Linkage disequilibrium and historical effective population size in the Thoroughbred horse

Citation for published version:
Corbin, LJ, Blott, SC, Swinburne, JE, Vaudin, M, Bishop, SC & Woolliams, JA 2010, 'Linkage disequilibrium and historical effective population size in the Thoroughbred horse', *Animal Genetics*, vol. 41, no. Suppt. 2, pp. 8-15. https://doi.org/10.1111/j.1365-2052.2010.02092.x

Digital Object Identifier (DOI):
10.1111/j.1365-2052.2010.02092.x

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Animal Genetics

Linkage disequilibrium and historical effective population size in the Thoroughbred horse

L. J. Corbin[1], S. C. Blott[2], J. E. Swinburne[2], M. Vaudin[2], S. C. Bishop[1], J. A. Woolliams[1]

[1] Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin Biocentre, EH25 9PS, UK.
[2] Animal Health Trust, Newmarket, CB8 7UU, UK

Keywords: effective population size; horse; linkage disequilibrium; population history; single nucleotide polymorphism; thoroughbred

Abstract

Many genomic methodologies rely on the presence and extent of linkage disequilibrium (LD) between markers and genetic variants underlying traits of interest, but the extent of LD in the horse has yet to be comprehensively characterized. In this study, we evaluate the extent and decay of LD in a sample of 817 Thoroughbreds. Horses were genotyped for over 50,000 single nucleotide polymorphism (SNP) markers across the genome, with 34,848 autosomal SNPs used in the final analysis. Linkage disequilibrium, as measured by the squared correlation coefficient (r(2)), was found to be relatively high between closely linked markers (>0.6 at 5 kb) and to extend over long distances, with average r(2) maintained above non-syntenic levels for single nucleotide polymorphisms (SNPs) up to 20 Mb apart. Using formulae which relate expected LD to effective population size (N(e)), and assuming a constant actual population size, N(e) was estimated to be 100 in our population. Values of historical N(e), calculated assuming linear population growth, suggested a decrease in N(e) since the distant past, reaching a minimum twenty generations ago, followed by a subsequent increase until the present time. The qualitative trends observed in N(e) can be rationalized by current knowledge of the history of the Thoroughbred breed, and inbreeding statistics obtained from published pedigree analyses are in agreement with observed values of N(e). Given the high LD observed and the small estimated N(e), genomic methodologies such as genomic selection could feasibly be applied to this population using the existing SNP marker set.

Introduction

Linkage disequilibrium (LD) describes the non-random association of alleles at different loci and can result from processes such as migration, selection and genetic drift in finite populations (Wang 2005). The efficacy of genomic techniques such as genome-wide association studies (GWAS), marker-assisted selection (MAS) and genomic selection is dependent on the extent of LD and its rate of decline with distance between loci within the population under study. The recent release of the Illumina Equine SNP50 Genotyping BeadChip has increased the potential for such techniques to be applied to the horse. Researchers have already begun to make use of the single nucleotide polymorphism (SNP) chip in GWAS (Bannasch *et al.* 2009; Blott *et al.* 2009; Lykkjen *et al.* 2009), and the opportunity exists to use validated loci for MAS in the future, as some success has already been seen in the localization of QTL for simple diseases (Drögemuller *et al.* 2009; Eberth *et al.* 2009; Gabreski *et al.* 2009). However, as has been shown in human studies, when applied to complex diseases, the outcomes of GWAS are generally less successful (Manolio *et al.* 2009), and thus the genomic selection techniques of Meuwissen *et al.* (2001) may become more attractive. The opportunities will depend on the extent of LD, and therefore the characterization of LD exhibited with the current SNP50 BeadChip will assist in planning future studies of complex traits and in the development of genomic tools.

Linkage disequilibrium structure can also provide insights into the evolutionary history of a population. The strength of LD at different genetic distances between loci can be used to infer ancestral effective population size (N_e), where N_e is the number of individuals in an idealized population that would give rise to the same rate of inbreeding as observed in the actual breeding population (Falconer & Mackay 1996). Deterministic equations derived by Daetwyler *et al.* (2009) show that, once the N_e for a population is known, the accuracy of genomic selection for a range of scenarios can be calculated. The pattern of historical N_e in domestic livestock populations can also help us to understand the impact of selective breeding strategies on the genetic variation present in populations and can provide an insight into the level of inbreeding in populations for which pedigrees are incomplete or unavailable.

The pattern of LD in the Thoroughbred has yet to be comprehensively characterized, and predictions of N_e are limited to those inferred from pedigree data, which itself may be inaccurate (Hill *et al.* 2002). An early study by Tozaki *et al.* (2005), based on 300 horses, concluded that useful LD in the Thoroughbred extends up to 7 cM, but this study covered only one small region of the genome. More recently, Wadeet *al.* (2009) investigated LD across ten 2- Mb regions in a number of different horse breeds, using sample sizes of 24 horses per breed. In contrast, genome-wide LD in livestock populations has been the focus of numerous studies (McRae *et al.* 2002; Heifetz *et al.* 2005;Khatkar *et al.* 2008). Studies have also been carried out to evaluate the historical N_e of a variety of cattle breeds, all of which suggest a continuous decrease in N_e since the time of domestication (Thévenon *et al.* 2007; de Roos *et al.* 2008; Qanbari *et al.* 2009).

The objective of this study was to characterize LD in a large sample of Thoroughbred horses using data generated from the Illumina Equine SNP50 BeadChip and to consider the results in the context of genomic methodologies. The decline of LD over distance is used to predict the effective population size both assuming a constant population size and assuming linear growth. These results are considered in the context of current knowledge of the establishment of the Thoroughbred breed.

