer a rapid expansion around 1,000 years ago to a present day N_e of around 20,000. Notably, the PSMC bootstrapping method (Li and Durbin 2011) gave a highly variable N_e estimate across the most recent 1,000 years (200 generations) with Mark's filtered sequence (supplementary fig. S10, Supplementary Material online). However, after applying our correction method to Mark's sequence, the PSMC estimate was much less variable in the most recent time period (supplementary fig. S11, Supplementary Material online). The PSMC inferred demography using corrected sequence was similar to our result (fig. 5), except in the most recent 1,000 years where PSMC estimates $N_e \approx 1{,}100$. However, in the PSMC method the final 1,000 years was modeled as a single fixed time period so there was no scope for more recent changes in N_e. Li and Durbin (2011) found that PSMC estimates of N_e in humans were not reliable more ancestrally than 120,000 generations ago and our estimates of N_e are similar at this time point (~600,000 years ago) but then diverge (fig. 5). We constrained our most ancestral estimate of N_e to a time period of approximately 1 million

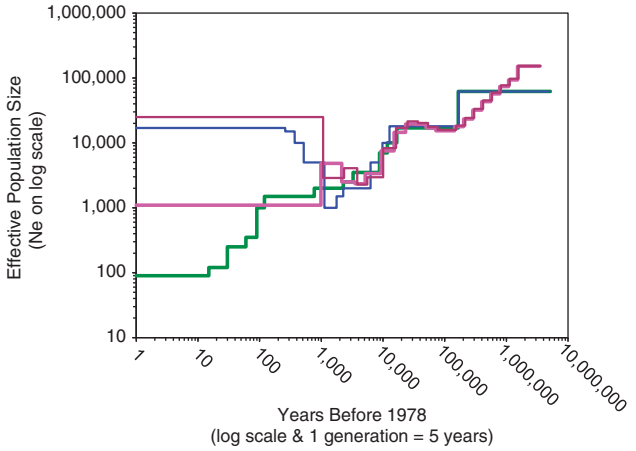

FIG. 5. Inferred demography for Mark's corrected sequence using our method (bold green line) and the Li and Durbin (2011) PSMC method (bold pink line). Also shown is the inferred demography from Mark's filtered sequence (with residual false-positive errors) using both our method (blue) and the PSMC method (maroon). There is a sharp contrast is the recent time N_e estimate between corrected and filtered sequence because of bias due to false-positive errors.

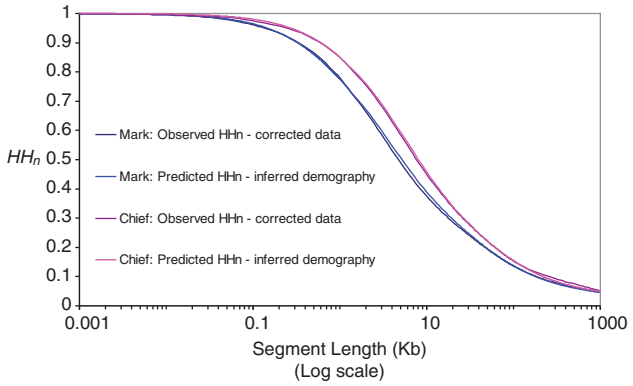

FIG. 6. Observed haplotype homozygosity (HH_n) in corrected sequence data for Mark (dark blue) and Chief (dark pink), compared with analytical predictions of HH_n for Mark (lighter blue line) and Chief (lighter pink line). Analytical predictions are based on the demographic model inferred from Mark's sequence only, but with mutation rates calibrated to match the observed single base heterozygosity in each bull. The close match between predicted and observed HH_n in Chief's data presents a validation of the inferred demographic model and data correction for false positives.

generations ago because the limited proportion of extremely short RoH limits the reliability of more ancestral N_e estimates.

A cross-validation of the inferred demography from Mark's corrected sequence was provided by using the inferred N_e parameters in the analytical model to predict the expected pattern of HH_n in Chief's corrected data, but re-scaling mutation rate to match Chief's observed single base heterozygosity (table 2). The recombination rate (r) was fixed at 1×10^{-8} between base pairs for both bulls, while rescaled mutation rates per site per generation were: $\mu_R = 4.1 \times 10^{-9}$ in Mark and $\mu_R = 2.15 \times 10^{-9}$. These mutation rates were lower than the assumed true mutation rate of 1×10^{-8} (Kumar and Subramanian 2002; Roach et al. 2010; Campbell et al. 2012) because they are scaled to account for the false-negative error rates in the sequence. The prediction of Chief's HH_n using Mark's inferred demography (fig. 6), deviated by $Q < 0.001$ when compared with observed HH_n across the range of segment sizes up to 0.01 Morgan. The pattern of observed and predicted HH_n for each bull in figure 6 demonstrates the closeness of fit for the inferred demographic model HH_n.

We also used Chief's corrected sequence to independently infer a demography, fixing the most ancestral N_e to 62,000 (fig. 7). The inferred demography, although broadly similar to Mark's, shows departures in the estimated time boundaries for changes in N_e and does not always fall within the upper and lower limits of N_e estimated with Mark's data. However, Chief's sequence has a very high false-negative rate (~70%, due to lower coverage than Mark) and we would therefore expect that this inferred demography is less accurate than Mark's. With such a high false-negative error rate the large proportion of missing heterozygous positions is likely to have a more variable effect on the distribution of observed RoH. We further explored the accuracy of the method to infer demography using simulated data with added false-negative

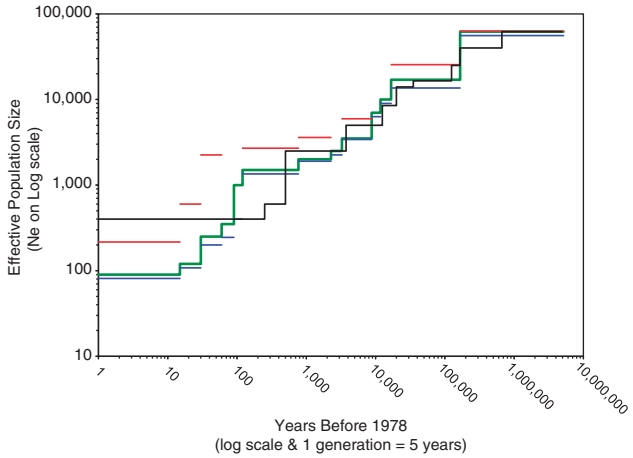

FIG. 7. Inferred ancestral demography from Chief's corrected sequence (black line) compared with the demography inferred using Mark's corrected sequence (green line). Also shown are Mark's upper (red) and lower (blue) limits for predicted N_e that met our goodness of fit threshold ($\delta \leq 0.001$).

errors (supplementary information, section 5, Supplementary Material online).

