THE UNIVERSITY
of EDINBURGH

The Application of Genomic Technologies to the Horse

Laura J. Corbin

Doctor of Philosophy – University of Edinburgh - 2013

Abstract

The publication of a draft equine genome sequence and the release by Illumina of a 50,000 marker single-nucleotide polymorphism (SNP) genotyping chip has provided equine researchers with the opportunity to use new approaches to study the relationships between genotype and phenotype. In particular, it is hoped that the use of high-density markers applied to population samples will enable progress to be made with regard to more complex diseases. The first objective of this thesis is to explore the potential for the equine SNP chip to enable such studies to be performed in the horse. The second objective is to investigate the genetic background of osteochondrosis (OC) in the horse. These objectives have been tackled using 348 Thoroughbreds from the US, divided into cases and controls, and a further 836 UK Thoroughbreds, the majority with no phenotype data. All horses had been genotyped with the Illumina Equine SNP50 BeadChip.

Linkage disequilibrium (LD) is the non-random association of alleles at neighbouring loci. The reliance of many genomic methodologies on LD between neutral markers and causal variants makes it an important characteristic of genome structure. In this thesis, the genomic data has been used to study the extent of LD in the Thoroughbred and the results considered in terms of genome coverage. Results suggest that the SNP chip offers good coverage of the genome. Published theoretical relationships between LD and historical effective population size (N_e) were exploited to enable accuracy predictions for genome-wide evaluation (GWE) to be made. A subsequent in-depth exploration of this theory cast some doubt on the reliability of this approach in the estimation of N_e, but the general conclusion that the Thoroughbred population has a small N_e which should enable GWE to be carried out efficiently in this population, remains valid. In the course of these studies, possible errors embedded within the current sequence assembly were identified using empirical approaches.

Osteochondrosis is a developmental orthopaedic disease which affects the joints of young horses. Osteochondrosis is considered multifactorial in origin with a variety of environmental factors and heredity having been implicated. In this thesis, a

genome-wide association study was carried out to identify quantitative trait loci (QTL) associated with OC. A single SNP was found to be significantly associated with OC. The low heritability of OC combined with the apparent lack of major QTL suggests GWE as an alternative approach to tackle this disease. A GWE analysis was carried out on the same dataset but the resulting genomic breeding values had no predictive ability for OC status. This, combined with the small number of significant QTL, indicates a lack of power which could be addressed in the future by increasing sample size. An alternative to genotyping more horses for the 50K SNP chip would be to use a low-density SNP panel and impute remaining genotypes. The final chapter of this thesis examines the feasibility of this approach in the Thoroughbred. Results suggest that genotyping only a subset of samples at high density and the remainder at lower density could be an effective strategy to enable greater progress to be made in the arena of equine genomics. Finally, this thesis provides an outlook on the future for genomics in the horse.

Contents

Preface ..... **i**

Acknowledgements ..... **ii**

List of Publications ..... **iii**

Chapter 1: General introduction **1**

1.1 From genetics to genomics in the horse 1

1.2 Genomics, disease and the horse 6

1.3 Introduction to the disorder osteochondrosis 11

1.3.1 Clinical background 11

1.3.2 Prevalence 13

1.3.3 Genetic control 13

1.3.4 The phenotype 16

1.3.5 Quantitative trait loci mapping 20

1.4 Thesis outline 22

Chapter 2: Description of dataset **25**

2.1 Data source 25

2.2 Genotyping 25

2.3 Preliminary data cleansing 28

2.4 Phenotypic enrichment of the osteochondrosis dataset 29

2.5 Phenotypic Summary 32

Chapter 3: Linkage disequilibrium and effective population size in the Thoroughbred horse **38**

3.1 Introduction 38

3.2 Materials and methods 39

3.2.1 Genotypic data 39

3.2.2 Linkage disequilibrium 40

3.2.3 Modelling decline of linkage disequilibrium with distance 41

3.2.4 Ancestral effective population size estimation 44

3.2.5 Sample size comparison 44

3.3 Results 46

3.3.1 Genotypic data 46

3.3.2 Linkage disequilibrium 46

3.3.3 Modelling of decline of linkage disequilibrium with distance 49

3.3.4 Ancestral effective population size 49
3.3.5 Sample size comparison 52
3.4 Discussion 55
3.4.1 Implications for genome-wide association studies, marker assisted selection and genome-wide evaluation 59
Chapter 4: The estimation of effective population size using linkage disequilibria: A methodological review 61
4.1 Introduction 61
4.2 Materials and methods 63
4.2.1 Theory for constant effective population size 63
4.2.1.1 An expression for $E[r^2]$ and its interpretation 63
4.2.1.2 Sampling effects 65
4.2.2 Extension to variable effective population size 66
4.2.3 Data sets investigated 67
4.2.3.1 Simulated data 67
4.2.3.2 Equine data 68
4.2.4 Data analysis 68
4.2.4.1 Estimating linkage disequilibrium 68
4.2.4.2 Estimating constant effective population size 69
4.2.4.3 Estimating variable effective population size 69
4.3 Results 71
4.3.1 Simulated data 71
4.3.1.1 Impact of minor allele frequency threshold 71
4.3.1.2 Impact of adjustment for sample size 72
4.3.1.3 Impact of form of $f(c)$ and value of a 77
4.3.1.4 Summary of models fitted to simulated data 77
4.3.2 Equine data 80
4.4 Discussion 83
Chapter 5: The estimation of historical effective population size using a composite-likelihood methodology 90
5.1 Introduction 90
5.2 Materials and methods 91
5.2.1 Data simulation 91
5.2.2 Data analysis 92
5.2.2.1 Composite-likelihood estimator (CLE) 92

5.2.2.1.1 Derivation of sample probabilities 92
5.2.2.1.2 Summary of linkage disequilibrium 94
5.2.2.1.3 Estimation of effective population size 95
5.2.2.2 Syntenic linkage disequilibrium (SLD) 95
5.2.2.2.1 Summary of linkage disequilibrium 95
5.2.2.2.2 Estimation of effective population size 95
5.3 Results 96
5.3.1 Linkage disequilibrium 96
5.3.2 Composite-likelihood estimator (CLE) 99
5.3.3 Syntenic linkage disequilibrium (SLD) 105
5.4 Discussion 109
Chapter 6: The identification of SNPs on the Equine SNP50 BeadChip with indeterminate positions 113
6.1 Introduction 113
6.2 Materials and methods 115
6.3 Results 116
6.3.1 SNPs assigned to the wrong chromosome 116
6.3.2 Within-chromosome discrepancies 120
6.4 Discussion 125
Chapter 7: A genome-wide association study of osteochondritis dissecans in the Thoroughbred 130
7.1 Introduction 130
7.2 Materials and methods 132
7.2.1 Quality control 132
7.2.2 Mixed model analysis 132
7.2.3 Genome-wide association study 134
7.2.4 Evaluation of ECA3 haplotype blocks 134
7.2.5 Testing previously published quantitative trait loci 135
7.3 Results 135
7.3.1 Mixed model analysis 135
7.3.2 Genome-wide association study 138
7.3.3 Evaluation of ECA3 haplotype blocks 139
7.3.4 Testing previously published quantitative trait loci 139
7.4 Discussion 142

Chapter 8: The use of genome-wide evaluation in the prediction of risk for osteochondritis dissecans in the Thoroughbred 148
8.1 Introduction 148
8.2 Materials and methods 150
8.2.1 Samples 150
8.2.2 Quality control 151
8.2.3 Simulation of additive genetic merit and phenotypic performance 151
8.2.4 Bayesian estimation of SNP effects 152
8.2.5 Genomic BLUP procedure 153
8.2.6 Cross-validation procedure 154
8.2.7 Evaluation of the accuracy of genomic estimated breeding values 155
8.3 Results 156
8.3.1 Osteochondritis dissecans trait results 156
8.3.1.1 Bayesian estimation of SNP effects 156
8.3.1.2 Genomic BLUP Procedure 157
8.3.2 Simulated trait 160
8.3.2.1 Bayesian estimation of SNP effects 160
8.3.2.2 Genomic BLUP 160
8.4 Discussion 163
Chapter 9: The utility of low-density genotyping in the Thoroughbred 169
9.1 Introduction 169
9.2 Materials and methods 171
9.2.1 Genotypes 171
9.2.2 Low-density panel SNP selection 173
9.2.3 Imputation 178
9.3 Results 178
9.3.1 Within population assessment 178
9.3.2 Between-population assessment 188
9.3.3 Relative Cost 188
9.4 Discussion 191
Chapter 10: General discussion 199
10.1 Summary 199
10.2 Linkage disequilibrium and effective population size: Theory and practice 200

10.3 Genetic control of osteochondrosis 203
10.4 The future for genomics in the horse 206
References 210
Web references 230
Appendix A: Supplementary documents A-1
A.i Details of Meta-Analysis A-1
A.ii Details of simulation validation A-2
A.iii Description of contemporary groups A-4
A.iv Quantitative trait loci A-5
Appendix B: Equine biobank documents B-1
B.i Project brief B-1
B.ii Online survey questions B-5
B.iii Online survey results B-9

Figures

Figure 1-1 A timeline of the main developments in the generation of an equine genetic map and genome sequence 2

Figure 1-2 Joints affected by osteochondrosis in the horse 12

Figure 2-1 Number of cases and controls by sex (*n*=348) 34

Figure 2-2 Distribution of cases and controls by farm of origin (*n*=348) 34

Figure 2-3 Percentage of case horses affected by OCD in the fetlock, hock, shoulder and stifle (*n*=169) 34

Figure 2-4 Number of joints affected by OCD in case horses (*n*=169) 34

Figure 2-5 Number of OCD lesions in case horses (*n*=169) 35

Figure 2-6 Percentage of case and control horses affected by non-OC conditions (*n*=348) 35

Figure 2-7 Non-OC conditions 36

Figure 3-1 Chromosome length and the number of SNPs per chromosome (average of UK and US) 46

Figure 3-2 Average LD measured by r^2, pooled over autosomes and plotted against the median of the distance bin range 48

Figure 3-3 Predicted r^2 versus observed r^2 against mean distance between markers (on a log scale) 50

Figure 3-4 Parameter estimates from the modelling of Equation 3-3 by chromosome, plotted against chromosome length 50

Figure 3-5 Average $\hat{N}_T$ plotted against generations in the past 51

Figure 3-6 Boxplot representing $\hat{N}_T$ plotted against generations in the past (on a non-linear scale), truncated at 1,000 generations 52

Figure 3-7 Average LD on ECA1 measured by r^2 and plotted against the median of the distance bin range 54

Figure 3-8 Average $\hat{N}_T$ plotted against generations in the past 55

Figure 4-1 Simulated data: Average $\hat{N}_T$ (truncated at N_T=500) plotted against average generations in the past 75

Figure 4-2 Simulated data: Boxplots of $\hat{N}_T$ plotted against generations in the past 76

Figure 4-3 Simulated data: Average $\hat{N}_T$ plotted against average generations in the past 78

Figure 4-4 Simulated data: Average $\hat{N}_T$ plotted against average generations in the past 78

Figure 4-5 Predicted and observed r^2 plotted against mean marker distance 79

Figure 4-6 Equine data: Average $\hat{N}_T$ plotted against average generations in the past 82

Figure 5-1 Sample configurations and marginal allele frequencies 94

Figure 5-2 Observed r^2 (pooled across 30 replicates) (O) and $E_c[r^2]$ (E) for marker pairs with different marginal allele frequencies 98

Figure 5-3 Histogram showing the distribution of N_e estimates by the CLE and SLD methods 101

Figure 5-4 Composite log likelihood curves for replicates 1 to 30 102

Figure 5-5 Estimates of N_e by CLE from marker pairs binned by distance 103

Figure 5-6 Mean and standard deviation of estimates of N_e by CLE from marker pairs binned by distance 104

Figure 5-7 Relationship between estimates of N_e by CLE and SLD (MAF threshold of 0.05) 106

Figure 5-8 Estimates of N_e by SLD from marker pairs binned by distance 107

Figure 5-9 Mean and standard deviation of estimates of N_e by SLD from marker pairs binned by distance 108

Figure 6-1 LD plot of ECA28: 0 – 1,342,640 with r^2x10^2 values shown 118

Figure 6-2 LD and synteny on ECA5 119

Figure 6-3 LD and synteny on ECA7 123

Figure 6-4 LD and synteny on ECA8 124

Figure 7-1 Distribution of genomic kinship between pairs of horses 137

Figure 7-2 Distribution of residuals from mixed model analysis 137

Figure 7-3 A Manhattan plot showing association results for ECA3 140

Figure 7-4 LD plot [203] of ECA3 region 1Mb either side of BIEC2-799865 141

Figure 8-1 The distribution of correlations between predicted and true phenotypes ($r_{y\hat{y}}$) following the permutation of GEBV amongst individuals 159

Figure 8-2 Correlation between corrected true phenotypes and genomic estimated breeding values ($r_{p\hat{p}}$) for BayesCπ and for GBLUP methods 159

Figure 9-1 Data flow for analysis 173

Figure 9-2 LD maps 176

Figure 9-3 LD maps 177

Figure 9-4 Mean proportion of correctly imputed genotypes by marker and its variance 181

Figure 9-5 Mean proportion of correctly imputed genotypes by marker and its variance 182

Figure 9-6 Proportion of correctly imputed genotypes plotted on the MAF of the SNPs (calculated in the reference panel) for ECA1 (bpEQ) 183