Materials and Methods

Genotypic data

The data for this study originated from two disease association studies, and the dataset comprises genotype data for 817 UK Thoroughbreds. Whilst the original data collection required horses to be categorized as cases or controls for the diseases of interest, for the purpose of this study, the horses were treated as a single population sample. Blood samples were collected in EDTA, and DNA was extracted either by Tepnel (http://www.tepnel.com/dna-extraction-service.asp) or at the AHT using Nucleon BACC DNA extraction kits (http://www.tepnel.com/dna-extraction-kits-blood-and-cell-culture.asp). A small dilution of each sample was prepared at 70 ng/ul and submitted for genotyping to Cambridge Genomic Services (http://www.cgs.path.cam.ac.uk/services/snp-genotyping/services.html). The Illumina Equine SNP50 Genotyping BeadChip (http://www.illumina.com/documents/products/datasheets/datasheet_equine_snp50.pdf) was used. This comprises 54 602 single nucleotide polymorphisms (SNPs) located across all autosomes and the X chromosome. These were selected from the database of over one million SNPs (http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2/) generated during the sequencing of the horse genome (http://www.broadinstitute.org/mammals/horse).

Genotyping data was analysed using the Illumina GenomeStudio genotyping module, and a series of quality control metrics were used to identify poorly performing SNPs. Quality control (QC) at this stage led to the removal of 7.1% (*n* = 3895) of SNPs from the analysis owing to poor genotyping quality (see Table S1). Further QC carried out as part of this study led to the removal of an additional 21 SNPs, which were genotyped in less than 95% of samples. The genotyping rate once these exclusions had been made was greater than 99%, with no individuals having more than 10% of SNPs missing. Markers which deviated significantly from Hardy–Weinberg equilibrium (HWE) ($P < 0.0001$) were excluded (Purcell *et al.* 2007; Purcell 2009). Previous studies have demonstrated that including markers with low minor allele frequencies (MAF) can bias LD estimates (Goddard *et al.* 2000; Qanbari *et al.* 2009; Toosi *et al.* 2010), therefore a MAF threshold of 0.10 was imposed on the data. Outcomes of the HWE and MAF screening are given in the results. This study used only autosomal markers.

Linkage disequilibrium

The measurement of LD used throughout this study is the squared correlation coefficient between SNP pairs (r^2) (Hill & Robertson 1968), computed as:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}, \quad (1)$$

where $D = pAB\text{-}pApB$ and *pA*, *pa*, *pB* and *pb* are the frequencies of alleles A, a, B and b, respectively. An EM algorithm (Weir 1996) was implemented to estimate haplotype frequencies. r^2 was calculated (to four decimal places) for all syntenic marker pairs. Individuals with a missing genotype for a given marker were excluded when calculating LD for that marker. Details of the physical position of the markers can be found in Illumina product literature (http://www.illumina.com/documents/products/marker_lists/marker_list_equineSNP50.xls).

To accommodate the large range of marker distances observed and to enable clear presentation of results showing LD in relation to physical distance between markers, SNP pairs were divided into three distance classes and subsequently put into 87 distance bins, with bin ranges dependent on the class (see Table S2). The mean r^2 for each of the distance bins was then plotted against the median of the distance bin range (Mb). This analysis was carried out on a chromosome by chromosome basis; the pooled results are presented here.r^2 was also calculated for a random selection of non-syntenic SNPs. Thirty SNPs per autosome were randomly selected, and r^2 values were calculated for all non-syntenic markers, resulting in a total of 418 500 pairwise comparisons.

Modelling of decline of linkage disequilibrium with distance

Under the assumption of an isolated population with random mating, Sved (1971) derived an approximate expression for the expectation of r^2:

$$E\ (\mathrm{r}^2) = (1+4N\mathrm{c})^{-1} \quad (2)$$

where N is effective population size, and c is the recombination frequency. In this paper, as in previous studies (Hayes *et al.* 2003; Tenesa*et al.* 2007; Thévenon *et al.* 2007; de Roos *et al.* 2008; Qanbari *et al.* 2009; Villa-Angulo *et al.* 2009), c is replaced by map distance in Morgans. This is justified by the approximation of the more precise equation for $E(r^2)$ given by Sved (1971), where $c(1\text{-}c/2)(1\text{-}c)^{-2}$ replaces$c$. This function is a reasonable approximation to both Haldane and Kosambi map distance for $0 \leq c \leq 0.5$. Based on this formula, a non-linear least squares approach to statistically model the observed r^2 was implemented within R (R Development Core Team 2009) using the following model:

$$y_\mathrm{i} = 1/(\mathrm{a}+4\mathrm{bd}_\mathrm{i}) + \mathrm{e}_\mathrm{i} \quad (3)$$

where *yi* is the value of r^2 for SNP pair *i*, at linkage distance *di* in Morgans. Parameters *a* and *b* were estimated iteratively using least squares. Chromosome-specific megabase-to-centimorgan conversion rates were calculated based on total physical chromosome length, as stated on the NCBI website (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9796) and total chromosome genetic length from the equine linkage map (Swinburne *et al.* 2006) (see Table S3). Marker pairs with <100 bp between them were excluded from this analysis, as it has been suggested that Sved's (1971) model is not appropriate for very small values of *c* (Hill 1981; de Roos *et al.*2008), and at small distances gene conversion contributes to the breakdown of LD (Frisse *et al.* 2001; Ardlie *et al.* 2002; Tenesa *et al.*2007). The minimum MAF threshold of 0.10 was also applied here, as Eq. 2 may be a poor approximation when allele frequencies are close to zero (Hill 1981; Hudson 1985). This model was applied to data for each autosome in turn, and parameter estimates were combined by meta-analysis in R (R Development Core Team 2009) using an inverse variance method for pooling and a random effects model based on the DerSimonian–Laird method (DerSimonian & Laird 1986) (see Appendix S1 for further details).

Ancestral effective population size estimation

Rearrangement of Eq. 2 allows the prediction of effective population size at a given point in time, expressed as generations in the past (Hayes *et al.* 2003; de Roos *et al.* 2008):

$$N_T(t) = (4c)^{-1} \left[(r_c^2)^{-1}-1\right] \quad (4)$$

where *NT* is the effective population size *t* generations ago, *c* is the distance between markers in Morgans, is the mean value of r^2 for markers *c* Morgans apart, and $c = (2t)^{-1}$ when assuming linear growth (Hayes *et al.* 2003). As previously mentioned, marker pairs with less than 100 bp between them and SNPs with MAF <0.10 were excluded from this analysis. To compute *NT*, the number of prior generations was selected and a suitable range for *c* was calculated (see Table S4). The binning process was designed to ensure sufficient SNP pair comparisons within each bin to get a representative estimate of r^2. The mean distance and mean r^2 between marker pairs in each bin was then computed. This process was carried out for each chromosome in turn and also for markers pooled across chromosomes, as is suggested by Hayes *et al.* (2003) to reduce the variability of estimates of N_T caused by finite population size.