Finally, we accounted for the false-negative error rate in corrected sequence by scaling up the mutation rate. We calibrated the analytical model to predict a match with the estimated true single base heterozygosity, resulting in a mutation rate for error free sequence data of $\mu = 9.4 \times 10^{-9}$ per base per generation (close to our assumed true mutation rate: 1×10^{-8}). Although this correction has no influence on the estimates of N_e, it is critical because it provides the final mutation rate required to produce a simulation of error free sequence data. Supplementary figure S1, Supplementary Material online, compares the predicted HH_n for Mark's corrected sequence with the predicted true HH_n if all

heterozygous SNP had been detected, given our inferred demography. The predicted true HH_n curve for Chief is the same as for Mark (not shown) because the estimated true heterozygosity for each bull differs by only 5×10^{-6} (table 2).

Discussion

In this study, we use whole-genome sequence to infer a stepwise demographic model by matching an empirical and analytically predicted summary statistic of the RoH distribution (HH_n). Importantly, our study shows that even a low level of residual errors in sequence data (after stringent filtering) can lead to a considerable bias in estimates of the more recent N_e. We therefore developed a simple but robust method to correct for false-positive heterozygous errors. Furthermore, we demonstrate that this error correction method considerably improved the accuracy of the Li and Durbin (2011) PSMC method of demographic inference.

Comparison with Previous Methods

Two recent studies have estimated ancestral patterns of demography using genome sequence from several humans with diverse ethnic backgrounds (Gronau et al. 2011; Li and Durbin 2011). Gronau et al. (2011) modified the Rannala and Yang (2003) method which implements a Bayesian Markov chain Monte Carlo (MCMC) approach, sampling from many possible genealogies to determine the likelihood that a given set of demographic parameters could have given rise to the observed properties of the sequence data. However, because this approach is too computationally demanding to implement with entire genome sequence, the authors selected 37,500 "neutral loci" from the sequence data, each of 1 kb length (~1.5% of each genome). They exploit the pattern of mutations to infer demographic history under the assumption that the "neutral loci" are in linkage equilibrium and that intra-locus recombination can be ignored. Importantly therefore, their method is not able to capture additional demographic information from LD and they emphasize that their primary focus is to estimate divergence times and migration rates between diverse human populations. Their model only allows for changes in N_e at the time of divergence, but otherwise N_e remains constant.

Li and Durbin (2011) inferred ancestral demography from sequence by assessing the distribution of heterozygosity along the genome sequence of a single individual. Using a hidden Markov model, they infer time to most recent common ancestor and use the distribution of coalescent times to estimate a stepwise demographic pattern (PSMC). This shares some similarities with our approach, however they first condense the sequence data by redefining nonoverlapping windows of 100 bp as a "single locus" being either heterozygous or homozygous. We use all heterozygous sites in our observed sequence to measure RoH so that no data are "lost." The Li and Durbin binning of 100 bp into a single locus may incur some loss of data where there are stretches of very high heterozygosity because these are potentially a result of very distant ancestral coalescent events. Generally their N_e estimates for human data showed reasonable accuracy between 800 and 120,000 generations ago but were unreliable for more distant or more recent time periods. Recent N_e in human populations may be more difficult to reliably estimate due to the expanding N_e resulting in relatively few recent coalescent events present in the sequence (Li and Durbin 2011). Our relatively narrow estimates for recent time N_e in Holstein cattle were potentially aided by the rapidly reducing N_e in Holsteins resulting in many more recent coalescent events compared with human populations.

Li and Durbin (2011) do not correct their sequence data for residual false-positive errors, although using simulations they show that these errors result in an upward bias in their more recent N_e estimates. Importantly, we confirm that our method of correcting for residual false positives had a strong impact on improving the precision of the PSMC method to estimate recent time N_e (supplementary information and figs. S10 and S11, Supplementary Material online). There was reasonable agreement between the inferred bovine demography using PSMC and our method between 200 and 120,000 generations ago. However, using simulated data we show that time boundaries for significant changes in N_e are not very precise with either our method or PSMC (supplementary information, section 5, and fig. S4A and B, Supplementary Material online). Our method can also be extended to estimate population divergence times in a similar way as the Li and Durbin (2011) approach, by combining data from male X chromosomes of two individuals from different populations. At the time of divergence, we would also expect to find a sudden increase in the estimated N_e, therefore we believe that it would be important to first correct for false positives.

We believe that a strength of our approach is the modeling of multiloci LD across a wide range of segment sizes, rather than pairwise LD. Schaffner et al. (2005) calibrated a human demographic model to match pairwise LD and empirical allele frequencies for SNP across a wide range of genomic segments. Recently, Pool et al. (2010) used this human demography to simulate data and compared the distribution of long RoH in simulated data with empirical HapMap data. They found that the simulations considerably underestimated the proportion of longer RoH observed in empirical human data. Pool et al. (2010) argue that more detailed population genetic information for recent times may be gleaned by considering patterns of multiloci LD. Hayes et al. (2003) demonstrated that their multiloci measure of LD could be used for estimating past N_e, and had a lower coefficient of variation compared with the pairwise r^2 LD measure. However, their methodology is computationally more demanding than ours and their analytical model does not include mutation.