Figure 9-7 Proportion of correctly imputed genotypes and mean LD 184

Figure 9-8 Correlation between true and imputed genotypes plotted on: a) proportion of correctly imputed genotypes; b) proportion of correctly imputed genotypes, scaled by random expectation 187

Figure 9-9 The cost of genotyping relative to the maximum (high-density) and plotted against average accuracy 189

Tables

Table 1-1 Equine monogenic diseases with available genetic test 3

Table 1-2 Summary of pedigree based heritability studies of OC 15

Table 2-1 Quality control criteria implemented on genotype data and the number of SNPs discarded at each step 27

Table 2-2 Radiographic surveys 31

Table 2-3 A description of data used in subsequent chapters 37

Table 3-1 SNP exclusions made during quality control 40

Table 3-2 Distance classes and bin ranges for LD summary 41

Table 3-3 Chromosome specific megabase to centiMorgan conversion ratios 43

Table 3-4 Description of generation binning process 45

Table 3-5 Parameter estimates from non-linear regression modelling of ECA1 data 53

Table 3-6 Predictions of accuracy for genome-wide evaluation 60

Table 4-1 Description of formulae used in the estimation of variable effective population size 71

Table 4-2 Estimates resulting from the non-linear least squares modelling of the simulated dataset 73

Table 4-3 Estimates resulting from the non-linear least squares modelling of the simulated dataset 74

Table 4-4 Estimates resulting from the non-linear least squares modelling of the equine dataset. 81

Table 5-1 N_e estimates by CLE and SLD under various models 100

Table 6-1 MAF of SNPs identified as having r_g^2>0.25 with SNP(s) on different chromosomes 118

Table 6-2 SNPs for which the maximum recorded LD value (in at least one direction) was r_g^2>0.25 and with a SNP >10Mb away 121

Table 7-1 A description of conditions (other than OC) for which horses were treated 134

Table 7-2 Genotype frequencies of BIEC2-799865 and results of chi-square tests for association with OCD 140

Table 8-1 OCD: BayesCπ 158

Table 8-2 OCD: BayesC (π=0.10) 158

Table 8-3 Simulation: BayesCπ 161

Table 8-4 Simulation: BayesC (π=0.693) 161

Table 8-5 Simulation: GBLUP 162

Table 9-1 Mean (min., max.) proportion of correctly imputed genotypes by sample (ECA1) 180

Table 9-2 Mean correlation (min., max.) between true and predicted genotypes by sample (ECA1) 186

Table 9-3 Mean (standard error across SNPs/samples) proportion of correctly imputed genotypes on ECA1/ECA26 190

Table A-1 Comparison of observed and expected number of segregating sites A-2

Table A-2 Observed heterozygosity A-3

Table A-3 The distribution of samples across contemporary groups A-4

Table A-4 Details of QTL regions A-5

Preface

The work contained within this thesis is my own and has not been done in collaboration, except where otherwise stated. No part of this thesis has been submitted to any other university in application for a higher degree.

Laura J. Corbin

Acknowledgements

Firstly, I would like to express my heartfelt thanks to my principal supervisor, Professor John Woolliams, for his unwavering guidance throughout the last four years. Without his help and crucially, his confidence in me, I am sure I would not have achieved nearly as much during my PhD. I also thank my secondary supervisors, Professor Stephen Bishop for, amongst other things, his straight talking and down to earth attitude, and Doctor Sarah Blott for her help and advice on everything horsey. My thanks also go to my industrial supervisor, Jan Rogers, for her support throughout. I would like to thank all those who were involved in the collection of the EGR dataset, without which this PhD would not have been possible. In particular, I would like to acknowledge the work of Doctor Larry Bramlage and his assistant, Sheri Miller, in collecting the osteochondrosis samples and thank them for the warm welcome they gave me when I visited the Rood and Riddle Equine Hospital in December 2010. I also acknowledge financial support from the Biotechnology and Biological Science Research Council, the Bioscience KTN and the British Equestrian Federation.

The Roslin Institute has been a great place to start my research career and I would like to thank all my colleagues, especially those in the Genetics and Genomics Department, for their help and support over the last four years. In particular, I would like to thank: Ricardo Pong-Wong for his help with all things Linux and for writing so many useful programs; Gib Hemani and Joseph Powell for their guidance on programming in R; Bill Hill for helpful discussions about LD; and Hans Daetwyler for helping me to get started. I would also like to thank all my office pals for keeping me amused, especially David Telford, Ross Houston, Suzanne Rowe, Ricardo Pong-Wong, Vinca Russell and Debby Lipschutz-Powell.

Finally, thanks to my family, the Braddock's and the Corbin's. And thanks to my husband Adam, for following me to Edinburgh and for supporting me throughout. And last, but not least, thanks to the Australians (you know who you are) for all the banter!

List of Publications

REFEREED:

1. L. J. Corbin, S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop and J. A. Woolliams (2010) **Linkage disequilibrium and historical effective population size in the Thoroughbred horse.** *Animal Genetics,* **41**(s2):8-15. (Based on Ch. 3)
2. L. J. Corbin, S. C. Blott, J. E. Swinburne, C. Sibbons, L. Y. Fox-Clipsham, M. Helwegen, T. D. H. Parkin, J. R. Newton, L. R. Bramlage, C. W. McIlwraith, S. C. Bishop, J. A. Woolliams and M. Vaudin. (2011) **A genome-wide association study of osteochondritis dissecans in the Thoroughbred.** *Mammalian Genome,* **23**(3-4):294-303. (Based on Ch. 7)
3. L. J. Corbin, S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop and J. A. Woolliams (2012) **The identification of SNPs on the Equine SNP50 BeadChip with indeterminate positions.** *Animal Genetics,* **43**(3):337-339. (Based on Ch. 6)
4. L. J. Corbin, A. Y. H. Liu, S. C. Bishop and J. A. Woolliams (2012) **Estimation of historical effective population size using linkage disequilibria with marker data.** *Journal of Animal Breeding and Genetics,* **129**(4):257-270. (Based on Ch. 4)

CONFERENCE ABSTRACTS:

1. L. J. Corbin, S. C. Blott, M. Vaudin , J. E. Swinburne, C. W. McIlwraith, L. R. Bramlage, and J. A. Woolliams. (2010) **Genomic prediction of risk for osteochondrosis in the Thoroughbred horse.** *Plant and Animal Genome XVIII Conference*, San Diego, 9th-13th January 2010.
2. L. J. Corbin, S. C. Bishop, J. E. Swinburne, M. Vaudin, S. C. Blott and J. A. Woolliams. (2010) **Characterisation of linkage disequilibrium and subsequent estimation of effective population size in Thoroughbred horses using SNP markers.** *British Society of Animal Science Annual Conference*, Belfast, 11th-12th April 2010.
3. L. J. Corbin, S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop and J. A. Woolliams. (2010) **Can we accurately predict the effective population size of a Thoroughbred horse population using SNP50 marker data?** *32nd Annual International Society of Animal Genetics Conference*, Edinburgh, 26th-30th July 2010.
4. L. J. Corbin, S.C. Bishop, J. E. Swinburne, M. Vaudin, S.C. Blott and J. A. Woolliams. (2010) **The impact of method on the estimated effective population size of a Thoroughbred population using genotype data.** *9th World Congress on Genetics Applied to Livestock Production*, Leipzig, 1st-6th August 2010.
5. L. J. Corbin, S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop and J. A. Woolliams. (2011) **A haplotype library for the horse.** *1st BBSRC Institute's Conference*, 5th-7th July 2011.

Chapter 1: General introduction

1.1 From genetics to genomics in the horse

Genetics is defined as the branch of science that deals with heredity and the variation of inherited characteristics in living organisms [1]. Since the first description of the action of genes by Crick in 1958 [2], tremendous effort has been devoted to improving our understanding of the relationship between genotype and phenotype. Genetic mapping in the horse began in the 1970's with the discovery of the first genetically linked markers which consisted of protein polymorphisms and blood group systems [3-6]. Figure 1-1 provides a summary of the progress that has subsequently been made (for a more detailed review, see Chowdhary & Raudsepp (2008) [7]). As more markers were discovered, their potential utility in parentage testing was recognised and probability of exclusion statistics were commonly presented [8-10]. During the 1990s, microsatellites were added to parentage testing panels for horses [11, 12]. At the same time, reports began accumulating on the identification of genes or chromosomal regions that influenced traits of economic importance in a range of livestock species such as fecundity in sheep [13, 14] and double muscling in cattle [15], and equine geneticists realised the potential utility of a genetic map of the horse in tackling common diseases.

From this enthusiasm was borne the Horse Genome Project (University of Kentucky (2012). *Horse Genome Project*. [Online] Available from: http://www.uky.edu/Ag/Horsemap/abthgp.html), the initial goal of which was to identify landmarks on each chromosome that could be used to establish points of reference with the human genome sequence, which had already begun to be sequenced. The value of this comparative mapping approach was soon demonstrated, with comparative candidate gene studies leading to the development of DNA tests for several single-gene disorders affecting particular breeds of horse (see Table 1-1).

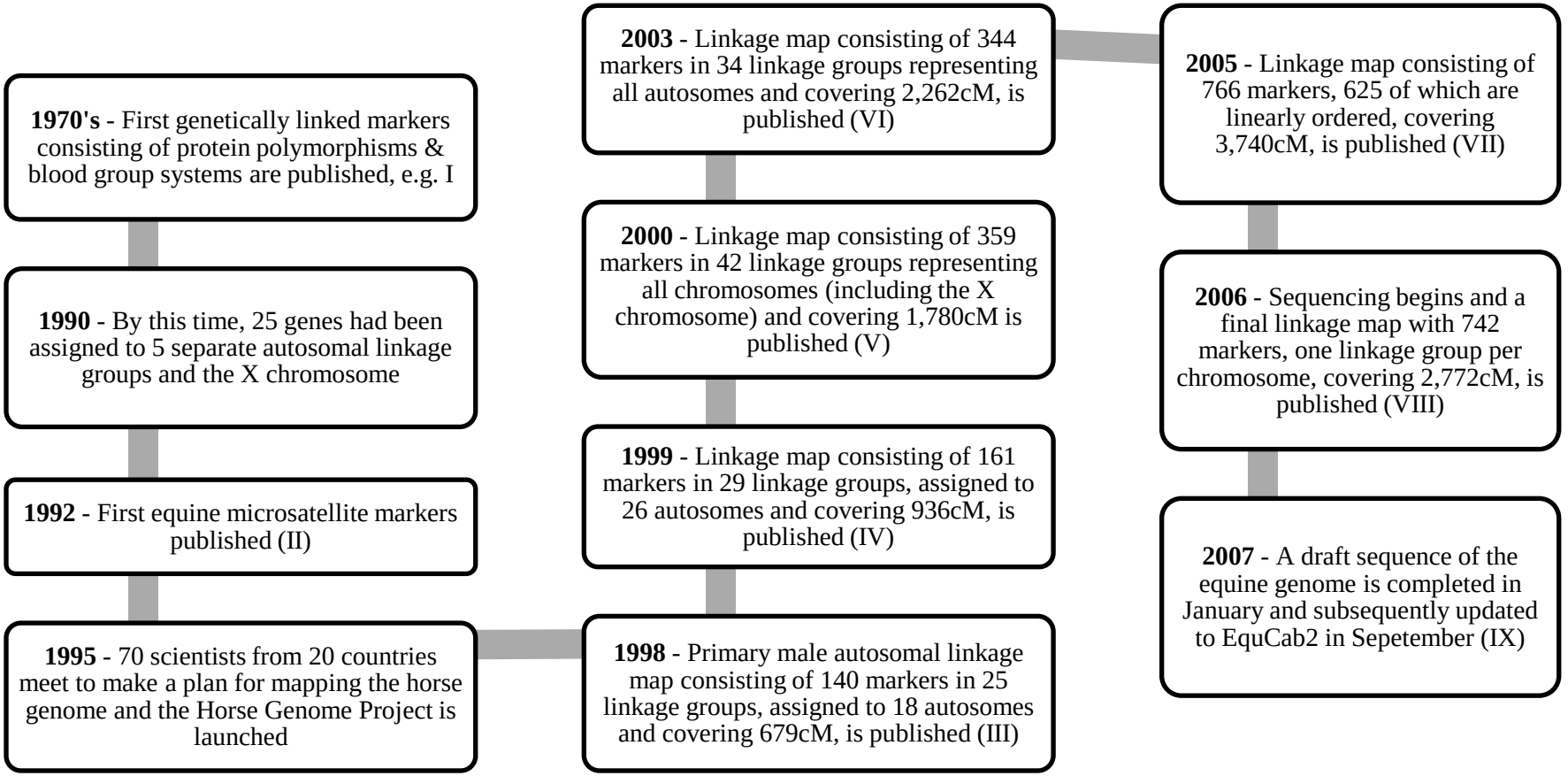

Figure 1-1 A timeline of the main developments in the generation of an equine genetic map and genome sequence. I = [3]; II = [16]; III = [17]; IV = [18]; V = [19]; VI = [20]; VII = [21]; VIII = [22]; IX = [23].