Results

Genotypic data

Of the 52 603 autosomal SNPs genotyped, 34 848 (66.2%) remained after filtering, resulting in more than 20 million pairwise comparisons. Of those SNPs excluded, 173 SNPs were excluded for not being in HWE and a further 13 372 for having a MAF <0.10 (4086 of these were in fact monomorphic in our sample). The number of SNPs per autosome remaining after exclusions ranged from 416 to 2,760 and was closely related to chromosome length, as shown in Fig. 1. The average distance between adjacent markers (±SD) was 64.05 ± 86.84 kb, with the distance between adjacent SNPs ranging from 1 bp to 2849 kb. The MAF of remaining SNPs followed a uniform distribution and averaged (±SD) 0.30 ± 0.12.

Linkage disequilibrium

Linkage disequilibrium declined with increasing distance between SNP pairs, as shown in Fig. 2a, b and c. The most rapid decline was seen over the first 0.2 Mb, with the mean r^2 decreasing by more than half over this period. The mean r^2 then decreased more slowly with increasing distance, and the decline in LD was almost linear with logtransformed distance (Fig. 3). The coefficient of variation of r^2increased from 0.6 at 5 kb to a maximum of 2.2 at 15 Mb, subsequently decreasing and remaining below 1.5 for distances greater than 50 Mb. A total of 10 130 SNP pairs were in complete LD ($r^2 = 1$); 5139 of these were adjacent pairs. The mean (±SD) r^2 between random non-syntenic markers was $0.0018 \pm 2.49 \times 10^{-3}$ and was similar to that observed between syntenic markers at distances greater than 100 Mb (Fig. 2c).

Modelling of decline of linkage disequilibrium with distance

The non-linear regression modelling of the decline of LD with distance resulted in both *a* and *b* being significantly different from zero. The mean estimate and 95% confidence interval by meta-analysis across autosomes for *a* was 2.25 [2.18; 2.33] and for *b* was 103.1 [95.8; 110.3]. The line of predicted r^2 from the non-linear regression equation only approximately follows that of the mean observed r^2, with the greatest discrepancy occurring at distances less than 0.03 Mb, as shown in Fig. 3. Parameter *b* showed greater variability between chromosomes than parameter *a*, although estimates for both parameters showed an approximately symmetrical distribution about the median. A significant negative correlation ($P < 0.01$) was observed between estimates of *b* and chromosome length (cM), but there was no such relationship between estimates of *a* and chromosome length (cM) (Fig. 4a, b). The interpretation of *b* as an estimate of effective population size is considered in the discussion.

Ancestral effective population size

We observed an initial pattern of decreasing N_e with values of over 3000 estimated in the distant past (see Figure S1) and values closer to 100 estimated at 20 generations ago (Fig. 5). Our results suggest that an increase in N_e has occurred over the past ten generations, with a maximum of approximately 190 observed two generations ago. Variation in predicted N_e across chromosomes was greatest for estimates corresponding to the most recent ten generations and those corresponding to the most distant generations (over 800 generations ago) (see Figure S2).

Discussion

This study provides an overview of LD in the Thoroughbred using a high density SNP panel. Validation work by Khatkar *et al.* (2008) on their cattle data suggests that our sample size of more than 800 horses is more than sufficient to obtain an unbiased picture of LD in our population. The pattern of decline of LD with distance in this population is consistent with that reported by Wade *et al.* (2009) in a sample of 24 Thoroughbreds, with both data sets exhibiting a decrease in r^2 from ~0.6 to 0.2 when the distance between markers is increased to 0.5 Mb. The LD observed is higher at short distances and more extensive than that observed in human populations (Shifman *et al.* 2003). Linkage disequilibrium declines more slowly in our population than in the range of cattle populations studied by de Roos *et al.* (2008), with r^2 remaining above 0.3 for distances up to 185 kb in our data, compared with a maximum distance of 35 kb in the cattle data.

The mean value of r^2 between non-syntenic SNPs was 0.0018, and this provides an approximation of the LD that can be expected by chance, assuming that the markers used have not undergone simultaneous selection. The value observed here is lower than, but of a similar magnitude to, that observed by Khatkar *et al.* (2008) in a sample of over 1500 cattle (0.0032). The mean non-syntenic r^2 value reflects both sampling of animals and genetic sampling (drift), and hence may be expected to decrease with increases in both sample size and N_e. Therefore, the larger non-syntenic value in Australian Holstein–Friesian cattle may more reflect a lower N_e in this cattle population. The low LD seen between non-syntenic SNPs in our population suggests that the LD created by admixture during breed formation (Hill *et al.* 2002) has declined to negligible levels for these markers. A similar decline of LD between non-syntenic markers was observed in Coopworth sheep approximately ten generations after the foundation of the breed through crossing (McRae *et al.* 2002). At distances greater than 100 Mb, average r^2 between syntenic SNPs is reduced to non-syntenic or background levels, and is no longer a function of distance. This is expected, as the recombination rate at such distances approaches 0.5.