Correcting for Errors and Potential Biases

Our results demonstrate the importance of correcting for even very low levels of false-positive errors. This is likely to be most important in species where there have been recent sharp reductions in N_e such as cattle or substantial recent bottlenecks. We estimated error rates and corrected the

sequence data of each bull independently, allowing us to test the effectiveness of our stochastic false-positive error correction, as well as our demographic model. Although read depth in Chief was low (~7×) and approximately half that in Mark, our inferred demography from Mark's HH_n predicted a close match with Chief's observed corrected HH_n.

When we tested our false-positive correction method in simulated sequence data, the true distribution of RoH was not completely restored (supplementary information, section 6, and fig. S8, Supplementary Material online). In contrast, error correction in Mark's data appears to restore the length of RoH closer to the true length (assuming true length = SNP50 RoH) than for simulated data (supplementary information, section 6, Supplementary Material online). We believe that our correction method worked better in real sequence data because rather than a very uniform spread of errors, there was some marked clustering of heterozygous SNP in small subregions within several sequence regions that matched a long SNP50 RoH. These clusters of heterozygotes may have arisen for several reasons: they may be real heterozygotes or false positives due to mapping errors of, for example, segmental duplications (SD; supplementary figs. S5 and S7, Supplementary Material online). If some clusters of heterozygosity were true heterozygotes the residual false-positive error rate may have been slightly overestimated, and this could have resulted in some RoH lengths being overestimated. Although this would not have had a major impact on the distribution of the longer RoH in our data, it may have resulted in a small downward bias in more recent N_e estimates. Conversely, if error rates were correctly estimated but we have only partially restored the true distribution of RoH to that of error free sequence, then our more recent N_e estimates will tend to have an upward bias.

Although we found some evidence of clusters of heterozygotes in SD > 1 kb regions (SD as identified by Liu et al. 2009), the filtering of sequence at least partially addressed likely errors in more highly repeated SD regions by excluding SNP with excessively high coverage. This removed 43% (Mark) and 37% (Chief) of heterozygous SNP within SD > 1 kb. Furthermore, these regions appear to only account for approximately 3% of the bovine genome (Liu et al. 2009), so overall these errors are unlikely to have had a very significant impact on our estimates of N_e.

In studies with larger SNP arrays, higher coverage and/or more individuals sequenced, it is possible to use more sophisticated methods to minimize false-positive heterozygous errors but an estimate of residual errors would still be required. A range of error detection strategies (such as machine learning and hidden Markov models) may be applied depending on the data available (Lynch 2008; Hoberman et al. 2009). Validation with individual SNP on commercial arrays should be used with caution because these SNP are generally chosen as reliably "well behaved" and may therefore result in a downward bias in estimated false-positive rates, because they are also less likely to produce errors in sequence data (due to other variants in close proximity for example).

In theory, false negatives (missing heterozygotes) should not affect the demographic estimates except as a timescale effect, so can be corrected for by scaling the mutation rate (Li and Durbin 2011). To test the theory, we simulated sequence data with 50% false-positive errors and then used this data to infer the demography. Having inferred the demographic model from simulated data with 50% false-negative errors, we were then able to analytically predict a close match to observed HH_n in error-free simulated sequence by scaling up the mutation rate (supplementary information, section 5, Supplementary Material online). However, the higher the false-negative error rate in sequence the more difficult it will be to adequately interpret and account for the effect of large amounts of missing heterozygous data. Although the independently inferred demography using Chief's sequence with 70% false-negative error rate was similar to Mark's, we advise caution in using sequence data with any more than 50% false negatives.

Our analytical method of predicting HH_n and inferring demography assumes a known and constant genome wide recombination and mutation rate. If our rescaled average mutation rate is lower or higher than the true value, this could result in some timescale changes but would still be expected to predict a pattern of demography reducing in size to present day (supplementary fig. S3, Supplementary Material online). Variable recombination rates due to hotspots could arguably increase the proportion of some lengths of RoH, resulting in more variable estimates of N_e across time. Evidence to date indicates that although hotspots occur more in intergenic regions, their density and intensity varies across the genome and they appear to be evolving quite rapidly (International HapMap Consortium 2007). Therefore, overall we expect variable recombination rate to have a minimal impact on the overall demographic model, because we are using whole-genome sequence and a summary statistic of LD. We suggest a similar argument holds for variable mutation rates across the genome, but this could be further tested through simulations.

A limitation of our study is that we used data from two bulls only, one for inference and one for validation. Our estimate of more ancestral demography is unlikely to be affected because there are hundreds of thousands of small homozygous segments in the sequence of these animals that are inherited independently from much more distant ancestors. Our estimates of very recent effective population size (N_e) could be biased because, by chance, these bulls could be more or less inbred than other cattle of this breed. Very long RoH arise from one or a combination of; small recent N_e, inbreeding and recent intense selection. However, both bulls have made a major contribution to the North American Holstein population so should be representative of the current population (Young and Seykora 1996). Furthermore, our recent time N_e estimate is close to that expected from major pedigree analysis of the Holstein breed (Weigel 2001; Stachowicz et al. 2011). If multiple genome sequences are available for a population (with the same coverage and false-negative rates), then a combined estimate of the distribution of RoH would be preferable to improve the recent time N_e estimates.

Inferred Bovine Demography

From approximately 1,500 generations ago to present, our predicted demography broadly follows a similar pattern of decreasing N_e found in several other studies based on pairwise or multiloci LD in marker data (Hayes et al. 2003; Gautier et al. 2007; Kim and Kirkpatrick 2009; MacLeod et al. 2009; Villa-Angulo et al. 2009). However, our full sequence data should provide better estimates of LD across very short segments compared with the less dense marker data used in these previous studies. This should allow more certainty of parameter estimates in the more distant past because the very short homozygous segments trace back to very distant ancestors. MacLeod et al. (2009) used RoH in SNP50 data to estimate bovine demography and found very low sensitivity for estimates of N_e beyond 3,600 years ago (720 generations), because their data contained no SNP in very close proximity.