Table 1-1 Equine monogenic diseases with available genetic test

Disease	Breed affected	Mode of inheritance	Candidate gene	Summary	References
Hyperkalaemic Periodic Paralysis (HYPP)	Quarter Horse	Autosomal dominant	Adult skeletal muscle sodium channel α-subunit – responsible for HYPP in humans	Linkage analysis showed segregation of the gene with HYPP status. Subsequent sequencing of the gene revealed a Phe to Leu substitution in affected horses.	[24, 25]
Severe Combined Immune Deficiency (SCID)	Arab	Autosomal recessive	DNA-dependent kinase, catalytic subunit - responsible for SCID in mice	Sequencing of the gene revealed a frame-shift mutation in affected horses.	[26-28]
Overo Lethal White Syndrome (OLWS)	American Paint Horse	Autosomal recessive	Endothelin receptor B - responsible for similar phenotypes in mice and humans	Sequencing of the gene revealed an Ile to Lys substitution in affected horses.	[29]

Throughout the 1990's, linkage mapping played an increasingly important role in the generation of genetic maps for a range of species. Linkage groups identified through this approach could be assigned to chromosomes using molecular approaches such as Fluorescence In-situ Hybridisation (FISH), allowing the integration of genetical and physical data. In horses, this approach was first demonstrated in 1997 [30], but by this time, primary and in some cases second generation linkage maps were already available for a number of other domestic species. The slow progress in the horse has been attributed to the high costs and demands of breeding horses and to their somewhat unique position between companion animal and livestock [19, 22]. When combined with a long gestation period, monoparous births, and a lack of success of the super-ovulation techniques practiced in cattle, these factors make it particularly challenging to establish the large full sib families required for efficient linkage mapping [17]. Researchers were thus forced to rely on existing family material such as the large half sib families present for popular stallions. This approach required greater sample numbers and could only produce male linkage maps. Despite these challenges, the horse linkage map continued to be developed alongside synteny and radiation hybrid (RH) maps. A major breakthrough came in 2000 as a result of developments in reproductive technologies. Using embryo splitting, Swinburne *et al.* (2000) [19] produced two 3-generation full-sibling families, crossing four breeds to maximise heterozygosity, enabling the generation of a much more comprehensive linkage map which for the first time included linkage groups on the X chromosome. The final and most comprehensive linkage map for the horse was published in 2006 and consisted of 742 markers covering 2,772cM [22], this being remarkably close to an earlier estimate made from chiasma counts of 2,720cM [31].

By 2005, it had become clear that a revision of strategy was necessary. Research had shown that genes, whilst highly conserved across species, typically made up a very small proportion of mammalian genomes with only 1-2% of the human genome sequence encoding proteins [32]. It became apparent that the remaining 98% of the genome was also likely to be important in regulating gene function [32, 33]. Consequently, members of the Horse Genome Project developed a new goal of sequencing the entire horse genome, which given the timing, was very ambitious.

The consortium was successful in convincing the National Human Genome Research Institute (NHGRI) to prioritise the horse as one of 24 mammals to be sequenced as part of their efforts to identify the similarities and differences of the human genome compared with those of other mammals (University of Kentucky (2012). *Horse Genome Project*. [Online] Available from: http://www.uky.edu/Ag/Horsemap/abthgp.html).

The equine genome was sequenced by the Broad Institute at MIT and Harvard in conjunction with the Equine Genome Sequencing Consortium. Sequencing began in February 2006 and in January 2007 the first draft of the genome sequence was released. This sequence was subsequently updated in September of the same year (EquCab2) [23]. Initial sequencing was based on the DNA of a female Thoroughbred, Twilight, chosen for her low heterozygosity rate, which reduces the complexity of the task. A Whole Genome Shotgun (WGS) approach was used to produce 6.8X coverage of the genome, similar to that achieved for *Bos taurus*.

The equine genome comprises 31 pairs of autosomes and two sex chromosomes (2*n*=64), commonly denoted using the prefix ECA which stands for *Equus caballus*. Chromosomes one to 13 are metacentric[1] or submetacentric[2], with the remainder being acrocentric[3]. The average N50[4] contig size during assembly was 112.38kb, considerably longer than the N50 size reported for *B.taurus* (48.7kb), with the total

[1] Metacentric – having the centromere medially situated so that the two chromosomal arms are of roughly equal length (Merriam-Webster (2012). *Dictionary*. [Online] Available from: http://www.merriam-webster.com/dictionary).

[2] Submetacentric – having the centromere situated so that one chromosome arm is somewhat shorter than the other (Merriam-Webster (2012). *Dictionary*. [Online] Available from: http://www.merriam-webster.com/dictionary).

[3] Acrocentric – having the centromere situated so that one chromosomal arm is much shorter than the other (Merriam-Webster (2012). *Dictionary*. [Online] Available from: http://www.merriam-webster.com/dictionary).

[4] N50 - The N50 size is the median sequence length of the contigs.

length of the contigs reaching 2.43Gb. The N50 can be viewed as one measure of the quality of the draft. When the gaps between the contigs in scaffolds were included, the total length of the Thoroughbred mare's genome was 2.68Gb. Repetitive sequences, many equine-specific, make up 46% of the genome assembly [23]. The current gene set comprises 22,900 genes and 20,646 proteins (National Center for Biotechnology Information (2012). *Equus caballus (Horse)* [Online]. Available from: http://www.ncbi.nlm.nih.gov/genome?term=equus%20caballus). By 2008, approximately 92% of the sequence had been ordered and oriented on the chromosomes using linkage maps, RH maps and FISH mapping data [7]. The draft genome sequence is subject to continued efforts by the community to improve its veracity both in terms of the sequence assembly and with respect to the associated annotation. With this in mind, in Chapter 6, genotype data is used to identify possible sequence errors.

This process of map development has led equine geneticists into the era of genomics. Considerable progress has been made in mapping genomes over the last ten to twelve years, but huge challenges remain in understanding how differences in sequence relate to the variation that is observed both within and across different species. The focus is now on learning how to utilise what is known about the equine genome to tackle disease and, potentially, to improve performance.

1.2 Genomics, disease and the horse

Genomics is defined as the scientific study of nucleotide sequencing, gene mapping, and the analysis of genome [1] and new tools are constantly being developed and made available to advance this field. In the horse, approximately 1.5 million single-nucleotide polymorphisms (SNP) were identified during both the assembly of the genome and the generation of approximately 100,000 WGS reads from each of seven horse breeds, chosen to represent a mixture of ancient and more modern populations. These SNP markers were then used to produce a DNA microarray, the Illumina Equine SNP50 BeadChip ('50K SNP chip'), which contains over 50,000 SNPs and was released in 2007. This technology prompted and then enabled the current project, although the fact that this chip has since been superseded by a chip with 74,000 SNPs shows the speed of development in this field.

In the late 1990's, the potential for such dense SNP genotypes to be used in so-called genome-wide association studies (GWAS) to identify regions of the genome associated with quantitative traits, including disease, was recognised [34]. Today, GWAS are widely used to map susceptibility loci, so-called quantitative trait loci (QTL), for complex disease, typically using either a population or case-control study design. Population studies involve selecting individuals at random from the population and recording details of the trait of interest, the result being a sample in which disease occurs at the same rate as in the wider population. In contrast, in case-control studies, individuals are selected for the presence or absence of a particular condition. This enables the enrichment of the sample for cases which is particularly useful when studying rare diseases. A basic analysis then involves the comparison of allele (or genotype) frequencies between groups of affected (cases) and unaffected individuals (controls), under the assumption that the cases will have a higher prevalence of susceptibility alleles. Assuming that causal loci themselves are unlikely to be genotyped with currently-available SNP chips, GWAS rely on linkage disequilibrium (see below) between causal loci and markers to enable the identification of susceptibility through indirect association [35]. The framework for statistical inference of QTL mapping is challenging and new statistical approaches to infer such relationships are continually being developed, e.g. [36-38]. Assuming an associated region has been identified and validated, fine-mapping using additional markers and possibly across-population (or breed) analysis, is used to search for the causal variant. The identification of such causal variants can help to improve our understanding of the underlying biology, illuminating new biological pathways and in turn informing new treatments and management strategies. In the context of domestic animals, QTL can also be utilised in breeding strategies and this is referred to as marker assisted selection (MAS).

Whilst GWAS have led to the discovery of numerous common genetic variants, many QTL identified in preliminary analyses have not been validated in subsequent analyses [39] and much of the additive genetic variation remains unexplained [40-42]. This latter issue, coined 'the missing heritability' by Maher in 2008 [41], has been the feature of many articles since in which researchers attempt to 'find' or at

least to explain this phenomenon [43-45]. Other potential issues of the GWAS methodology which have been identified relate to sample properties, shortcomings in statistical inference, and general underlying assumptions of the method [39, 40, 42, 46-48]; these have been discussed in the literature at length and therefore will not be considered in any further detail here, although some of these issues are addressed in Chapters 7 and 8. The lack of QTL of large effect, combined with issues relating to implementation, has meant that whilst MAS has been beneficial in the case of monogenic diseases [49], there are only a small number of examples of its successful implementation with respect to more complex traits [50].

In recognition of the limited ability of single QTL to help address more complex production and disease traits, an alternative approach to marker based selection was suggested [51] and subsequently demonstrated [52]. Genomic or genome-wide evaluation (GWE) permits the estimation of the breeding value of animals using genotypic information without having to identify individual causal variants, thus avoiding the aforementioned issues of statistical inference. The method involves the simultaneous estimation of a large number of marker effects from a limited number of phenotypic records, with the expectation that most QTL will be in linkage disequilibrium (see below) with at least one marker. Genomic breeding values (GEBV) are generated by first collecting genotype and phenotype data for a so-called reference population and estimating allele effects for each marker or haplotype. Subsequently, breeding values for selection candidates can be estimated from genotypes alone by summing all the allele effects for the individual. A variety of methods have been proposed to estimate the allele effects, which poses a challenge due to the fact that there are usually considerably more allelic effects to be estimated than there are observations to estimate the effects from. The most commonly used methodologies to date are those of BLUP estimation (a ridge regression method), where all SNP effects are assumed to follow the same distribution, and Bayesian methods which make use of prior knowledge regarding the distribution of QTL effects [52]. Alternative non-parametric and semi-parametric methods including the use of machine learning algorithms have also been proposed [53-56]. The use of

GEBV to select animals for breeding is referred to as genomic selection (GS), and has already been applied within the dairy industry with encouraging results [57, 58].

Whilst first considered in the context of selection in animal breeding, human geneticists have relatively recently recognised the potential of the GWE methodology to help to explain the aforementioned 'missing heritability' of complex traits [59, 60]. Genome-wide evaluation can also be used to derive an individual's genetic predisposition to a particular disease [61]. In this context, the emphasis is on identifying high risk individuals (human or animal) so that their environment can be managed to reduce the likelihood of them developing the disease. Furthermore, whilst in animal breeding the focus has traditionally been on the additive genetic component and the narrow sense heritability, in the context of risk prediction there might also be value in estimating the non-additive components that contribute to the broad sense heritability, for example, dominance and epistasis. In this case, the nonparametric methods of Gianola *et al.* (2009) [62] might be beneficial. This ability to generate genetic risk predictions from birth offers great potential to reduce the incidence of common diseases that are known to also have a significant environmental component.

The success of these new SNP-based approaches in the horse will depend on how well the available markers capture the sequence variation present in the population. This in turn depends on the informativeness of the markers and on particular structural properties of the equine genome. Trait-specific properties such as the distribution of underlying gene effects, in particular the joint distribution of effect size and allele frequency, are also important although generally not known. Important SNP statistics when predicting likely success of genomic approaches include genotyping rate, the distribution of minor allele frequencies (MAF) and the extent of linkage disequilibrium. Linkage disequilibrium (LD) describes the tendency for particular alleles at neighbouring loci to be co-inherited or in a statistical sense, is the non-random association of alleles at two (or more) loci [35]. As previously mentioned, genomic methodologies such as GWAS and GWE are dependent on the LD that exists between causal variants and SNP markers; greater

LD between SNPs and causal variants increases the power to detect QTL in GWAS and increases the accuracy with which GEBV can be estimated in GWE.

Preliminary work undertaken as part of the equine sequencing project provided some insight into these areas. The characterisation of haplotype structure within and across breeds showed that major haplotypes were frequently shared among diverse populations [23], increasing the possibility of disease associations being replicated across breed groups. Within-breed LD was found to be moderate with the majority of breeds showing a similar level of LD [23]. These early findings allowed some comparisons to be made with other species, as well as enabling some power predictions to be made regarding the number of SNPs that would be needed for future gene mapping studies to be successful. However, this analysis was not comprehensive, involving only 24 samples per breed and SNP genotypes from ten 2Mb regions, and the results cannot be assumed to be representative of all equine populations.

The equine population investigated in this thesis is the Thoroughbred, an important breed both in the UK and across the world. As well as having a considerable impact on the equine industry through horseracing, the Thoroughbred has contributed genetically to a wide range of sport horse breeds and continues to be used both in its purebred form and as a crossbred in a wide range of disciplines. A comprehensive within-Thoroughbred analysis of marker informativeness and genome structure has not previously been done and is presented in Chapter 3. The extent of LD is at least partly dependent on past effective population size and this relationship is explored more fully in Chapters 3, 4 and 5.

At the outset of this PhD, several groups had published preliminary findings of GWAS performed using the 50K SNP chip. Two of these studies focused on diseases thought to be simply inherited and monogenic, Lavender Foal Syndrome in Arabian horses [63] and dwarfism in Miniature horses [64]. A further study addressed a condition with a previously unknown genetic background, Adolescent Idiopathic Lordosis in American Saddlebred Horses [65], which has since been hypothesised to be caused by a recessive gene [66]. Whilst success has already been

seen with respect to Lavender Foal Syndrome, the causal mutation having been identified [66, 67], fine-mapping and candidate gene studies are apparently on-going in the case of the other diseases. Somewhat in contrast to these early studies, this PhD considers how the 50K SNP chip can be used to investigate more common complex diseases, using osteochondrosis in the Thoroughbred as an example of such a condition. This disease is more typical of the multifactorial, lowly heritable conditions that GWAS and GWE have been developed to study.