By using Sved's (1971) formula for the expectation of r^2, a non-linear regression model was fitted to the data to describe the relationship between linkage distance and LD. Without making any assumptions about the value of r^2 at the intercept, estimates of a and b, as predicted using Eq. 3 and averaged over all autosomes, were 2.25 and 103, respectively. Parameter a determines the value of expected r^2 when the line crosses the y-axis (i.e. when the distance between markers is effectively zero). Our estimate of a supports an alternative version of Sved's (1971) equation, derived by Tenesa *et al.* (2007), which takes into account mutation and puts a equal to two, whilst at the same time raising the question of whether fixing a to unity in the model (as in Abasht *et al.* (2009), Toosi *et al.* (2010) and Zhao *et al.*(2005)) is appropriate. The impact of such model assumptions are explored in Corbin *et al.* (2010). The heterogeneity of variance associated with the observed r^2, such that the variance of r^2 declined with increasing distance between markers, may also have impacted on our results. We observed a significant negative relationship between chromosome length (cM) and estimates of *b* from the non-linear model, suggesting LD is higher in longer chromosomes. This contrasts with the findings of Tenesa *et al.* (2007), who observed a positive relationship, but is in keeping with the observations of Khatkar *et al.* (2008) and Muir *et al.* (2008) in domestic livestock species. Our estimate of *b* (103) is an estimate of N_e assuming constant population size. However, this assumption is difficult to sustain, and therefore, *b* more likely represents a conceptual average N_e over the period inferred from the marker distance range, for example seeToosi *et al.* (2010). For this reason, Fig. 5 shows the results following the approach of Hayes *et al.* (2003) by calculating historical N_e, assuming linear population growth. The pattern observed shows a decrease in N_e up until around 20 generations ago, followed by an increase until one generation ago. The interpretation of such trends is difficult, with the observed dip in N_e potentially representing any one of a number of scenarios, including a founder event, an immigration event, a hybridization event or any combination of these (Wang 2005). Therefore, it is useful to consider our observation in the context of what is known about the Thoroughbred's demographic history.

Documentary evidence suggests that the Thoroughbred was derived from a cross between sires originating from the Mediterranean Middle East and British native breeds, and the breed was established during the seventeenth century (Hill *et al.* 2002). It is not clear from published literature what effects an admixture like this would have on patterns of estimated N_e prior to the crossing event, although clues may be observed in Toosi *et al.* (2010). However, what may be predicted is that such a crossing event would appear as a bottleneck in the population, creating an initially high level of LD in the beginning. Therefore, one might infer from our results that the lowest point of the curve reflects the point at which the breed was formed; this approximately coincides with the findings of Mahon & Cunningham (1982) that Thoroughbreds born in the 1960s were separated from seventeenth century founders by an average of 21.5 generations. Cunningham*et al.* (2001) also found evidence for a population bottleneck at the time of breed formation.

The reliability of this method depends both on the technical implementation (Corbin *et al.* 2010) and, as discussed above, on the demographic history of the breed. Some calibration of the accuracy of the N_e profile presented can be obtained by comparison with values obtained from pedigree analyses. For example, Cunningham *et al.* (2001) calculate the effective number of studbook founders of the Thoroughbred to be 28.2. As this relies on calculating the long-term contributions of the founders, quantitative genetic theory (Woolliams & Bijma 2000) suggests that the N_e for this generation is twice this value if in HWE, providing an estimate of 56 soon after breed formation. This may be compared with the minimum N_e of 88 obtained in this analysis, which gives fair agreement. A further estimate of reliability can be obtained by comparing the mean inbreeding of 0.125 (SE 0.005) obtained by Mahon & Cunningham (1982) for the 21.5 generations from breed foundation to 1964 with the accumulated inbreeding for generations four to 25 (assuming four generations since 1964) using $1-\prod_{4}^{25}(1-1/2N_e)$, with N_e being estimated from Fig. 5. The value obtained of 0.112 is remarkably close. Therefore, our minimum of $N_e \approx 90$ is of the correct magnitude, and the increase in N_e over the last ten generations may be explained by an increase in actual population size. In Thoroughbreds, with low reproductive rate of the mare and the ban upon use of artificial insemination, there is a greater likelihood that increases in census size will be translated into effective population sizes. The trend in N_e observed in the most recent generations should be interpreted with caution because of the technical limitations of the methods.

Implications for genome-wide association studies, marker-assisted selection and genomic selection

The extent of LD in a population can be used to estimate the SNP density required for GWAS studies to be effective, as well as giving some indication as to the likely precision with which the QTL region will be located. The required sample size is said to be inflated by $1/r^2$, when it is necessary to rely on marker-QTL LD, rather than on the QTL itself (Du *et al.* 2007), and this has prompted authors to propose thresholds for useful LD. The term ‘useful LD’ has been described as the proportion of QTL variance explained by a marker (Zhao *et al.*2005), and the consensus is that an average $r^2 > 0.3$ will permit reasonable sample sizes to be employed for GWAS (Ardlie *et al.* 2002; Du*et al.* 2007; Khatkar *et al.* 2008). In this dataset, markers 185 kb apart achieve an average LD of $r^2 = 0.3$, and this corresponds to approximately 14 500 evenly spaced markers across the genome. However, because markers with $r^2 = 1$ will likely be excluded in genomic selection, and given the high variability of r^2 values at small distances, this is likely to be an underestimation of the actual number of SNPs needed. Indeed, in this study, whilst markers separated by less than 250 kb had a mean r^2 of 0.32 (after the exclusion of those pairs in complete LD), less than half the SNP pairs exhibited r^2 values of greater than 0.3. With MAS also relying on close and consistent linkage between markers and QTL, the high LD observed here is promising. Genomic selection (GS) appears to be effective at lower average r^2 than that required for GWAS, with simulation results demonstrating accuracies of up to 0.65 with an average r^2 between adjacent SNPs as low as 0.2 and a trait heritability of 0.1 (Calus *et al.* 2008). Deterministic equations derived by (Daetwyler *et al.* 2009) demonstrate that the accuracy of GS can be expressed as a function of the effective number of loci (M_e) in a population. M_e relates to the number of independent chromosome segments and, given our current N_e estimate of ∼180 and assuming a random mating population, the M_e for our population is ∼1500 (Meuwissen 2009). Thus, we are now able to predict the potential accuracy of GS in this population for a range of scenarios.

In summary, we used dense SNP genotype data to characterize LD and make inferences regarding ancestral N_e for a large sample of Thoroughbred horses. In the population studied, LD extended for long distances, reaching baseline levels at around 50 Mb. From the decay in LD with distance, we inferred ancestral N_e and observed a decrease in N_e since the distant past, which reached a minimum of ∼90 20 generations ago, followed by an increase until the present time. Such an approach could be used to investigate the demographic histories and rates of inbreeding of horse breeds with less extensive pedigree records than the Thoroughbred. The results indicate that genomic methodologies which are reliant on LD between markers and QTL have the potential to perform well within Thoroughbred populations genotyped for the 50-K SNP chip.