We inferred a sharp reduction in N_e approximately 170,000 years ago, possibly marking the period of divergence between African and European *Bos taurus* cattle estimated to have taken place between 26,000 and 250,000 years ago (Bradley et al. 1996; MacHugh et al. 1997; Troy et al. 2001). Alternatively, it may be linked to the divergence between *B. indicus* and *B. taurus* cattle estimated to have occurred between 100,000 and 500,000 years ago (Ritz et al. 2000; Ho et al. 2008; Murray et al. 2010). To test if this was simply a reflection of our most ancestral N_e being over-estimated, we also attempted to infer a bovine demography beginning with a fixed most ancestral N_e of 15,000, rather than 62,000. However, the demography could not be inferred without the estimated true mutation rate rising to above 5×10^{-8}, which is most unlikely.

In figure 4*A*, we also infer a steep decline in N_e (9,000 to 17,000 years ago) around the time of cattle domestication, estimated to have taken place in Neolithic times some 10,000 years before present in the Near East (Perkins 1969; Bruford et al. 2003). It is possible that this bottleneck begins a little earlier in our demography due to wild *B. primigenius* populations being affected by post glacial warming and expansion of human populations that were taking place at this time (Soares et al. 2010). A severe population bottleneck around the time of domestication might be expected due to the difficulties involved in the initial capture and taming of wild cattle, and also because evidence suggests that *B. taurus* domestication was confined to one or two very local regions (Bruford et al. 2003).

A recent genetic study of cattle domestication in the Near East using mitochondrial DNA (mtDNA) samples from ancient and modern domestic cattle, from the Near East only, estimated the domestication founder female N_e to be around 80 with confidence limits of 23 to 452 (Bollongino et al. 2012). Although this is a very severe bottleneck, it may not be in conflict with our results because this is a mitochondrial DNA estimate of founding female N_e. Bollongino et al (2012) assumed a single domestication event and mitochondrial DNA samples were restricted to ancient and modern cattle sampled from close to the original site of domestication. It is likely that their estimate is therefore for a more localized period than would be possible using our method. Our method would not detect a relatively brief bottleneck nor can it estimate female only N_e, and it is likely that this female founder population was quite rapidly increased by keeping and rearing of female offspring. It seems highly plausible that the females were more easily managed than males, but it would be difficult to impossible to prevent tamed females and their female offspring from breeding at random with wild males. It is possible, given our N_e estimate around the period of domestication was approximately 3,500, that for some time the female N_e could have been as small as 1,200 while male N_e would then be 3,200 (based on the approximation: $N_{e\ Total} = 4N_{e\ female}N_{e\ male}/[N_{e\ female} + N_{e\ male}]$). Only after many generations of breeding for tameness would it have become easier to manage and retain males in a domestic setting. Thus, for quite some time during the domestication process the male N_e was likely larger than the female N_e: a situation that would become gradually reversed up to modern day cattle breeding where the very small elite male population now limits the overall N_e.

In apparent conflict with our results, Murray et al. (2010) detected no domestication bottleneck for taurine cattle, although they did detect a severe bottleneck around 30,000 to 50,000 years ago. There are a range of difficulties in determining the accurate timing of bottleneck events, including knowledge of mutation rates for the loci studied. The study by Murray et al. (2010) may have lacked power because they used gene regions which may have been subject to selection, and their methodology inferred demography from the summary statistic of "site frequency spectrum" (SFS) only, which exploits information regarding mutation but not recombination. Furthermore, they combined genomic data across a range of *B. taurus* breeds from Europe and Africa, which may have concealed any domestication bottleneck because there is some molecular evidence suggesting different domestication origins for European and African taurine cattle (Bradley et al. 1996; Beja-Pereira et al. 2006). The wide variation in past estimates of the timing of bovine bottlenecks highlights the need for further studies to confirm the accuracy of our estimates.

Following the domestication period our estimates of further gradual decline in N_e are potentially due to increasing genetic isolation from the wild population, limited numbers of domestic cattle being taken to northern Europe, and the start of breed formation (Beja-Pereira et al. 2006). In the last 100 years, the further decrease in N_e is likely a result of breed registration rules requiring that animals are purebred, as well as the high selection intensity in modern breeding programs through very extensive use of artificial insemination (Goddard 1992; Young and Seykora 1996; Stachowicz et al. 2011). Although Finlay et al. (2007) report a sharp increase in cattle population size from domestication to present day, their study is based on mtDNA and is therefore indicative of the expanding female population size with no adjustment made for the likely decreasing to very small male effective population size. They also combined mtDNA from a number of different cattle breeds and analyzed this as one population which therefore reflects the present day population size of

domestic cattle generally, such that in this context an expanding population is not surprising.

Similarly, Murray et al. (2010) estimated a large present day N_e for domestic cattle but they also combined genomic data from a range of breeds to define a "taurine" population. Our estimate of present day population size of between 80 and 220 is close to several independent estimates for this breed using both LD methods (Hayes et al. 2003; de Roos et al. 2008; Kim and Kirkpatrick 2009; MacLeod et al. 2009) as well as extensive pedigree records (Weigel 2001; Stachowicz et al. 2011). This current day N_e indicates the importance of taking steps to monitor and minimize inbreeding in Holstein cattle to avoid potential negative effects of inbreeding depression on economic traits. Two traits of particular concern in modern Holstein dairy cattle breeding are fertility and longevity (VanRaden 2004) both of which could potentially be affected by inbreeding depression as a result of low N_e.

Conclusions

Our study demonstrates a computationally efficient method of inferring ancestral demography from empirical observations of the distribution of RoH in whole-genome sequence. We also demonstrate a method to correct for the potential serious bias of residual false-positive errors on recent estimates of N_e. Our inference method can be applied to any outbred diploid species for one or multiple individuals without the need to phase the data into haplotypes. That is, HH_n can be summarized from all RoH measured within individuals, provided their sequences have similar false-negative error rates. If sequence haplotypes are available from unrelated individuals the RoH can be measured both within and between pairs of individuals. Our method provides a demographic model that can be used to simulate sequence data generated under a null hypothesis with realistic multiloci LD patterns, to calibrate significance tests for evidence of selection or for a range of other genetic studies.