1.3 Introduction to the disorder osteochondrosis

1.3.1 Clinical background

Osteochondrosis (OC) is a disease of the locomotory system which affects many animals, but is most frequently observed in pigs, horses and dogs [68]. In horses OC is classified as a developmental orthopaedic disease (DOD) and has been defined as a disturbance in the physiological process of endochondral ossification that occurs in young, growing individuals [69]. The normal process of endochondral ossification involves several stages – cartilage formation, hypertrophy, degradation and replacement by bone [70], a system which ensures all structures of the skeletal system remain functional during growth. In long bones, there are two regions at which endochondral ossification occurs, the growth plate (physis) and the articular-epiphyseal cartilage complex at the end of the bone. Osteochondrosis can affect both of these areas but in horses the term is most often used with respect to the articular form of the disease and this convention is upheld here. Articulations most commonly affected by OC are the joints of the fetlock, hock and stifle, but OC can also affect the shoulder and cervical spine [71] (see Figure 1-2). Specific so-called 'predilection sites' exist within each of these joints [72] and these sites are the focus of radiological examinations in clinical investigations of OC.

In horses with OC, irregular ossification leads to the formation of primary lesions of necrotic tissue within the growth cartilage [68, 70]. These lesions may then progress into the adjacent structures of the articular cartilage and/or subchondral bone [68, 70]. In cases where the lesion extends through the articular cartilage to form cartilage flaps or loose fragments at the joint surface, the condition is generally

referred to as osteochondrosis dissecans [68, 70] or osteochondritis dissecans [72], with these terms apparently being used interchangeably and abbreviated to OCD. This range in severity of disease leads to considerable variability in clinical presentation. Whilst the primary lesions may be subclinical in nature, as the disease progresses clinical signs include synovitis and pain, accompanied by varying degrees of lameness [73]. The management of clinical cases of OC depends on the site and severity of signs. Approaches range from conservative measures, such as restricted exercise, to surgical intervention in the form of arthroscopy to remove damaged cartilage [71].

Whilst the underlying cause of OC is not known, it is considered multifactorial in origin and factors which have been implicated in the horse include high growth rate, nutrition, endocrinological factors, biomechanical factors relating to both conformation and exercise, and heredity [71, 74-76]. However, there is much conflicting evidence regarding the influence of these factors and based on a review of available literature, it has been suggested that hereditary and anatomic factors are the most important contributors to the pathogenesis of OC [68].

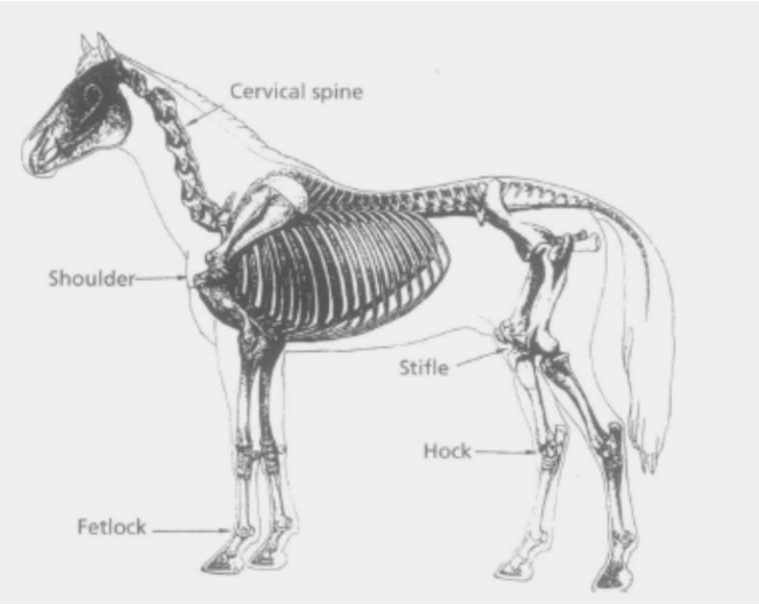

Figure 1-2 Joints affected by osteochondrosis in the horse, taken from Jeffcott (1997) [71]

1.3.2 Prevalence

There appears to be considerable variation in the prevalence of OC across breeds. Estimates vary from 16.6% in Italian Warmbloods (Maremmano) [77] to 70% in Dutch Warmbloods [69] (for a comprehensive list of prevalence estimates, see van Grevenhof *et al.* (2009) [69]). However, a large proportion of this variation may be attributed to differences in the methods used, including different disease definitions as discussed in detail below. Of particular relevance is the fact that OC can be subclinical and for a horse to be confidently designated as unaffected, radiographs must be taken. Studies based on clinical instances of OC alone will inevitably record lower prevalence rates than those in which all horses have undergone a radiographic survey. However, the radiographs are another source of between-study variability as the number and position of the radiographs are likely to influence the number and severity of lesions observed, with fewer examined sites potentially resulting in missed cases. Age at examination is also relevant because of the transient nature of the disease. Primary lesions often repair spontaneously in horses under the age of five and eight months in the case of the hock and stifle joints, respectively [78], and therefore a definitive diagnosis can only be made after this stage. Regardless of the exact prevalence, OC is recognised as a significant problem in horses, especially since it has been suggested that the incidence of OC in horses may be increasing [71, 79].

1.3.3 Genetic control

Observations that the progeny of certain sires suffered more frequently from the disease than those descended from other sires [80] provided the initial evidence of an additive genetic component to OC. Subsequent studies have attempted to determine the heritability of OC in a number of different populations. Heritability estimates vary widely, with this variation likely to be at least partly attributable to issues relating to phenotypic definition which are discussed in more detail below. Heritability estimates based on the binary classification and calculated using a linear model are expected to be lower than those calculated on the underlying scale, with the scaling factor dependent on disease prevalence, such that the rarer the disease, the greater the underestimation of heritability under the linear model. However, the

transformation of observed heritabilities according to Dempster & Lerner (1950) [81], produces estimates on the liability scale which are expected to be comparable to those calculated by fitting a threshold model, for example a logit model, in the first instance. Variability in heritability estimates is also expected because they are by definition both population and environment specific. Table 1-2 presents the results of a sample of heritability studies undertaken in a variety of breeds, along with a meta-analysis of these studies (see A.i for meta-analysis methodology).

Large standard errors are fairly characteristic of heritability estimates for OC. This is likely due both to the small sample sizes and the non-ideal family structures typical of such studies. Theory dictates that the optimum half-sib family size for estimating genetic parameters is 20-30 or, more precisely in cases where the heritability is known, $^4/_{h^2}$ [82]. Assuming heritability for OC of 0.10, this represents 40 offspring per sire, a figure that is not typically achieved in horse samples. A high variation in the number of offspring, with typical pedigrees containing a small number of sires with a relatively large number of offspring, and a large number of sires with a small number of offspring, is also expected to increase the variance of the estimate [83]. The results of simulation work done by Stock *et al.* (2007) [84, 85], designed to explore the power of typical horse samples, support these theoretical predictions. In this thesis, a genomic approach to heritability estimation is presented in Chapters 7 and 8.

Despite these issues, there is strong evidence to suggest a genetic component to OC and meta-analysis results in Table 1-2 indicate heritabilities which are significantly different from zero. Therefore, implementation of selection strategies should, in principle, help to reduce the incidence of OC in equine populations. Simulations based on a heritability for OC of 0.14, which appears reasonable given the meta-analysis results in Table 1-2, and five generations of active selection on males and females, demonstrated the potential to reduce the incidence of OCD in Maremmano horses from its current level (16.6%) to a level of around 1.8% [77].

Table 1-2 Summary of pedigree based heritability studies of OC. The result of a meta-analysis of the heritability estimates on an underlying scale and for each phenotype is presented.

Sample description	Method used	Heritability estimates (SE)	Reference
All joints OC/OCD	Meta-analysis	**0.21 (0.08)**	
n=350 Maremmano Italian WB	LAM	0.14 (0.23) [I]	[77]
n=167 SGC	LAM	0.17 (0.19) [I]	[86]
n=811 Dutch WB	LAM	0.23 (0.09) [II]	[87]
Fetlock joint OC/OCD	Meta-analysis	**0.12 (0.06)**	
n=167 SGC (OC)	LAM	0.16 (0.16) [I]	[86]
n=167 SGC (OCD)	LAM	0.08 (0.09) [I]	[86]
n=811 Dutch WB	LAM	0.14 (0.08) [II]	[87]
Hock joint OC/OCD	Meta-analysis	**0.20 (0.09)**	
n=325 Danish Trotters	SM	0.26 (0.14) [III]	[80]
n=793 Swedish SB Trotters	LSM	0.24 (0.19) [I]	[88]
n=167 SGC (OC)	LAM	0.04 (0.07) [I]	[86]
n=811 Dutch WB	LAM	0.36 (0.11) [II]	[87]
Stifle joint OC/OCD	Meta-analysis	**0.05 (0.05)**	
n=350 Maremmano Italian WB	LAM	0.09 (0.24) [IV]	[77]
n=811 Dutch WB	LAM	0.05 (0.05) [II]	[87]

SB = Standardbred; SGC = South German Coldblood; WB = Warmblood; OC = osteochondrosis; OCD = osteochondrosis dissecans; LAM = linear animal model; SM = sire model; LSM = linear sire model. [I]Heritability estimates using a linear model then transformed following Dempster & Lerner (1950) [81]; [II]Heritability estimates after prior transformation to a continuous liability scale; [III]Heritability estimates using a threshold model; [IV]Heritability estimate taken from Table 1 in van Grevenhof *et al.* (2009) [87].

1.3.4 The phenotype

A key area of concern when conducting genetic studies is the phenotypic definition used. Optimal phenotypic definitions are those with strict inclusion criteria, ideally in the form of a clinical diagnostic test. In the case of OC, the phenotypic definition is far from standardised, making results from both epidemiological and genetic studies difficult to interpret. There are a large number of radiographic findings that have been associated with OC over the last 20 to 30 years, some of which have subsequently been attributed to other causes such as trauma, for example, plantar osteochondral fragments (POFs) found in the fetlock. This has led to some confusion amongst practitioners and in some cases, whether a horse is diagnosed with OC or not will depend on the training, experience and perspective of the veterinarian. The exact location of the lesions and/or fragments within joints is generally considered to be an important factor in distinguishing between horses with OC and horses suffering from an injury of a traumatic nature. However, OC derived fragments are often studied alongside more generally described 'osteochondral fragments' which occur in regions not generally associated with OC, for example, the aforementioned POFs, and this has increased the confusion. For example, in a review of heritability estimates for OC/OCD, van Grevenhof *et al.* (2009) [87] present results from Stock *et al.* (2005) [89] and Stock *et al.* (2006) [90], neither of which restricted cases to OC. Rather these were studies of osseous fragments, not restricted by joint location or cause, and their interpretation as OC/OCD is despite Stock *et al.* (2005) [89] stating that 'osseous fragments as analysed in the present study cannot generally be attributed to the OC syndrome.' This confusion around disease diagnosis has led to authors defining their own, study-specific criteria for OC case inclusion. However, further issues surround the way in which OC cases are then considered in the analysis itself. Two principal areas of conflict between studies are: (i) the appropriateness of the binary classification for OC; and (ii) the treatment of OC in different joints as separate traits. These issues are discussed in more detail below.

The variability in the severity and manifestation of the disease from primary lesions that are barely visible, through loose bodies associated with OCD to subchondral

fractures, makes recording phenotypic descriptions within studies challenging, and across studies even more so. This has led to some debate as to whether OC should be described as a categorical trait that reflects severity or a simple binary classification based on presence or absence of disease. The system of grading cases according to severity has its origins in a clinical setting, where it can be used to help determine the best treatment option. However, its relative usefulness when compared with a binary classification in the genetic context remains unclear. Whilst it is assumed that there is an underlying disease liability, it is currently unknown what the relationship is between an individual's genetic liability for OC and the severity of the phenotype displayed. For example, does an increase in the number of susceptibility alleles in an individual cause lesions to become bigger or does it cause them to become more numerous? This ambiguity makes the alignment of severity and liability scales questionable, and undermines a standard assumption that a more quantitative trait is preferable. Furthermore, the severity of disease is likely to be more strongly confounded with environmental factors, such as the age of the horse at examination and its management up to that point. In a study of environmental risk factors of DOD (including OC), Lepeule *et al.* (2011) [91] demonstrated that risk factors are different for the presence and for the extent of disease. Whilst such factors may need to be present in models used to evaluate the genetic variance on any scale, the interpretation of a categorical scoring system may depend more heavily on the correct environmental model.

The empirical evidence in favour of using a categorical definition for OC in genetics studies remains limited. The recommendations of van Grevenhof *et al.* (2009) [87] that OC should be scored in more than two categories were based on results that showed higher heritability estimates could be obtained when using such an approach. These results are largely in line with genetics theory which predicts that the estimation of heritability using a logistic model and binary classification will be lower than that estimated using a threshold model and do not indicate any increase in genetic variance *per se*. However, the decrease in the standard error of the estimates based on the categorical definition demonstrates an increase in power which,

provided the aforementioned problem relating to the alignment of severity and liability scales could be overcome, would also be beneficial in QTL mapping studies.