Acknowledgements

The authors are financially supported by the British Equestrian Federation, the Biosciences Knowledge Transfer Network and the Biotechnology and Biological Sciences Research Council (BBSRC). Genotyping was funded by the Horserace Betting Levy Board and the Thoroughbred Breeders' Association. LJC would like to thank W.G. Hill and I.M.S. White for helpful discussions.

Conflicts of interest

L.J. Corbin received a bioscience CASE Studentship PhD. S.C. Blott received grants from Horserace Betting Levy Board, The Kennel Club. S.C. Bishop received grants from Quality Meat Scotland, EBLEX. J.A.W. received grants from Aviagen, Cobb, Quality Meat Scotland, EBLEX. The remaining authors have no conflicts to declare.

References

- Abasht B., Sandford E., Arango J. et al. (2009) Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. BMC Genomics 10(Suppl. 2), S2.

- Ardlie K.G., Kruglyak L. & Seielstad M. (2002) Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3, 299–309.

- Bannasch D., Lohi H., Wade C.M. et al. (2009) Genome wide association analysis of a behavioural vice in horses. In: Proceedings of the 8th Dorothy Havemeyer Foundation International Equine Genome Mapping Workshop, July 2009, Newmarket, p. 15. R&W Communications, Newmarket.

- Blott S., Boursnell M., Bramlage L. et al. (2009) Whole genome association mapping for catastrophic fracture, RER and OCD in Thoroughbred horses. In: Proceedings of the 8th Dorothy Havemeyer Foundation International Equine Genome Mapping Workshop, July 2009, Newmarket, p. 17. R&W Communications, Newmarket.

- Calus M.P.L., Meuwissen T.H.E., De Roos A.P.W. & Veerkamp R.F. (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178, 553–61.

- Corbin L.J., Bishop S.C., Swinburne J.E., Vaudin M., Blott S.C. & Woolliams J.A. (2010) The impact of method on the estimated effective population size of a Thoroughbred population using genotype data. In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, August 2010. Leipzig.

- Cunningham E.P., Dooley J.J., Splan R.K. & Bradley D.G. (2001) Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. Animal Genetics 32, 360–4.

- Daetwyler H.D., Pong-Wong R., Villanueva B. & Woolliams J.A. (2009) The impact of genetic architecture on genome-wide evaluation methods. In: Genome-Wide Evaluation of Populations, Thesis (PhD) of Daetwyler, H. D., Wageningen University, NL.

- DerSimonian R. & Laird N. (1986) Meta-analysis in clinical trials. Controlled Clinical Trials 7, 177–88.

- Drögemuller M., Drögemuller C., Welle M., Straub R., Poncet P.-A., Rieder S. & Leeb T. (2009) Mapping of Caroli liver fibrosis (CLF) in Franches-Montagnes horses. In: Proceedings of the 8th Dorothy Havemeyer Foundation International Equine Genome Mapping Workshop, July 2009, Newmarket, p. 16. R&W Communications, Newmarket.

- Du F.-X., Clutter A.C. & Lohuis M.M. (2007) Characterizing linkage disequilibrium in pig populations. International Journal of Biological Sciences 3, 166–78.

- Eberth J., Swerczak T. & Bailey E. (2009) Investigation of Dwarfism Among Miniature Horses using the Illumina Horse SNP50 Bead Chip. Journal of Equine Veterinary Science 29, 315.

- Falconer D.S. & Mackay T.F.C. (1996) Introduction to Quantitative Genetics, 4th edn. Pearson Education Limited, Malaysia.

- Frisse L., Hudson R.R., Bartoszewicz A., Wall J.D., Donfack J. & Di Rienzo A. (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. American Journal of Human Genetics 69, 831–43.

- Gabreski N., Brooks S., Miller D. & Anczak D. (2009) Mapping of Lavender Foal Syndrome using the EquineSNP50 Chip. Journal of Equine Veterinary Science 29, 321–2.

- Goddard K.A.B., Hopkins P.J., Hall J.M. & Witte J.S. (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. American Journal of Human Genetics 66, 216–34.

- 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, 635–43.

- Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C.M. & Soller M. (2005) Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics 171, 1173–81.

- Hill W.G. (1981) Estimation of Effective Population-Size from Data on Linkage Disequilibrium. Genetical Research 38, 209–16.

- Hill W.G. & Robertson A. (1968) Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38, 226–31.

- Hill E.W., Bradley D.G., Al-Barody M., Ertugrul O., Splan R.K., Zakharov I. & Cunningham E.P. (2002) History and integrity of thoroughbred dam lines revealed in equine mtDNA variation. Animal Genetics 33, 287–94.

- Hudson R.R. (1985) The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics 109, 611–31.

- Khatkar M., Nicholas F., Collins A., Zenger K., Cavanagh J., Barris W., Schnabel R., Taylor J. & Raadsma H. (2008) Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics 9, 187.

- Lykkjen S., Dolvik N.I., Mickelson J.R. & Roed K.H. (2009) Genetic studies of osteochondrosis dissecans (OCD) in the hock and proximoplantar osteochondral fragments (POF) in the fetlock joints of Norwegian Standardbred trotters. In: Proceedings of the 8th Dorothy Havemeyer Foundation International Equine Genome Mapping Workshop, July 2009, Newmarket, p. 20. R&W Communications, Newmarket.

- Mahon G.A.T. & Cunningham E.P. (1982) Inbreeding and the inheritance of fertility in the thoroughbred mare. Livestock Production Science 9, 743–54.

- Manolio T.A., Collins F.S., Cox N.J. et al. (2009) Finding the missing heritability of complex diseases. Nature 461, 747–53.

- McRae A.F., McEwan J.C., Dodds K.G., Wilson T., Crawford A.M. & Slate J. (2002) Linkage disequilibrium in domestic sheep. Genetics 160, 1113–22.

- Meuwissen T.H.E. (2009) Accuracy of breeding values of ‘unrelated’ individuals predicted by dense SNP genotyping. Genetics Selection Evolution 41, 35.

- Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–29.

- Muir W.M., Wong G.K., Zhang Y. et al. (2008) Review of the initial validation and characterization of a 3K chicken SNP array. World’s Poultry Science Journal 64, 219–26.