Materials and Methods

For this study, our definition of RoH in sequence data refers to an observed unbroken run of homozygous (i.e., IBS) base pairs along a pair of homologous chromosome segments within a diploid individual. We define our HH_n summary statistic in sequence data as the probability that a pair of homologous chromosomes are observed IBS for at least n base pairs to the right of a site chosen uniformly at random. Three key steps in our methodology for demographic prediction are as follows:

1) Detection of heterozygous SNP and error rates in whole-genome sequence,
2) Correction of false positives and calculation of the summary statistic, HH_n, from the distribution of RoH for a range of segment sizes (n), and
3) Inference of a demographic model that predicts HH_n matching that observed in corrected sequence.

A summary of the workflow is given in figure 1.

Whole-Genome Sequence

Whole-genome sequence was generated for two Holstein–Friesian bulls using a 454 FLX-Titanium platform (Larkin et al. 2012). The bulls were Pawnee Farm Arlinda Chief ("Chief") and one of his offspring, Walkway Chief Mark ("Mark"). Mark was sequenced at approximately 13× coverage, while Chief was sequenced at approximately 7× coverage. Details of sequencing, alignment, mapping, and SNP discovery are published in Larkin et al. (2012). For our study, we extracted all information on heterozygous SNP within each animal's autosomal genome sequence, and applied stringent quality control filters to remove likely errors (Larkin et al. 2012, supplement). Filtered heterozygous sequence SNP were then used to measure base pair (bp) length of all intervening RoH within each animal's sequence.

Genotyping with SNP50 BeadChip

DNA from Chief, Mark and 92 of Mark's offspring was used to generate approximately 50,000 SNP genotypes each, using the Infinium BovineSNP50 BeadChip (http://www.illumina.com/products/bovine_snp50_whole-genome_genotyping_kits.ilmn, last accessed July 23, 2013) (described in Larkin et al. 2012). The "SNP50" genotypes, generated independently of the sequence data, were used to validate the corresponding sequence SNP genotypes, but were first subject to strict quality control. SNP50 genotypes from Mark's offspring were used to identify and remove inconsistencies in Mark's SNP50 genotypes, and checks were also made for any discordant genotypes between Chief and Mark. Additionally, based on preliminary checks of concordance between SNP50 and sequence SNP, SNP50 genotypes were discarded if the Illumina "Gen Train" and "GenCall" scores were less than 0.8 (http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf, last accessed July 23, 2013). All SNP50 genotypes with unknown reference genome position were also eliminated. There remained 38,956 and 39,026 of Mark and Chief's SNP50 genotypes. The lengths of all RoH between SNP50 heterozygous positions were recorded within each animal.

Correcting Sequence for Errors

Two types of error in the sequence data must be quantified; "false-negative" errors (heterozygous positions missed in the sequence data and called as homozygous) and "false-positive" errors (homozygous positions wrongly called as heterozygous in the sequence data).

The pattern of heterozygous sites (clustered or not, dense to sparse) occurring along a pair of sequences has a direct impact on the distribution of observed lengths of RoH. In a neutral model, the pattern of heterozygous sites across the genome can be affected by changes in past N_e, such as bottlenecks, and recombination rate. If false-positive errors arise randomly across the genome they will break up long RoH into a number of shorter RoH, whereas in regions with more dense true heterozygous sites the false-positive errors will have much less impact on the shorter RoH. Thus false positives

may particularly bias the estimation of more recent N_e, because the longer RoH coalesce to more recent ancestors.

The distribution of false negatives (missed heterozygous sites) should follow the same distribution as discovered true heterozygous sites assuming no bias in the sequencing method or SNP discovery. These errors will therefore more similarly affect both shorter and longer RoH and thus will have little impact on our estimates of ancestral N_e. However, the number of heterozygous sites present in the genome under a given demography, is directly proportional to the mutation rate per site per generation, and false-negative errors reduce the number of observed heterozygous sites. These missed heterozygous sites represent an independent thinning of the frequency of true heterozygous sites, which is equivalent to a reduced mutation rate. It should therefore be possible to correct for false negatives by simply rescaling the true mutation rate (μ_T) before inferring the demography. The scaled mutation rate (μ_R) should be in the order of: $\mu_R = \mu_T(1-q)$, where the false-negative error rate (q) was estimated by validation with SNP50 data (assumed error free), so that $1-q$ is the proportion of sequence SNP observed as heterozygous given the SNP50 position was heterozygous. We assumed $\mu_T \approx 1.0 \times 10^{-8}$ per base per generation based on recent estimates of mammalian mutation rates (Kumar and Subramanian 2002; 1000 Genomes Project Consortium 2010; Roach et al. 2010; Campbell et al. 2012). Similarly, we estimated true single base heterozygosity (H_T) in the sequence of each bull by scaling the observed heterozygosity (H_O): $H_T = (H_O)/(1-q)$.

The residual false-positive error rate in our filtered sequence was expected to be very low and therefore the accuracy of estimated error rate by direct validation with SNP50 data was limited by the small number of SNP50 heterozygous positions (table 1). There is also potential bias due to unidentified errors in SNP50 data and the nonrandom choice of "better behaved" SNP on commercial SNP arrays. We therefore developed a simple method to quantify residual false-positive error rates in filtered sequence data, using long RoH in the SNP50 data. An RoH in our SNP50 data was defined as a run of adjacent homozygous SNP genotypes within an individual. We identified several SNP50 RoH spanning >10 Mb; five in Chief and four in Mark. There is a high probability that these SNP50 long RoH identify IBD chromosome segments, that is, these regions are also very likely to be homozygous in sequence data. This assumption is justified because within the pedigree of each bull there are several inbreeding loops to recent common ancestors three to six generations ago (https://www.holstein.ca/, last accessed July 23, 2013). In cattle, one Morgan is approximately equal to 1×10^8 base pairs (Arias et al. 2009). Although mutations may occur in the generations since inheriting chromosome segments IBD from these common ancestors, this is likely to account for only a single heterozygous position per 10 Mb every 5 generations, assuming a mutation rate of 1.0×10^{-8} per base per generation. Average single base heterozygosity across these SNP50 long RoH in the filtered sequence data therefore provided an estimate of the residual false-positive error rate. We excluded the outer 0.15 Mb of the SNP50 RoH region when estimating the error rate to avoid the possibility of including a region beyond the end of the sequence RoH, that by chance appeared as part of the SNP50 long RoH due to the average distance between SNP50 being 0.07 Mb.