The occurrence of OC in different joints and at different sites within joints further complicates the phenotypic definition. The degree of lameness can vary by joint affected and is therefore clinically relevant [92]. Literature from the 1990s, tends to refer to OC as a generalised condition but following a review of possible models of pathogenesis, Ytrehus *et al.* (2007) [68] concluded that the focal nature of OC lesions, which most likely originate due to ischemia, mean that studies should focus on site-specific traits. However, generalised factors were also acknowledged as being potentially influential, with the lesions then being secondary to the generalised condition. In a diagram of causal factors of OC, Ytrehus *et al.* 2007 [68] acknowledged a role for heredity only at the site-specific level, for example, relating to an anatomical predisposition to the disease. However, it seems equally plausible that genetic factors could also be involved at the generalised level. For example, if heritable generalised factors predispose a horse to OC, then further heritable anatomical characteristics could determine in which joint lesions will appear. The heritability estimates in Table 1-2, which indicate a genetic component to OC both at the generalised level and at the joint level, appear to lend credence to this argument. Therefore, there is no biological consensus on whether cases should be divided by joint affected in genetic studies.

Several studies have attempted to address the question of whether OC in different joints should be treated as different traits by calculating genetic correlations of OC in different joints [79, 86, 87, 93]. However, at least one of these studies, a study of South German Coldblood horses by Wittwer *et al.* (2007) [86], had an insufficient number of samples to allow any valid conclusions to be made. As predicted by theory [82, 83, 94], the dataset of 167 horses from 30 half-sib families, giving a mean number of offspring of 5.6 per sire (range 1 to 28), resulted in such large standard errors that it was impossible to exclude genetic correlations of 0 or 1 or -1. Furthermore, theory and simulation work by Stock *et al.* (2007) [84, 85] suggest the sample sizes of <500 horses used by Lykkjen *et al.* (2012) [79], may also be underpowered to detect significant genetic correlations. Whilst van Grevenhof *et al.*

(2009) [87] had only a moderate sample of 811 Dutch Warmbloods from 32 sires, with a greater mean number of offspring per sire (25.3) it benefitted from a much lower variance in family size (22 to 28 offspring per sire). Despite this more powerful study design, the standard errors of the genetic correlations were still relatively large, such that a correlation of zero could not be excluded in two out of three cases. Similar studies in pigs also appear to show conflicting results, with both positive and zero correlations having been recorded between OC in different joint locations, despite sample sizes being typically much larger. For example, in a study of around four thousand pigs from two different breeds Jørgensen & Andersen (2000) [95] recorded both positive and negative correlations between OC in a range of joints and there was little correspondence in estimates between the two breeds.

Interestingly, both Jørgensen & Andersen (2000) [95] and van Grevenhof *et al.* (2009) [87] conclude from their results that OC should not be considered a generalised disease and that OC in different joints should be considered as different traits. Indeed, the combination of low and non-significant correlations observed to date has led to this being the general consensus of researchers at the current time. However, it would seem that the tendency of these authors and others to err on the side of caution and make recommendations such that OC in different joints should be considered as different traits, is due to the way in which the hypotheses tests are formulated. By setting the null hypothesis (H_0) equal to there being no correlation between traits, if a significant correlation is not found the conclusion must be that the traits are separate. However, if what we are really looking for is evidence that OC in different joints is *not* the same disease, this hypothesis seems incorrect. Combining the three genetic correlations from Table 3 in van Wittwer *et al.* (2007) [86] and the two from Table 4 in van Grevenhof *et al.* (2009) [87] by meta-analysis (see A.i), gives an estimated genetic correlation of 0.32 (0.20). Although the validity of this approach (combining correlation estimates across traits) is questionable, there does not seem to be sufficient evidence to reject the idea that there is shared genetic variance across the disease in different joints. Therefore, in my opinion, the question of whether or not OC in different joints represents genetically distinct diseases remains unanswered.

In conclusion, there remains a great deal of uncertainty surrounding the issue of optimal phenotypic definition. Unfortunately, the issues raised above in the context of prevalence and heritability studies are now impacting on genomic analyses.

1.3.5 Quantitative trait loci mapping

The search for QTL associated with OC in horses began with family-based linkage analyses using genome-wide microsatellite markers. The first such study was carried out on a sample of 123 Hanoverian Warmbloods (HWB) in 2003 [96]. Two further studies involving a larger HWB cohort and a sample of South German Coldblood (SGC) horses were conducted in 2007 [97, 98]. Following the release of the 50K SNP chip, GWAS have been performed in three different horse breeds to date: HWB [99, 100], French Trotters [101] and Norwegian Standardbreds [102]. These studies have led to a large number of putative QTL being identified although no validated markers for OC have yet been published. In many cases, authors have gone on to discuss the relative merit of nearby genes as candidate genes for OC. However, the point made by Cardon & Bell (2001) [103] that, due to a lack of understanding of the mechanisms of action of complex trait loci, a plausible argument can be made for most associated alleles, most likely still holds true. Furthermore, the results of these mapping studies are once again strongly influenced by study design and phenotypic definition.

Presumably due to the uncertainty surrounding the optimal definition for OC, all but one of the aforementioned QTL mapping studies have analysed multiple disease phenotypes. Studies conducted on the HWB and SGC cohorts involved the analysis of up to six phenotypes, representing two different manifestations – OC and OCD, and three different locations – hock, fetlock and hock and fetlock combined. A similar approach was taken by Teyssèdre *et al.* (2011) [101] who analysed hock, fetlock and other cases separately and as a combined trait which also accounted for the number and severity of the lesions. This repeated testing using different clinical phenotypes was recognised by Cardon & Bell (2001) [103] as a common error in association studies because it leads to subgroups that are small, providing less robust results and creating a substantial risk of false positives. This could well explain the variable correspondence between QTL locations for the different phenotypes tested

in these studies. An alternative approach to this retrospective subdivision of cases is to collect more uniform cases in the first instance. For example, Lykkjen *et al.* (2010) [102] preferentially chose cases with OCD just at the intermediate ridge of the distal tibia (part of the hock joint). When compared to the testing of multiple phenotypes within a dataset, this approach is instinctively more statistically robust, leaving readers with little doubt as to the researcher's intentions at the outset of the study.

It is also common for researchers to use multiple significance thresholds and several different statistical approaches to analyse the data. For example, Lykkjen *et al.* (2010) [102] presented results for a chi-square-test, a logistic regression, a Cochran-Mantel-Haenszel (CMH) test and a mixed-model analysis on the same dataset. Whilst these kinds of approaches might be considered thorough, as in the case of testing multiple phenotypes, they also cast doubt on the statistical validity of what has been done and make interpretation of the results more challenging. Whilst these issues are not specific to studies of OC in the horse, they do seem to be exacerbated by small sample sizes, the tendency for related horses to be used, and by the aforementioned issues surrounding phenotypic definition. In Chapter 7, whilst mindful of these issues, a GWAS for OCD is performed.

The large numbers of different results that are published as a result of these uncertainties make comparing results across studies extremely challenging. Three instances of potential QTL correspondence have so far been reported by the authors. These are for:

- A QTL for hock OCD observed by Lykkjen *et al.* (2010) [102] at 77.4Mb on ECA5 and corresponding QTL for fetlock OCD observed by Lampe *et al.* (2009) [104] at 78.0-90.2Mb on ECA5.
- A QTL for hock OC observed by Teyssèdre *et al.* (2011) [101] at 100.4-107.9Mb on ECA3 and corresponding QTL for hock OCD observed by Lykkjen *et al.* (2010) [102] at 113.5Mb on ECA3.

- A QTL for fetlock OC observed by Teyssèdre *et al.* (2011) [101] at 6.9-12.9Mb on ECA13 and corresponding QTL for fetlock OCD observed by Lampe (2009) [99] at 15.3Mb on ECA13.

Whether these are viewed as true validations depends not only on an apparently subjective interpretation of statistical significance and phenotype equivalence, but also on the maximum distance at which QTL are considered to overlap. A more direct assessment of QTL correspondence is conducted in Chapter 7.

In summary, several genetic analyses and QTL mapping studies for OC have been published. However, due to differences in populations, trait definitions and methodologies, it is very difficult to get an overview of the results and make general conclusions. The evidence for a genetic component of OC is compelling, providing justification for conducting genomic analyses, but how a genetic susceptibility for OC translates into the phenotype we observe is not clear. The lack of a standardised phenotype for OC, combined with the wider issues associated with the implementation of QTL mapping studies, has resulted in a largely incoherent body of literature on the topic of the genetic background of OC.

1.4 Thesis outline

The main objective of this thesis is to investigate the potential for new genotyping technologies to enhance our understanding of the genetic component of complex diseases in the horse. As well as exploring the properties of the 50K SNP chip with respect to the Thoroughbred horse genome, osteochondrosis is used as an exemplar for the application of the 50K SNP chip to a complex disease in the horse.

Chapter 2 describes the datasets used throughout this thesis.

Chapter 3 summarises the extent and decline of LD in two Thoroughbred populations, considering the results in the context of the genome coverage provided by the Illumina Equine SNP50 BeadChip. The theoretical relationship between LD and N_e is used to predict the N_e of the two Thoroughbred populations when N_e is assumed to be constant *a priori* and when it is not. The results are considered in the context of the efficacy of genomic methodologies such as GWAS and GWE.

Chapter 4 evaluates the reliability of the methods used in Chapter 3 to predict N_e from molecular data. A thorough review of the theory that describes the relationship between the rate of decline of LD, the distance between markers and the N_e is presented. The impact of several developments on the estimation of N_e using both simulated and equine SNP data, when N_e is assumed to be constant *a priori* and when it is not, are considered.

Chapter 5 considers an alternative approach to that used in Chapters 3 and 4 for predicting N_e of a population using marker data. Preliminary findings only are presented since this approach was not progressed.

Chapter 6 uses LD to identify SNP on the Illumina Equine SNP50 BeadChip which may be incorrectly positioned on the genome map. Linkage disequilibrium was evaluated in a pair-wise fashion between all autosomal SNPs, both within and across chromosomes. Filters were then applied to the data firstly, to identify SNPs which may have been mapped to the wrong chromosome and secondly, to identify SNPs which may have been incorrectly positioned within chromosomes. The likely impact of the findings on a range of commonly used genomic approaches is discussed.

Chapter 7 details a GWAS for QTL associated with OCD in the Thoroughbred. A secondary objective was to test the effect of previously identified QTL in the current population. Genotype data for 348 horses which had been classified as cases or controls according to clinical findings was used. The effects of 24 SNPs, representing QTL previously identified in a sample of Hanoverian Warmblood horses, were tested directly using the data.

Chapter 8 evaluates the potential utility of GWE in tackling common complex diseases in the horse. Genotype data is analysed using both true OCD phenotypes and using a simulated binary trait intended to resemble OCD. Two methodologies are used to predict GEBV and the predictive ability of the models assessed.

Chapter 9 evaluates the potential for low-density genotyping panels to be used to reduce genotyping costs in the future. Three different approaches are used to select SNPs for panels of varying density and their efficacy judged according to the

accuracy with which missing genotypes could be imputed. The results are considered with respect to both the cost of each of the panels and the intended use of the imputed genotypes.

Chapter 10 features a final summary of the research undertaken in this thesis. A critical review of the results is conducted and the limitations of what has been done are considered. Considerations regarding the future for genomics research and its application to the horse are then presented.

Chapter 2: Description of dataset

2.1 Data source

Genotypic and phenotypic data were provided by the Animal Health Trust (AHT) as part of the Equine Genetics Research (EGR) programme, a joint initiative between the AHT and the British Horseracing Board (now the British Horseracing Authority). The aim of this wider project was to identify genomic regions contributing to complex musculo-skeletal disease in the Thoroughbred, specifically focusing on osteochondrosis (OC), recurrent exertional rhabdomyolysis and the risk of fracture as phenotypes. Samples for the project were collected from affected and unaffected Thoroughbred horses from the United Kingdom (UK) and in the case of OC, additionally from the United States (US). Collections took place over several years from 2006 onwards and involved a number of people[5]. Samples collected in the UK came from horses from both National Hunt and flat racing stables in the UK. As such, these samples are expected to comprise primarily of relatively unrelated horses originating from a wide geographical area. Samples collected in the US came from flat racehorses admitted to the Rood and Riddle Equine Hospital (RREH) (further details below). Genotyping was funded by the Horserace Betting Levy Board (HBLB) and the Thoroughbred Breeders' Association.

2.2 Genotyping

Blood samples were collected in ethylene-diamine-tetra-acetic (EDTA), and deoxyribonucleic acid (DNA) 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 70ng/ul and submitted for genotyping to Cambridge Genomic Services

[5] Sarah C. Blott (SCB), June E. Swinburne (JES), Charlene Sibbons (CS), Laura Y. Fox-Clipsham (LYFC), Maud Helwegen (MH), J. Richard Newton (JRN) and Mark Vaudin (MV) of the AHT, Tim D. H. Parkin (TDHP) of the University of Glasgow, Lawrence R. Bramlage (LRB) of RREH, and C. Wayne McIlwraith (CWM) of Colorado State University.

(http://www.cgs.path.cam.ac.uk/services/snp-genotyping/services.html). The Illumina Equine SNP50 Genotyping BeadChip (www.illumina.com/documents/products/datasheets/datasheet_equine_snp50.pdf) was used to genotype all samples. This BeadChip comprises 54,602 single nucleotide polymorphisms (SNP) located across all autosomes and the X chromosome. These were selected from the database of over one million SNP (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 at the AHT 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=3,895) of SNPs from the analysis due to poor genotyping quality and these SNPs were subsequently set to missing in all samples (see Table 2-1).