- Purcell S. (2009) PLINK (v1.06) http://pngu.mgh.harvard.edu/purcell/plink/

- Purcell S., Neale B., Todd-Brown K. et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559–75.

- Qanbari S., Pimentel E.C.G., Tetens J., Thaller G., Lichtner P., Sharifi A.R. & Simianer H. (2009) The pattern of linkage disequilibrium in German Holstein cattle. Animal Genetics, 10.1111/j.1365-2052.2009.02011.x .

- R Development Core Team (2009) R: A Language and Environment for Statistical Computing (v2.10.0). R Foundation for Statistical Computing, Vienna. http://www.R-project.org

- 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, 1503–12.

- Shifman S., Kuypers J., Kokoris M., Yakir B. & Darvasi A. (2003) Linkage disequilibrium patterns of the human genome across populations. Human Molecular Genetics 12, 771–6.

- Sved J.A. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology 2, 125–41.

- Swinburne J.E., Boursnell M., Hill G. et al. (2006) Single linkage group per chromosome genetic linkage map for the horse, based on two-three-generation, full-sibling, crossbred horse reference families. Genomics 87, 1–29.

- 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, 520–6.

- Thévenon S., Dayo G.K., Sylla S., Sidibe I., Berthier D., Legros H., Boichard D., Eggen A. & Gautier M. (2007) The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. Animal Genetics 38, 277–86.

- Toosi A., Fernando R.L. & Dekkers J.C.M. (2010) Genomic selection in admixed and crossbred populations. Journal of Animal Science 88, 32–46.

- Tozaki T., Hirota K., Hasegawa T., Tomita M. & Kurosawa M. (2005) Prospects for whole genome linkage disequilibrium mapping in thoroughbreds. Gene 346, 127–32.

- Villa-Angulo R., Matukumalli L.K., Gill C.A., Choi J., Van Tassell C.P. & Grefenstette J.J. (2009) High-resolution haplotype block structure in the cattle genome. BMC Genetics 10, 19.

- Wade C.M., Giulotto E., Sigurdsson S. et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–7.

- Wang J.L. (2005) Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society B-Biological Sciences 360, 1395–409.

- Weir B.S. (1996) Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer Associates, Inc, Canada.

- Woolliams J.A. & Bijma P. (2000) Predicting rates of inbreeding in populations undergoing selection. Genetics 154, 1851–64.

- Zhao H., Nettleton D., Soller M. & Dekkers J.C.M. (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genetical Research 86, 77–87.

Legends

Figure 1. Chromosome length (Mb) and the number of single nucleotide polymorphisms (SNPs) per chromosome.

Figure 2. Average linkage disequilibrium (solid line) and the 25th and 75th percentiles (dashed lines) (measured by r^2), plotted against the median of the distance bin range (Mb). Percentiles Calculated for each bin in turn. (a) Distance range from 0 to 0.5 Mb. r^2 values averaged using bins of 0.01 Mb and pooled over autosomes (minimum of 4900 SNP pairs per distance bin). (b) Distance range from 0.5 to 20.5 Mb. r^2 values averaged using bins of 1.0 Mb and pooled over autosomes (minimum of 196 000 SNP pairs per distance bin). (c) Distance range from 20.5 to 190 Mb. r^2 values averaged using bins of 10.0 Mb and pooled over autosomes (minimum of 4375 SNP pairs per distance bin).

Figure 3. Predicted r^2 versus observed r^2 against mean distance between markers (cM) (on a log scale). Predicted r^2 calculated using Eq. 3 with$a$ = 2.25 and b = 103.

Figure 4. Parameter estimates from model (3) plotted against chromosome length (cM) according to the equine linkage map (Swinburne *et al.* 2006). (a) Estimates of a plotted against chromosome length (cM). (b) Estimates of $b$plotted against chromosome length (cM).

Figure 5. Average estimated effective population size plotted against generations in the past, truncated at 100 generations.

Fig.1

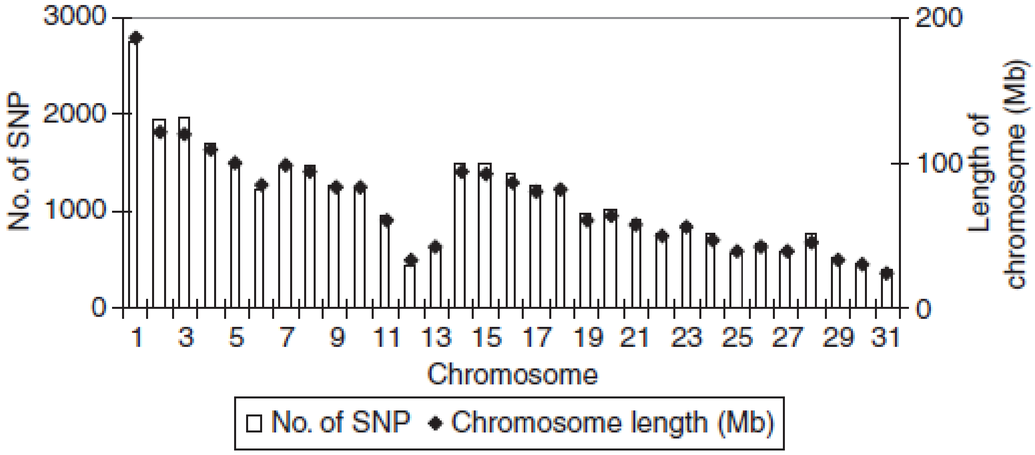

3000
2000
1000
0
No. of SNP
200
100
0
Length of chromosome (Mb)
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
Chromosome
□ No. of SNP ◆ Chromosome length (Mb)