To remove bias due to false-positive heterozygous errors in the sequence, we aimed to restore the distribution of RoH rather than identify actual false positives remaining. Filtered sequence was "corrected" by removing the expected proportion of residual heterozygous errors, assuming uniform distribution across the genome. Thus, in nonoverlapping windows of three times the average length that contains one false heterozygous error, we randomly deleted three heterozygous SNP per window (or less if fewer existed in a window). We chose this window length having first tested the method with a range of window sizes (1 to 5) and compared the resulting RoH pattern across the entire genome of Mark and Chief with those of RoH in the SNP50 data (e.g., supplementary figs. S5, S6, and S7, Supplementary Material online). We also compared our "3 error window" correction method with randomly deleting the same proportion of heterozygous SNP without implementing uniform deletion from nonoverlapping windows: that is, removing the restriction that false-positive errors arise with equal probability across the genome.

We tested the variability in the resulting RoH distribution after applying the 3 error window correction method by replicating the data correction 25 times, resulting in 25 sets of "corrected sequence" for each bull. Henceforth, these data sets are referred to collectively as "corrected sequence" data. Further validations of the correction methods for false negatives and false positives were carried out using simulated data (supplementary information, section 6, Supplementary Material online).

Observed HH_n Summary Statistic in Sequence

We use the HH_n summary statistic (MacLeod et al. 2009) to describe the distribution of observed RoH in diploid whole-genome sequence. For any given value of n, HH_n is calculated as the proportion of sites in the diploid genome for which at least n bases to the right are observed homozygous, expressed relative to the total possible number of such sites if the entire genome were homozygous. Take a trivial example of calculating HH_5 (i.e., n is 5) in one individual with a single chromosome only 10 bp long in which we observe only one RoH of 6 bp. For this individual, $HH_5 = 2/6$ because moving left to right across each base pair on this chromosome there will be only two sites at which we will observe at least 5 homozygous base pairs to the right, and the maximum possible number of sites if the entire 10 bp had been homozygous is six.

We calculated HH_n in both filtered and corrected sequence data, for a range of segment sizes between 1 and 1,000,000 bp. This maximum segment length was chosen because this should be informative up to recent times given that LD on a segment is most influenced by N_e approximately $1/(2c)$ generations ago assuming a linearly changing N_e, where c is the segment length in Morgans (Hayes et al. 2003). Also, in cattle the RoH distribution becomes relatively flat at this length because although there are some rare much longer

RoH, the majority of RoH are shorter. The HH_n was calculated for each of the 25 replicates of "corrected sequence" and was then averaged across replicates for each bull. To evaluate the robustness of the correction method we measured the variability between replicates of the summary statistic, HH_n, across a range of segment sizes (n) up to 1 Mb, because HH_n provides the basis for inferring the demography. Variability was assessed as the coefficient of variation across the 25 replicates.

Analytical HH_n Prediction

For demographic inference, we compared the observed HH_n in sequence with an analytically predicted HH_n. The analytical HH_n prediction is based on a simplified coalescence method that accommodates stepwise changes in historical N_e, with constant mutation and recombination rates (MacLeod et al. 2009). We implemented a small modification to the original method to increase the computational speed of the calculation without compromising the accuracy of prediction (details are given in supplementary information, section 1, Supplementary Material online). With the modified method, we predicted HH_n for $n = 1, 2, 3\ldots$ to 1,000 bp, and then for 1,000, 2,000, 3,000,... to 1,000,000 bp. This allowed us to rapidly test a range of demographic models to search for one predicting a good match to observed HH_n.

Demographic Inference

The inference approach is similar to those where demographic parameters are sampled from a grid of prior parameters to determine those most likely to have given rise to summary statistics in observed data (Beaumont 2004). However, rather than simulating data with each new set of N_e parameters sampled, we use the analytical model to predict the HH_n summary statistic for any sampled set of demographic parameters and determine the goodness of fit with the observed HH_n across a range of segment lengths (MacLeod et al. 2009). We used the averaged HH_n from the 25 replicates of Mark's corrected sequence to infer the demography.

Parameters in the model include; effective population size over variable time periods, "Phases" ($N_{e\ \text{Phase i}}$) with time measured in generations ($G_{\text{Phase i}}$), as well as mutation rate (μ) and recombination rate (r). The model assumes a single population with no selection or migration. Coalescent time scales are dependent on N_e, μ, and r, therefore we assume both μ and r are constant across time and across the genome. We fixed r between any base pair as 1×10^{-8}, so that 1 Morgan was assumed to be approximately equal to 1×10^{8} base pairs (Arias et al. 2009). We assumed true mutation rate (μ_T) to be in the order of 1.0×10^{-8} per single base per generation based on recent mammalian estimates (Kumar and Subramanian 2002; Roach et al. 2010; Campbell et al. 2012). Therefore, the scaled mutation rate accounting for false negatives in the inferred demography was expected to be in the order of $\mu_R = \mu_T(1 - q)$. The accuracy of this scaling depends on how close the estimated false-negative rate, q, is to the true value.

The analytical model cannot infer a starting value for N_e, therefore we ran some preliminary checks with a range of simple models with the most ancestral N_e between 50,000 and 100,000 and compared single locus heterozygosity with that observed in the sequence. We used this N_e range because the ancestral N_e of domestic cattle has been estimated to be between 50,000 and 100,000 based on evidence from several independent studies (de Roos et al. 2008; MacEachern et al. 2009), around the time of *Bos* species divergence from *Bubalus* (buffalo) 1 to 5 Ma (Ritz et al. 2000). Based on these preliminary checks, our starting assumption was that the most ancestral $N_e = 62{,}000$. The method is not expected to predict N_e further back than around 1 to 1.5 million generations ago given a rough rule of thumb that LD on segment lengths of *c*Morgans will inform the N_e estimates $1/(2c)$ generations ago. It was therefore assumed that the most ancestral population had reached a drift-recombination-mutation equilibrium, over a time period fixed as $>10N_e$ generations.