Table 2-1 Quality control criteria implemented on genotype data and the number of SNPs discarded at each step

Metric Examined	Purpose of metric applied	Threshold applied	No. of SNP removed at this stage
Cluster separation[I]	Removal of SNPs with inadequately defined clusters	<0.25 discarded	1319
Call frequency	Removal of SNPs which were not called in a minimum number of samples	<98% call rate discarded	2380
AB R Mean[II]	Removal of SNPs with low intensity data	<0.15 discarded	7
AB T Mean[III]	Removal of SNPs where the heterozygote cluster is not well separated from the homozygote clusters	<0.2 and >0.8 discarded	20
Heterozygote excess	Removal of SNPs which deviate significantly from Hardy-Weinberg equilibrium	Heterozygosity of <-0.3 and >0.1 discarded	149
ECAX markers	Removal of X chromosome markers where the males have been called heterozygous	Each marker assessed individually	20

[I]Measures the separation between the three genotype clusters and varies from 0 to 1

[II]Mean normalized intensity of the heterozygote cluster, with values increasing from 0

[III]Mean of the normalized theta values of the heterozygote cluster, with values ranging from 0-1. Values of <0.2 and >0.8 indicate the heterozygote cluster has shifted toward the homozygotes.

2.3 Preliminary data cleansing

The output from the initial screening carried out by the AHT and described above provided the starting dataset for this PhD and consisted of 1,235 horses genotyped for up to 50,707 SNPs. Samples had been anonymised, each being given a unique individual identification number (IID) and no pedigree data was available. Of the 1,235 samples, 399 had associated phenotypic data which reflected the horse's osteochondrosis (OC) status, either case or control, which were mostly collected from the US and will be referred to as the OC dataset. The remainder, having been collected for the other disease studies to be carried out at the AHT, had no associated OC status, were collected from the UK, and will be referred to as the population control dataset. A limited number of individuals had also been assigned a sex.

The preliminary exploration of the dataset involved the investigation of SNP quality through calculation of: genotyping rate, defined as, the number of missing genotypes, both by SNP and by sample; Hardy-Weinberg equilibria; and the distribution of minor allele frequencies (MAF). In addition, genotype data was used to estimate genome-wide identity by descent (IBD) sharing coefficients between all pairs of samples. This was done using a technique within Plink whereby the prior probability of identity by state (IBS) sharing is first expressed in terms of the allele frequency and then a global IBD estimate is calculated for that pair by the method of moments [105, 106]. This analysis revealed 17 pairs of individuals who had IBD coefficients of one, indicating that they had all genotypes in common. This result is indicative of sample duplication, such that 17 horses had been sampled twice and given two different IIDs. In some cases, the phenotype for the two samples making up the pair differed and, therefore, a conservative approach was taken and all 34 samples involved were excluded from future analyses. Consequently, 15 individuals were removed from the OC dataset and 19 from the population control dataset, leaving 384 samples (190 cases and 194 controls) in the former and 817 in the latter. Of the remaining samples, 417 had sex assigned. Sex determination, based upon detecting heterozygosity of X-chromosome SNPs [105, 106], revealed 29 instances of mismatched sex. Seventeen of these represented instances whereby homozygosity levels (F) fell in the range between the default thresholds used to define males and

females (0.2<F<0.8). In the remainder (*n*=12), the sex associated with the sample in the data file contrasted with that indicated by the X chromosome SNP data; these were all samples from the population control dataset. Since these samples did not have associated phenotypic data, there would be no impact of incorrect sex assignment or sample misidentification on the results of the intended analyses and therefore these samples were not excluded. The previous investigation of SNP quality described above was then repeated on this reduced dataset. Whilst some preliminary analyses of the OC dataset were carried out at this stage, further phenotypic information was desirable prior to carrying out the final analyses.

2.4 Phenotypic enrichment of the osteochondrosis dataset

The OC dataset consisted primarily of samples collected by LRB at RREH in Lexington, Kentucky, in the first quarters of 2007 and 2008. Initially, a simple case/control categorisation was provided for samples in the dataset which did not allow consideration of any additional phenotypic criteria such as joint affected or severity (see 1.3.4 for further details). Furthermore, whilst it was known that the horses sampled for the OC study originated from several different horse farms in the vicinity of RREH, the initial dataset did not include the farm of origin. Farm of origin was relevant since environmental factors would most likely vary by farm.

To help address the data gaps, more data was provided from RREH and AHT with additional details about the samples in the OC dataset. The following details were provided: country of origin – indicating that 36 of the OC dataset samples were collected from Thoroughbreds in the UK with the remainder having come from RREH; phenotype details – including, for the US samples, a list of conditions each horse was treated for and the joint affected; sex; farm – for the US samples, farm of origin was indicated using an anonymous coding system.

The new information highlighted two potential areas of concern. Firstly, there were ambiguities in the OC status of samples between the original dataset and the updated version; 85 horses had opposing OC status across the two datasets. This finding suggested some mis-labelling of samples which needed to be resolved before any

further phenotypic analysis could be carried out. Secondly, it became clear from the phenotypic details provided that both cases and controls suffered from a range of orthopaedic conditions other than OC (see below for further details) which would need to be considered in any future analysis of OC using this dataset. These issues could not be simply resolved and therefore, a visit was made to RREH in order to examine the raw data collection forms for the 348 samples that originated from the US.

Discussions with LRB established that OC case samples consisted of horses, aged between nine and twelve months, which were diagnosed as having OC in at least one joint from radiographic surveys performed by referring veterinarians (see Table 2-2). Such surveys of youngstock are often carried out in the first few months of the year in preparation for the yearling sales. The diagnosis was then confirmed through repeat radiography of suspected OC affected regions on the admission of the horses to RREH and where necessary, arthroscopic surgery was performed by LRB. A typical arthroscopic surgery involves the removal of all fragments and the debridement of any separated articular cartilage and defective bone from the joint [107]. Therefore, to qualify for surgery and subsequent sampling, cartilage and/or bone fragments separated from the articular surface would have to be present. Consequently, according to the definition of Ytrehus *et al.* (2007) [68], the cases can be considered as suffering specifically from osteochondritis dissecans (OCD) (see 1.3.1 for further details) and represent a subset of OC affected horses suffering from a specific manifestation of the disease. This reduction in the phenotypic heterogeneity addresses some of the concerns discussed in 1.3.4 regarding phenotypic definition. Further details that were also recorded include: sex of the horse; the code for farm of origin; the number of OCD lesions observed; the specific location of the OCD lesions observed; the approximate size of the OCD lesions observed (where recorded); and details of the non-OC conditions for which horses were operated on. Of the 348 US horses, 169 had OCD (cases) and 179 were negative for OCD (controls).

Table 2-2 Radiographic surveys: 32 radiograph views as recommended by Keeneland Thoroughbred Racing and Sales, Lexington, and based on guidelines provided by the American Association of Equine Practitioners (AAEP). Description taken from Preston *et al.* (2010) [108].

Right and left metacarpophalangeal (fetlock) joints
Dorsoproximal-palmarodistal oblique (15° proximal to the supporting surface)
Dorsomedial-palmarolateral oblique (30° medial to the dorsopalmar line)
Dorsolateral-palmaromedial oblique (30° lateral to the dorsopalmar line)
Lateromedial (obtained with the joint flexed)
Right and left metatarsophalangeal (fetlock) joints
Dorsoproximal-plantarodistal oblique (15° proximal to the supporting surface)
Dorsoproximomedial-plantarodistolateral oblique (15° proximal to the supporting surface and 30° medial to the dorsoplantar line)
Dorsoproximolateral-plantarodistomedial oblique (15° proximal to the supporting surface and 30° lateral to the dorsoplantar line)
Lateromedial (obtained with the horse standing)
Right and left carpal joints
Dorsolateral-palmaromedial oblique (30° lateral to the dorsopalmar line)
Dorsolateral-plantaromedial oblique (30° medial to the dorsopalmar line)
Lateromedial (obtained with the joint flexed)
Right and left tarsal (hock) joints
Dorsomedial-plantarolateral oblique (65° medial to the dorsoplantar line)*
Dorsolateral-plantaromedial oblique (10° lateral to the dorsoplantar line)
Lateromedial
Right and left stifle joints
Lateromedial
Caudolateral-craniomedial oblique (20° lateral to the craniocaudal line; must include medial femoral condyle in its entirety)
*Alternatively, the plantarolateral-dorsomedial oblique view (25° lateral to the dorsoplantar line) may be used.

2.5 Phenotypic Summary

Using the newly enriched OC dataset, a more detailed phenotypic summary based on the US samples in this set was carried out. There was a slight difference in the proportion of cases and controls per sex (Figure 2-1). Furthermore, having sex assigned to all samples meant that sex checking against X chromosome genotype data could be used during subsequent QC. Horses originated from one of 19 surrounding horse farms. The number of horses per farm ranged from two to 89, with approximately equal numbers of cases and controls sourced from each farm (see Figure 2-2). Although the anonymity of samples prevented pedigree details from being known, it is hypothesised that the dataset consists of a mixture of half-sibs and less related horses. This hypothesis is based on the results of molecular analyses whereby relationships between pairs of individuals were predicted based on genomic data. The common practice in horse breeding of using a single stallion on several mares both within and across farms, and of using the same mares within farms to produce foals in subsequent years, adds further credence to this hypothesis.

Cases could now be categorised according to the joint(s) affected by OCD (see Figure 2-3). The most commonly affected joint was the hock, with over half of cases suffering from this particular form of OCD. The next most commonly affected joint was the stifle, followed closely by the fetlock. Only one horse suffered from OCD in the shoulder. The severity of the disease, as demonstrated by the number and size of lesions present, could now be assessed (Figure 2-4 and Figure 2-5), enabling further characterisation of cases and comparisons with other studies. However, for the reasons outlined in 1.3.4, notably the ambiguity surrounding the alignment of severity and liability scales, this information was not used directly in subsequent analyses.

Horses were categorised according to the non-OC conditions which they were treated for (Figure 2-6). The most commonly occurring conditions were angular limb deformities (ALD), osteochondral fractures of the proximal (first) phalanx in the fetlock joint (fetlock chips) and sesamoid fractures (see Figure 2-7). Angular limb deformities (ALD) involve deviations in a limb when viewed from the front or back, such that deviation is excessive from side to side [109]. The condition is the result of

one side of the physis growing faster than the other [110] and can be treated with transphyseal bridging to stop growth on one side of the plate [109]. Osteochondral fractures of the proximal (first) phalanx in the fetlock joint (fetlock chips) although previously associated with OC have since been shown not to be part of the OC complex [111, 112] (L. R. Bramlage 2010, Pers. Comm.), but instead to have traumatic origins [107, 113]. Concussion and overextension of the joint have been identified as the most likely reasons for bone fracture [114]. Sesamoid fractures are also believed to have traumatic origins with excessive fetlock extension due to muscle fatigue resulting in the overloading of the sesamoid bones at the back of the fetlock which then fail [114, 115]. The uneven distribution of these conditions between cases and controls (see Figure 2-6) means that the presence/absence of at least the three most common non-OC conditions should be accounted for in future analyses.

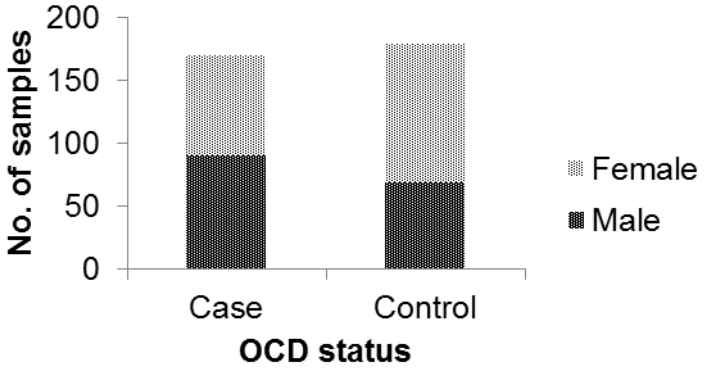

Figure 2-1 Number of cases and controls by sex (*n*=348)

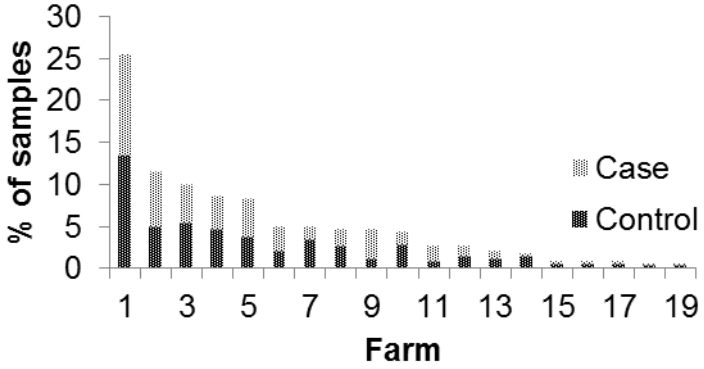

Figure 2-2 Distribution of cases and controls by farm of origin (*n*=348)

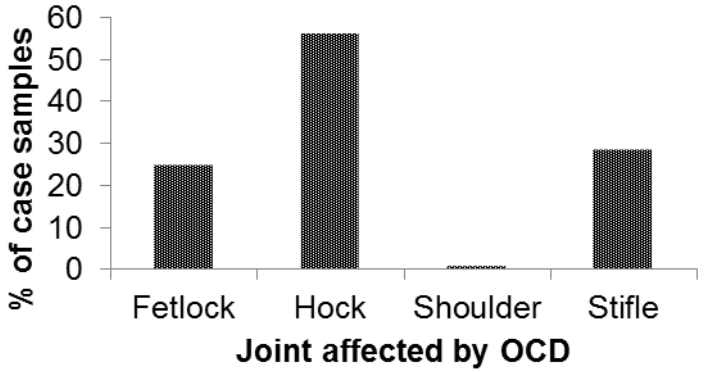

Figure 2-3 Percentage of case horses affected by OCD in the fetlock, hock, shoulder and stifle (*n*=169)