Fig.2

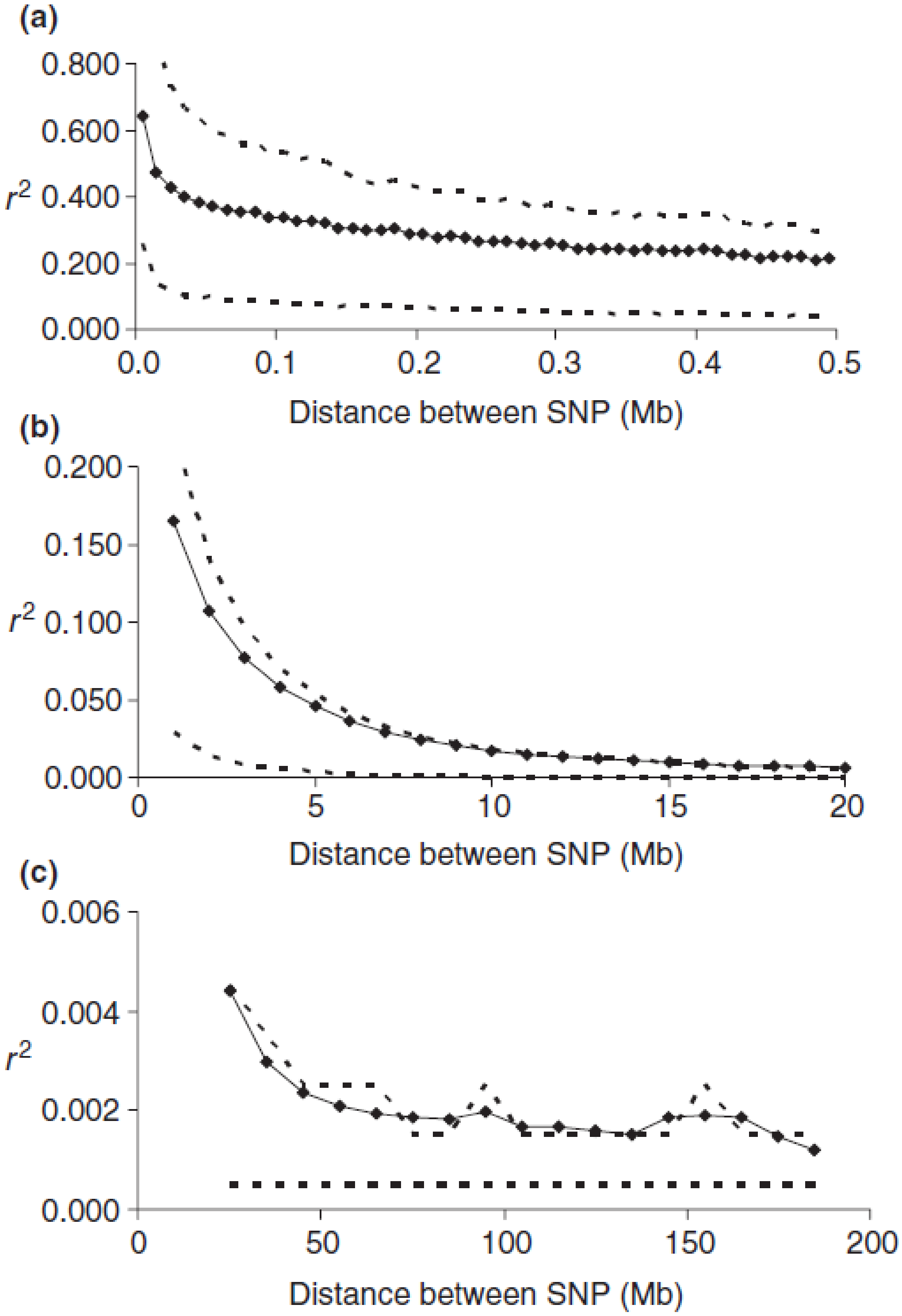

(a)
0.800
0.600
r^2 0.400
0.200
0.000
0.0
0.1
0.2
0.3
0.4
0.5
Distance between SNP (Mb)
(b)
0.200
0.150
r^2 0.100
0.050
0.000
0
5
10
15
20
Distance between SNP (Mb)
(c)
0.006
0.004
r^2
0.002
0.000
0
50
100
150
200
Distance between SNP (Mb)

Fig.3

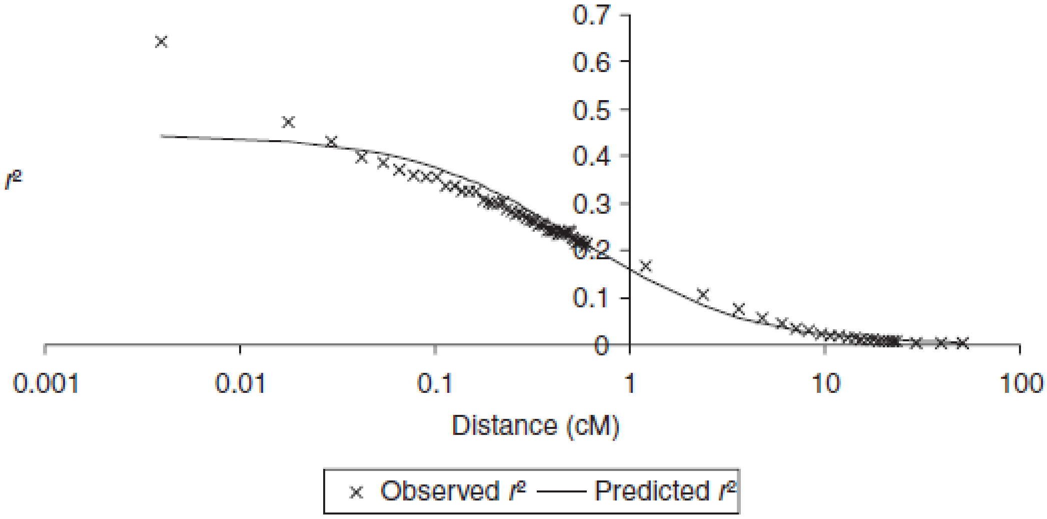

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
r^2
0.001
0.01
0.1
1
10
100
Distance (cM)
× Observed r^2 —— Predicted r^2