Our demographic inference makes no prior assumption of the maximum number of time intervals (Phases) or the specific boundary of time intervals at which there could be an instantaneous change in N_e. Rather, we begin with constant N_e and exploit the theory that LD over shorter distances reflects more ancestral population parameters than LD at larger distances. We employed the iterative approach of MacLeod et al. (2009) to search the parameter space for the best fit demography, with variable N_e:

1. Use the analytical model with the starting parameters to predict HH_n for segments of length n, where n was 1–1,000 bp, and then 1, 2,. . ., 1,000 kb.
2. Test the match between the predicted HH_n and the observed HH_n in the sequence, across the range of segment lengths. The HH_n summary statistic is a continuous variable for each segment length tested, therefore a "match" was defined as meeting a threshold goodness of fit test ($Q \leq \delta$) for each HH_n in the range of segment lengths tested):
$$Q = \frac{(HH_{n\ \text{Predicted}} - HH_{n\ \text{Observed}})}{HH_{n\ \text{Observed}}} \leq \delta, \qquad (1)$$
where δ is a stringent predetermined threshold that we set to 0.001. This threshold choice was based on our prior experience with the model using simulated data. It was also confirmed as reasonable because HH_n in replicated simulations using our inferred demography generally differed by $Q \leq \delta$ for all segment lengths up to 0.01 Morgan.
3. If the threshold was not met at any one or more HH_n for a given segment size n, the N_e was resampled over one or more time periods (Phases) since the ancestral population. It is expected that LD on a segment size of c Morgans is most affected by the population size approximately $1/(2c)$ generations in the past (Hayes et al. 2003). We therefore conditioned the re-sampling of N_e over a variable time period that corresponded to the range of segment lengths where there was a mismatch in observed and predicted HH_n. Therefore, the time boundaries for changes in N_e were not predetermined, but rather

were estimated as the approximate time periods corresponding to segment lengths n where HH_n mismatched. For example, if HH_n was under (over) predicted compared with the observed HH_n for segments of lengths of 1.0×10^{-5} to 0.01 Morgans, we assumed that from approximately present day to 50,000 generations ago $[1/(2c)]$ the N_e should be reduced (increased) in size from the current value to $N^*_{e\ \text{Phase i}}$. However, if the reduced (increased) N^*_e was close in value to $N_{e\ \text{Phase i}+1}$ or $N_{e\ \text{Phase i}-1}$ then we first tried to match this to minimize the number of additional Phases. If HH_n mismatched across one or more time periods between others that matched, the N_e was first adjusted in the poorest fitting, most ancestral time period followed by a return to Step 2.

Steps 2 and 3 were repeated until a demographic model was found that met the goodness of fit criteria for HH_n across all segments lengths tested. If required, the mutation rate was adjusted slightly to ensure a close match ($\pm 5.0 \times 10^{-5}$) with the single locus heterozygosity observed in the corrected sequence.

Step 2 and 3 could also be implemented with a systematic sampling from a grid of prior N_e values over a pre-defined number of Phases (each of G generations) with fixed time boundaries. The goodness-of-fit parameter (Q) can be summed across the range of HH_n for each tested demography ($\sum_{n=1}^{k} Q_{HH_n}$, where k is the total number of HH_n values tested) and minimized to provide a means of ranking each demographic model tested.

Having successfully inferred a demography from Mark's data that met our threshold δ, we estimated upper and lower limits for each stepwise change in N_e based on the maximum and minimum N_e value possible where the threshold of $\delta \leq 0.001$ was still met. All other stepwise values of N_e were held constant when estimating these upper and lower bounds. Slight adjustment was made to the mutation rate if necessary to ensure a match with observed single locus heterozygosity. For computational efficiency, we tested intervals of ± ~10% of each stepwise N_e, or less if the first increase/reduction resulted in $\delta > 0.001$. After inferring a good fit demography, we then attempted to combine two or more adjacent Phases of differing N_e to a single N_e value to test the resolution of determining time boundaries for changes in N_e.

Finally, using our inferred demographic model, we predicted expected HH_n for the corrected sequence data had there been no false-negative errors (missing heterozygous SNP). We did this by scaling up the mutation rate to match the estimated true single base heterozygosity in Mark's sequence. This step has no effect on the N_e estimates but is of considerable importance for estimating the true mutation rate required to simulate sequence that mimics error free sequence data, where all true heterozygous sites are discovered.

Validation of the Inferred Demographic Model

We cross validated the inferred demography, using the analytical model to predict the expected HH_n in Chief's independently corrected sequence. Assuming we have accurately corrected for the different false-positive error rates in the two bull sequences, the difference between the distribution of RoH in Mark and Chief's corrected sequence is due only to a higher false-negative rate in Chief's sequence (lower coverage). If false-negative errors are equivalent to a lowered mutation rate, then Mark's demographic model should predict Chief's HH_n by simply rescaling the mutation rate to account for the higher proportion of false negatives in Chief's sequence. We therefore rescaled the mutation rate to match Chief's observed single base heterozygosity, predicted HH_n and estimated the goodness of fit parameter (Q) between predicted and observed HH_n.

We also used Chief's corrected sequence HH_n to independently infer demography, although acknowledging that this data would be less reliable than Mark's because the lower coverage resulted in a high false-negative error rate (~70%). However, we also inferred demography using simulated data (based on Mark's inferred demography) in which we had randomly changed heterozygous SNP to homozygous to mimic the false-negative rate in Mark's corrected sequence (supplementary information, section 5, Supplementary Material online).

Supplementary Material

Supplementary information, table S1, and figures S1–S11 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to two anonymous reviewers for thoughtful suggestions to improve the manuscript. This work was supported by Australian Research Council's Discovery Projects funding scheme (grant number DP1093502) to M.E.G. and by the US Department of Agriculture Cooperative State Research Education and Extension Service, Livestock Genome Sequencing Initiative (grant numbers: 538 AG2009-34480-19875 and 538 AG 58-1265-0-031) to H.A.L.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

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

Beaumont MA. 2004. Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity* 92:365–379.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Beja-Pereira A, Caramelli D, Lalueza-Fox C, et al. (21 co-authors). 2006. The origin of European cattle: evidence from modern and ancient DNA. *Proc Natl Acad Sci U S A.* 103:8113–8118.