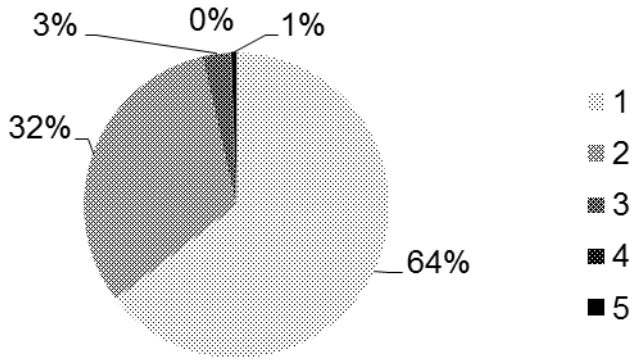

Figure 2-4 Number of joints affected by OCD in case horses (*n*=169)

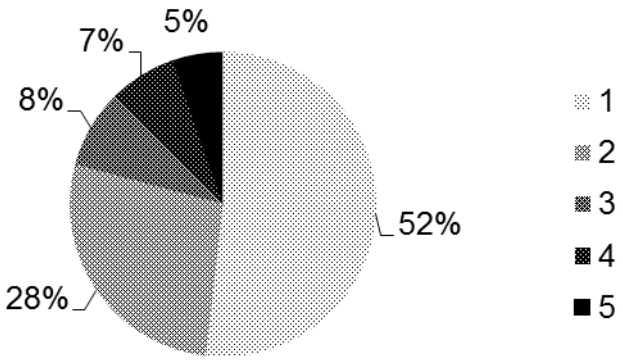

Figure 2-5 Number of OCD lesions in case horses (*n*=169)

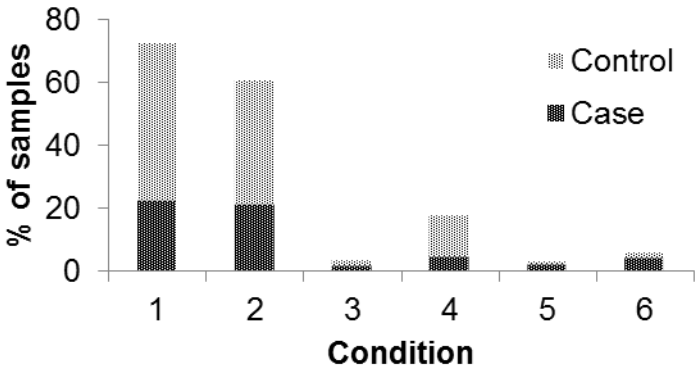

Figure 2-6 Percentage of case and control horses affected by non-OC conditions (*n*=348). Condition: 1 – angular limb deformity (ALD); 2 – fetlock chip(s); 3 – other chip(s); 4 – sesamoid fracture; 5 – other bone related conditions; 6 – other (non-bone related) conditions.

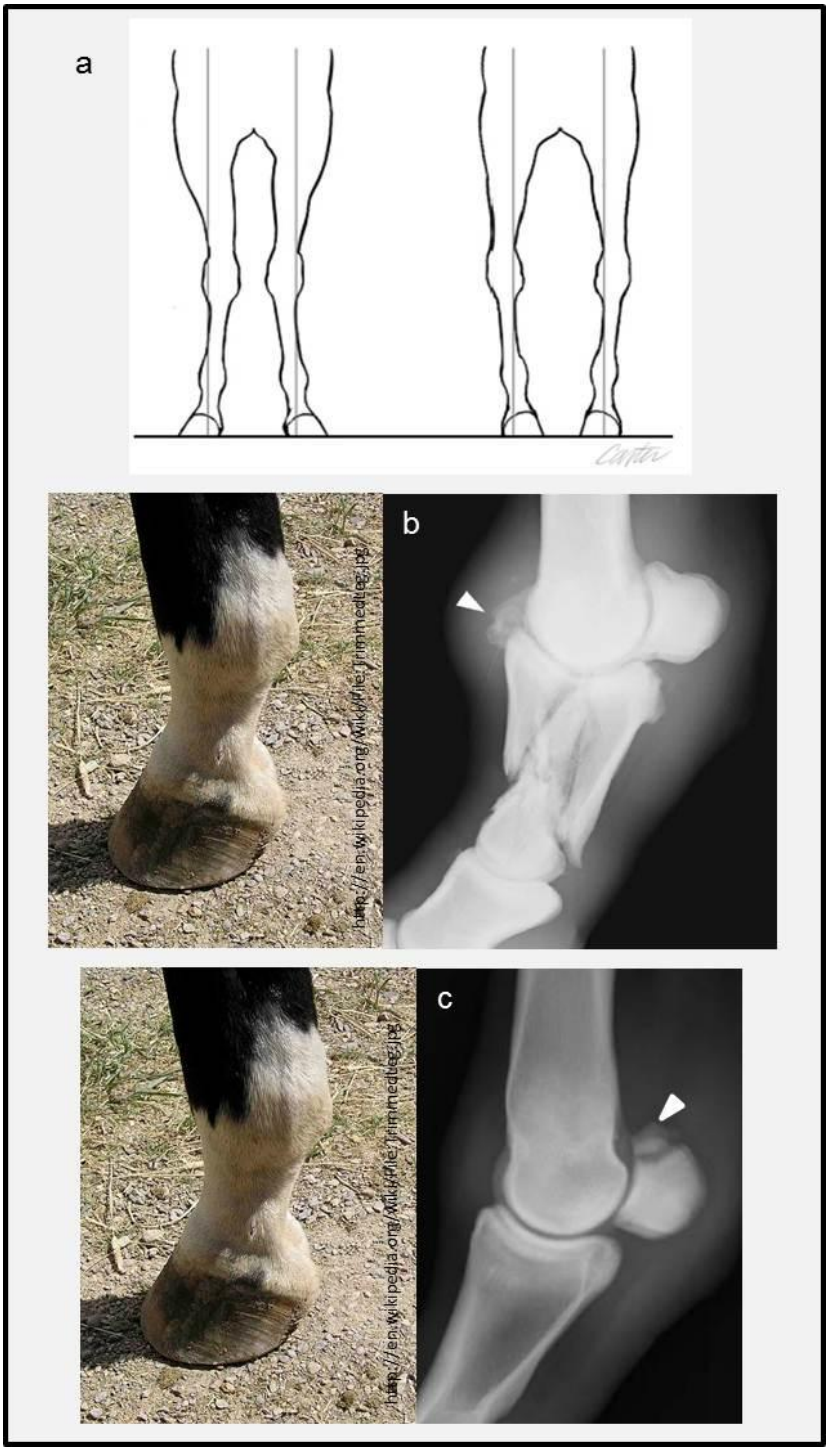

Figure 2-7 Non-OC conditions

a) Angular limb deformity: Drawing of a foal with a valgus deformity in the distal radius region (left) and a foal with a varus deformity of the same location (right). Taken from Bramlage & Auer (2006) [116]; b) Fetlock chip: A comminuted first phalanx fracture. Taken from Vanderperren & Saunders (2009) [117]; c) Sesamoid fracture: a lateral apical sesamoid fracture in a yearling. Taken from Schnabel *et al.* (2007) [118].

The additional data concerning, for example, sex of the horses, farm of origin and the presence of non-OC conditions made it possible to reduce potential confounding in subsequent analyses. However, since horses were pre-selected for their case/control status with an unspecified degree of matching, as opposed to being sampled at random from the population, no conclusions can be drawn with respect to the effect of these factors on the incidence of OC. The final phenotypically enriched dataset consisted of 384 samples with associated OC status, 348 from the US and 36 from the UK, and 817 samples with no OC status from the UK, giving a total of 1,201 samples. Whilst the original data collection of the population dataset samples required horses to be categorised as cases or controls for the diseases of interest, for all analyses in this thesis, the horses were treated as a single population sample. Table 2-3 provides details of which subsets of data have been used in the following chapters.

Table 2-3 A description of data used in subsequent chapters

Chapter	Subset used, no. of samples & their origin
3	Population control dataset – 817 UK samples OC dataset – 348 US samples
4	Population control dataset – 817 UK samples
5	Simulated data only
6	All 1,201 samples
7	OC dataset – 348 US samples
8	OC dataset – 348 US samples
9	All 1,201 samples, categorised by country of origin

Chapter 3: Linkage disequilibrium and effective population size in the Thoroughbred horse

3.1 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 [119]. The efficacy of genomic techniques such as genome-wide association studies (GWAS), marker assisted selection (MAS) and genome-wide evaluation (GWE) 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 [120-122] and the opportunity exists to use validated loci for MAS in the future as some success has already been seen in the localisation of QTL for simple diseases [63, 64, 123]. As has been shown in human studies, when applied to complex diseases the outcomes of GWAS are generally less successful [45] and therefore, the GWE techniques of Meuwissen *et al.* (2001) [52] may become more attractive. The opportunities will depend on the extent of LD and therefore the characterisation 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 idealised population that would give rise to the same rate of inbreeding as observed in the actual breeding population [124]. Deterministic equations derived by Daetwyler *et al.* (2008) [61] show that, once the N_e for a population is known, the accuracy of GWE 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 and indeed the horse more generally, has yet to be comprehensively characterised and predictions of N_e are limited to those inferred from pedigree data which itself may be inaccurate [125]. An early study by Tozaki *et al.* (2005) [126] based on 300 horses, concluded that useful LD in the Thoroughbred extends up to 7cM, but this study covered only one small region of the genome. More recently, Wade *et al.* (2009) [23] investigated LD across ten 2Mb 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 [127-129]. Studies have also been done 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 [130-132].

The objective of this study was to characterise LD in two geographically distinct samples 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 N_e 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.

3.2 Materials and methods

3.2.1 Genotypic data

The data for this study comprised the population control dataset sourced in the UK (n=817) and the OC dataset samples that were sourced in the US (n=348), genotyped at the 50,707 SNPs that passed preliminary quality control (QC). Further QC carried out as part of this study led to the removal of additional SNPs on the basis of poor genotyping quality, deviations from Hardy-Weinberg equilibrium (HWE) and monomorphism; these exclusions are detailed in Table 3-1. The genotyping rate once these exclusions had been made was greater than 99%, with no individuals having more than 10% SNPs missing. Previous studies have demonstrated that including markers with low minor allele frequencies (MAF) can bias LD estimates

[131, 133, 134], therefore a MAF threshold of 0.10 was also imposed on the data (see Table 3-1). This study used only autosomal markers.

Table 3-1 SNP exclusions made during quality control

	No. of SNPs excluded	
Criteria[I]	UK	US
>5% missing genotypes	21	8
Deviation from Hardy-Weinberg equilibrium (p<0.0001)	173[II]	25[III]
Monomorphic in sample	4,086	4,778
Minor allele frequency <0.10	9,286	9,361

[I]Quality control was carried out in Plink [105, 106]

[II]Since case/control groups were unknown, HWE was tested using all samples

[III]HWE tested in control samples

3.2.2 Linkage disequilibrium

The measurement of LD used throughout this study is the squared correlation coefficient between SNP pairs (r^2) [135], computed as:

Equation 3-1 $$r^2 = \frac{D^2}{p_A p_a p_B p_b},$$

where, $D = p_{AB} - p_A p_B$ and p_A, p_a, p_B and p_b, are the frequencies of alleles A, a, B, and b, respectively. An EM algorithm [136], written by Dr. Ricardo Pong-Wong, 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_equineSNP

50.xls). In order 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 3-2). 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 calculated for all non-syntenic markers, resulting in a total of 418,500 pairwise comparisons.

Table 3-2 Distance classes and bin ranges for LD summary

Class	Minimum distance (Mb)	Maximum distance (Mb)	Within class bin distance range (Mb)	No. of Bins
1	0	0.5	0.01	50
2	0.5	20.5	1	20
3	20.5	190	10	17

3.2.3 Modelling decline of linkage disequilibrium with distance

Under the assumption of an isolated population with random mating, Sved (1971) [137] derived an approximate expression for the expectation of r^2, such that:

Equation 3-2 $$E(r^2) = (1 + 4N_e c)^{-1},$$

where, N_e is effective population size and c is the recombination frequency. In this paper, as in previous studies [92, 130-132, 138, 139], c is replaced by map distance in Morgans. The validity of this replacement is explored in Chapter 4. Based on this formula, a non-linear least squares approach to statistically model the observed r^2 was implemented within R [140] using the following model:

Equation 3-3 $y_i = 1/(a + 4bd_i) + e_i$,

where, y_i is the value of r^2 for SNP pair *i*, at linkage distance d_i 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 (National Center for Biotechnology Information (2009). *Map View* [Online]. Available from: http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9796), and total chromosome genetic length inferred from the equine linkage map [22] (see Table 3-3). Marker pairs with less than 100bp 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* [130, 141] and at small distances gene conversion contributes to the breakdown of LD [35, 139, 142]. The minimum MAF threshold of 0.10 was also applied here as Equation 3-2 may be a poor approximation when allele frequencies are close to zero [141, 143]. This model was applied to data for each autosome in turn and parameter estimates combined by meta-analysis in R [140] using an inverse variance method for pooling and a random effects model based on the DerSimonian-Laird method [144] (see A.i for further details).