Fig.4

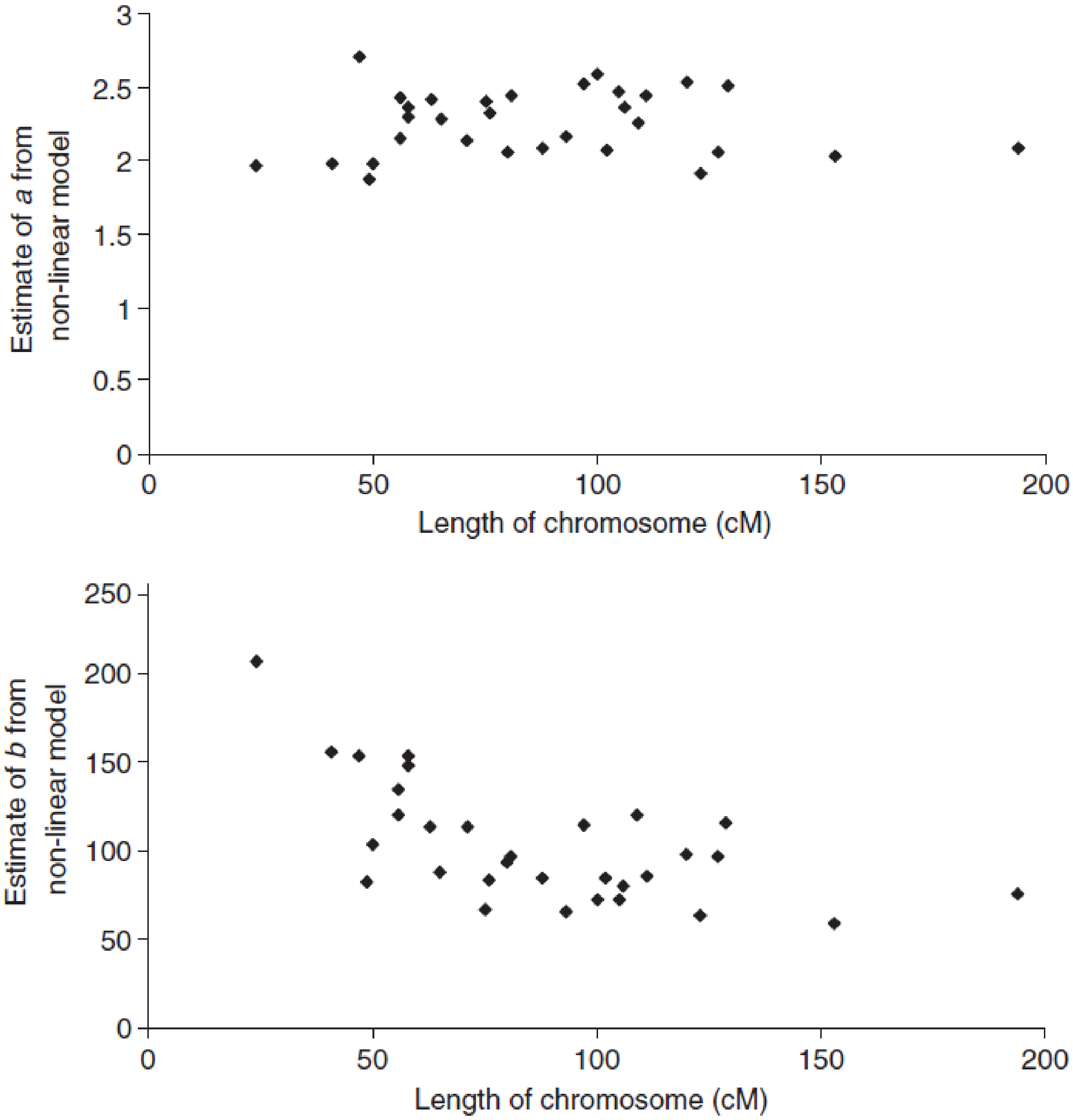
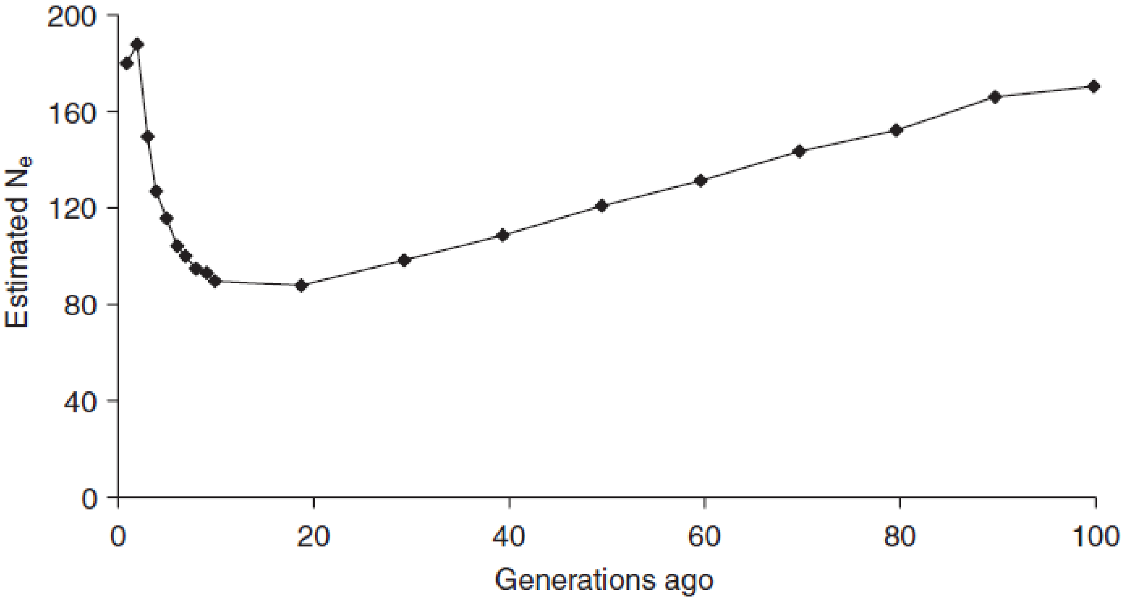

Estimate of *a* from non-linear model
3
2.5
2
1.5
1
0.5
0
0
50
100
150
200
Length of chromosome (cM)
Estimate of *b* from non-linear model
250
200
150
100
50
0
0
50
100
150
200
Length of chromosome (cM)

Fig.5



Estimated N_e
200
160
120
80
40
0
0
20
40
60
80
100
Generations ago