Bollongino R, Burger J, Powell A, Mashkour M, Vigne J-D, Thomas MG. 2012. Modern taurine cattle descended from small number of Near-Eastern founders. *Mol Biol Evol.* 29:2101–2104.

Bradley DG, MacHugh DE, Cunningham P, Loftus RT. 1996. Mitochondrial diversity and the origins of African and European cattle. *Proc Natl Acad Sci U S A.* 93:5131–5135.

Bruford MW, Bradley DG, Luikart G. 2003. DNA markers reveal the complexity of livestock domestication. *Nat Rev Genet.* 4:900–910.

Campbell CD, Chong JX, Malig M, et al. (13 co-authors). 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet.* 44:1277–1281.

de Roos APW, Hayes BJ, Spelman RJ, Goddard ME. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus Cattle. *Genetics* 179:1503–1512.

Finlay EK, Gaillard C, Vahidi SMF, Mirhoseini SZ, Jianlin H, Qi XB, El-Barody MAA, Baird JF, Healy BC, Bradley DG. 2007. Bayesian inference of population expansions in domestic bovines. *Biol Lett.* 3:449–452.

Gautier M, Faraut T, Moazami-Goudarzi K, et al. (12 co-authors). 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* 177:1059–1070.

Goddard ME. 1992. Optimal effective population size for the global population of black and white dairy cattle. *J Dairy Sci.* 75:2902–2911.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet.* 43:1031–1034.

Grossman SR, Shylakhter I, Karlsson EK, et al. (13 co-authors). 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327:883–886.

Gutierrez J, Altarriba J, Diaz C, Quintanilla R, Canon J, Piedrafita J. 2003. Pedigree analysis of eight Spanish beef cattle breeds. *Genet Sel Evol.* 35:43–63.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13:635–643.

Hill WG. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor Popul Biol.* 8: 117–126.

Hill WG. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genet Res.* 38:209–216.

Ho SYW, Larson G, Edwards CJ, Heupink TH, Lakin KE, Holland PWH, Shapiro B. 2008. Correlating Bayesian date estimates with climatic events and domestication using a bovine case study. *Biol Lett.* 4: 370–374.

Hoberman R, Dias J, Ge B, Harmsen E, Mayhew M, Verlaan DJ, Kwan T, Dewar K, Blanchette M, Pastinen T. 2009. A probabilistic approach for SNP discovery in high-throughput human resequencing data. *Genome Res.* 19:1542–1552.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.

Kim ES, Kirkpatrick BW. 2009. Linkage disequilibrium in the North American Holstein population. *Anim Genet.* 40:279–288.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A.* 99:803–808.

Larkin DM, Daetwyler HD, Hernandez AG, et al. (17 co-authors). 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc Natl Acad Sci U S A.* 109:7693–7698.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496.

Liu G, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, Li C, Song J, Eichler E. 2009. Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* 10:571.

Lohmueller KE, Bustamante CD, Clark AG. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 182: 217–231.

Lynch M. 2008. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol Biol Evol.* 25:2409–2419.

MacEachern S, McEwan J, Goddard M. 2009. Phylogenetic reconstruction and the identification of ancient polymorphism in the Bovini tribe (Bovidae, Bovinae). *BMC Genomics* 10:177.

MacHugh DE, Shriver MD, Loftus RT, Cunningham P, Bradley DG. 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of Taurine and Zebu cattle (Bos taurus and Bos indicus). *Genetics* 146:1071–1086.

MacLeod IM, Meuwissen THE, Hayes BJ, Goddard ME. 2009. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genet Res.* 91:413–426.

Mc Parland S, Kearney JF, Rath M, Berry DP. 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. *J Anim Sci.* 85:322–331.

Meuwissen THE, Goddard M. 2007. Multipoint IBD prediction using dense markers to map QTL and estimate effective population size. *Genetics* 176:2551–2560.

Murray C, Huerta-Sanchez E, Casey F, Bradley DG. 2010. Cattle demographic history modelled from autosomal sequence variation. *Philos Trans R Soc B Biol Sci.* 365:2531–2539.

Nordborg M, Tavare S. 2002. Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18:83–90.

Perkins D. 1969. Fauna of Çatal Hüyük: evidence for early cattle domestication in Anatolia. *Science* 164:177–179.

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Res.* 20: 291–300.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet.* 69:1–14.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–1656.

Ritz LR, Glowatzki-Mullis ML, MacHugh DE, Gaillard C. 2000. Phylogenetic analysis of the tribe Bovini using microsatellites. *Anim Genet.* 31:178–185.

Roach JC, Glusman G, Smit AFA, et al. (15 co-authors). 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H-J, Torroni A, Richards MB. 2010. The Archaeogenetics of Europe. *Curr Biol.* 20:R174–R183.

Stachowicz K, Sargolzaei M, Miglior F, Schenkel FS. 2011. Rates of inbreeding and genetic diversity in Canadian Holstein and Jersey cattle. *J Dairy Sci.* 94:5160–5175.

Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC, Bradley DG. 2001. Genetic evidence for Near-Eastern origins of European cattle. *Nature* 410:1088–1091.

VanRaden PM. 2004. Invited review: selection on net merit to improve lifetime profit. *J Dairy Sci.* 87:3125–3131.

Villa-Angulo R, Matukumalli L, Gill C, Choi J, Van Tassell C, Grefenstette J. 2009. High-resolution haplotype block structure in the cattle genome. *BMC Genetics* 10:19.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A.* 102:18508–18513.

Weigel KA. 2001. Controlling inbreeding in modern breeding programs. *J Dairy Sci.* 84(Suppl):E177–E184.

Young CW, Seykora AJ. 1996. Estimates of inbreeding and relationship among registered Holstein females in the United States. *J Dairy Sci.* 79:502–505.