Table 3-3 Chromosome specific megabase to centiMorgan conversion ratios

Chromosome	Length (Mb)	Length (cM)	cM/Mb Ratio
ECA1	186	194	1.04
ECA2	121	129	1.07
ECA3	119	120	1.01
ECA4	109	123	1.13
ECA5	100	100	1.00
ECA6	85	127	1.49
ECA7	99	102	1.03
ECA8	94	109	1.16
ECA9	84	105	1.25
ECA10	84	106	1.26
ECA11	61	65	1.07
ECA12	33	58	1.76
ECA13	43	58	1.35
ECA14	94	153	1.63
ECA15	92	97	1.05
ECA16	87	111	1.28
ECA17	81	71	0.88
ECA18	82	88	1.07
ECA19	60	56	0.93
ECA20	64	81	1.27
ECA21	58	76	1.31
ECA22	50	80	1.60
ECA23	56	56	1.00
ECA24	47	47	1.00
ECA25	40	49	1.23
ECA26	42	24	0.57
ECA27	40	93	2.33
ECA28	46	63	1.37
ECA29	34	75	2.21
ECA30	30	50	1.67
ECA31	25	41	1.64

3.2.4 Ancestral effective population size estimation

Rearrangement of Equation 3-2 allows the prediction of effective population at a given point in time, expressed as generations in the past [130, 138, 141]:

Equation 3-4 $$N_T(t) = (4c)^{-1}\left[\left(r_c^2\right)^{-1} - 1\right],$$

where, N_T is the effective population size t generations ago, c is the distance between markers in Morgans, r_c^2 is the mean value of r^2 for markers c Morgans apart and $c = (2t)^{-1}$, assuming linear growth [138]. As previously, marker pairs with less than 100bp between them and SNPs with MAF less than 0.10 were excluded from this analysis. To compute N_T, the number of prior generations was selected and a suitable range for c was calculated (see Table 3-4). The binning process was designed to ensure sufficient SNP pair comparisons within each bin to get a representative estimate of r^2. The 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 suggested by Hayes *et al.* (2003) [138] to reduce the variability of estimates of N_T caused by finite population size.

3.2.5 Sample size comparison

Sample size has been shown to bias LD estimates to varying degrees, depending on the distance between markers [128]. Whilst this relationship is explored more fully in Chapter 4, here, in order to differentiate between an effect of sample size versus a true difference between the UK and US population samples, a random sample of 348 horses was drawn from the UK dataset. The analyses described above were then repeated on this subset for ECA1 markers and the results compared to corresponding results from the full UK dataset and the US dataset.

Table 3-4 Description of generation binning process

Generation range applied to	No. of generations represented by each bin	Example for first bin		
		Generation	Generation range	Corresponding distance range (Morgans)
1 - 10	1	1	0.5 to 1.5	0.33 to 1.0 [6.7×10^{-1}]
20 - 100	10	20	15 to 25	0.02 to 0.033 [1.3×10^{-2}]
200 - 1,000	100	200	150 to 250	0.002 to 0.0033 [1.3×10^{-3}]
2,000 - 10,000	1,000	2,000	1,500 to 2,500	0.0002 to 0.00033 [1.3×10^{-4}]
20,000 - 100,000	10,000	20,000	15,000 to 25,000	0.00002 to 0.000033 [1.3×10^{-5}]
200,000 - 500,000	100,000	200,000	150,000 to 250,000	0.000002 to 0.0000033 [1.3×10^{-6}]

3.3 Results

3.3.1 Genotypic data

Of the 52,063 autosomal SNPs on the 50K SNP chip, 34,848 (66.9%) and 34,221 (65.7%) remained after filtering in the UK and US datasets, respectively. This resulted in more than 20 million pairwise comparisons being made per dataset. The number of SNPs per autosome remaining after exclusions ranged from 414 to 2,760 and was closely related to chromosome length, as shown in Figure 3-1. The average distance between adjacent markers (±SD) was 64.05±86.84kb in the UK dataset and 65.24±86.76kb in the US dataset, with the distance between adjacent SNPs ranging from 1bp to 2849kb. In both sample groups, the MAF of remaining SNPs followed a uniform distribution and averaged (±SD) 0.30±0.12.

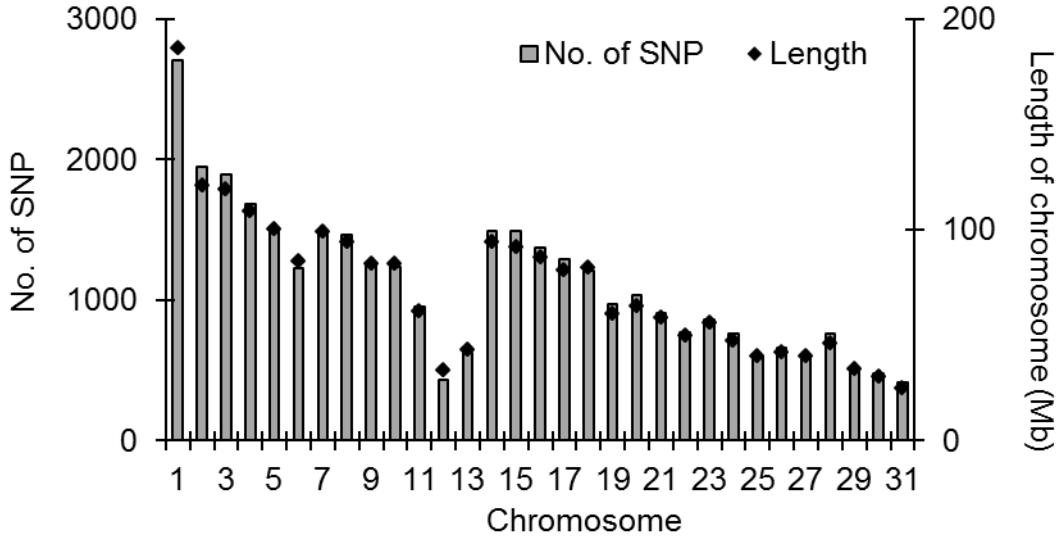

Figure 3-1 Chromosome length and the number of SNPs per chromosome (average of UK and US)

3.3.2 Linkage disequilibrium

Linkage disequilibrium declined with increasing distance between SNP pairs, as shown in Figure 3-2. The most rapid decline was seen over the first 0.2Mb 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 log-transformed distance (Figure 3-3). Linkage disequilibrium values were slightly higher in the US dataset across the full range of marker pair distances, with the relative difference increasing with increasing distance between markers. The coefficient of variation of r^2, as calculated in the UK dataset, increased from 0.6 at 5kb to a maximum of 2.2 at 15Mb, subsequently decreasing and remaining below 1.5 for distances greater than 50Mb. A total of 10,130 and 14,191 SNP pairs were in complete LD (r^2=1) in the UK and US datasets, respectively, and of these, 5,139 and 6,146 were adjacent pairs. The mean (±SD) r^2 between adjacent SNP markers was 0.46±0.36 and 0.47±0.37 in the UK and US datasets, respectively. Slightly more than half the adjacent SNP marker pairs exhibited $r^2 \geq 0.3$. The mean (±SD) r^2 between random non-syntenic markers in the UK dataset was $0.0018 \pm (2.49 \times 10^{-3})$, with the equivalent measure in the US dataset being slightly higher at $0.0044 \pm (6.20 \times 10^{-3})$. In both cases, the value was similar to that observed between syntenic markers at distances greater than 100Mb (Figure 3-2).

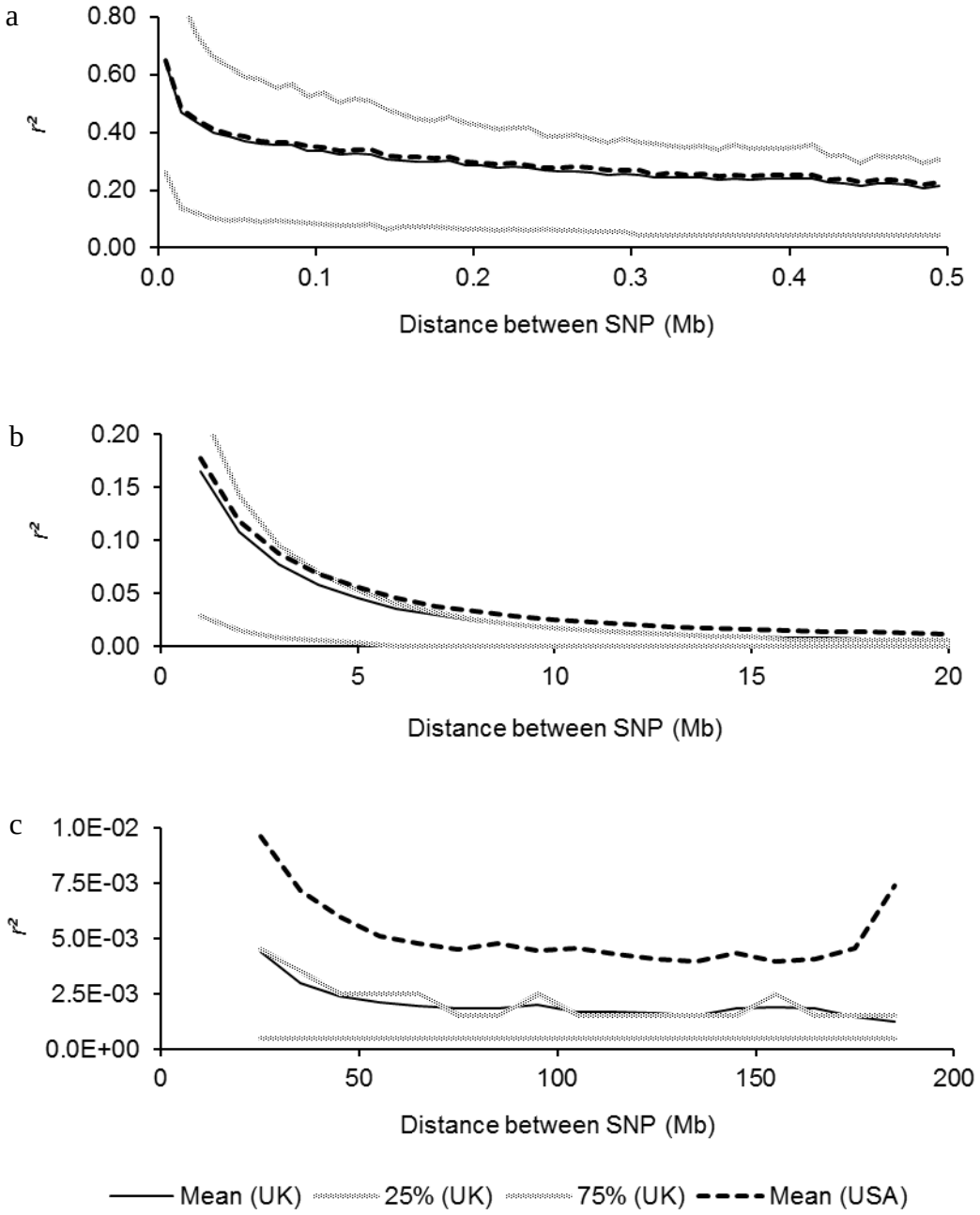

Figure 3-2 Average LD measured by r^2, pooled over autosomes and plotted against the median of the distance bin range

a) Distance range from 0 to 0.5Mb, in bins of 0.01Mb

b) Distance range from 0.5 to 20.5Mb, in bins of 1.0Mb

c) Distance range from 20.5 to 190Mb, in bins of 10.0Mb

3.3.3 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] for the UK dataset. For the US dataset, corresponding estimates were 2.31 [2.23; 2.40] and 84.4 [78.8; 89.9]. 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 0.03Mb, as shown in Figure 3-3. 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 chromosome length and estimates of *b*, but there was no such relationship with estimates of *a* (Figure 3-4). The interpretation of *b* as an estimate of effective population size is considered in the discussion.

3.3.4 Ancestral effective population size

An initial pattern of decreasing N_e, with values of over 3,000 estimated in the distant past and values closer to 100 estimated at 20 generations ago was observed (Figure 3-5). The results suggest that an increase in N_e has occurred over the past ten generations. This increase appears much greater in the UK population which reaches a maximum of approximately 190 two generations ago, compared to an estimate of just 90 in the US population. 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) (Figure 3-6).

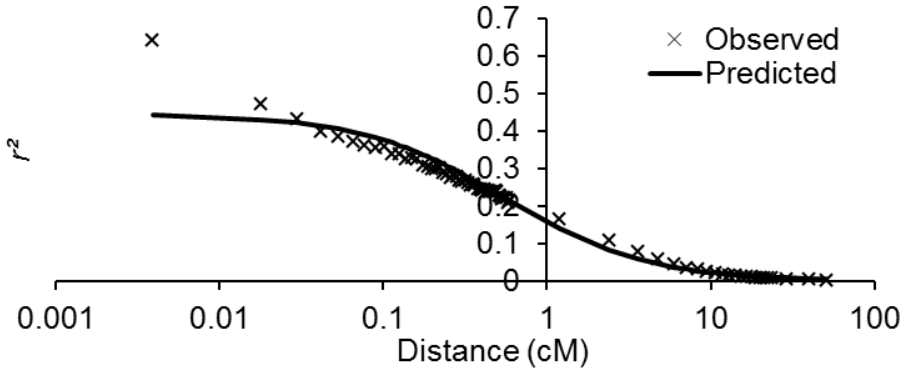

Figure 3-3 Predicted r^2 versus observed r^2 against mean distance between markers (on a log scale). Predicted r^2 calculated using Equation 3-3 with a=2.25 and b=103 (results for the UK dataset).

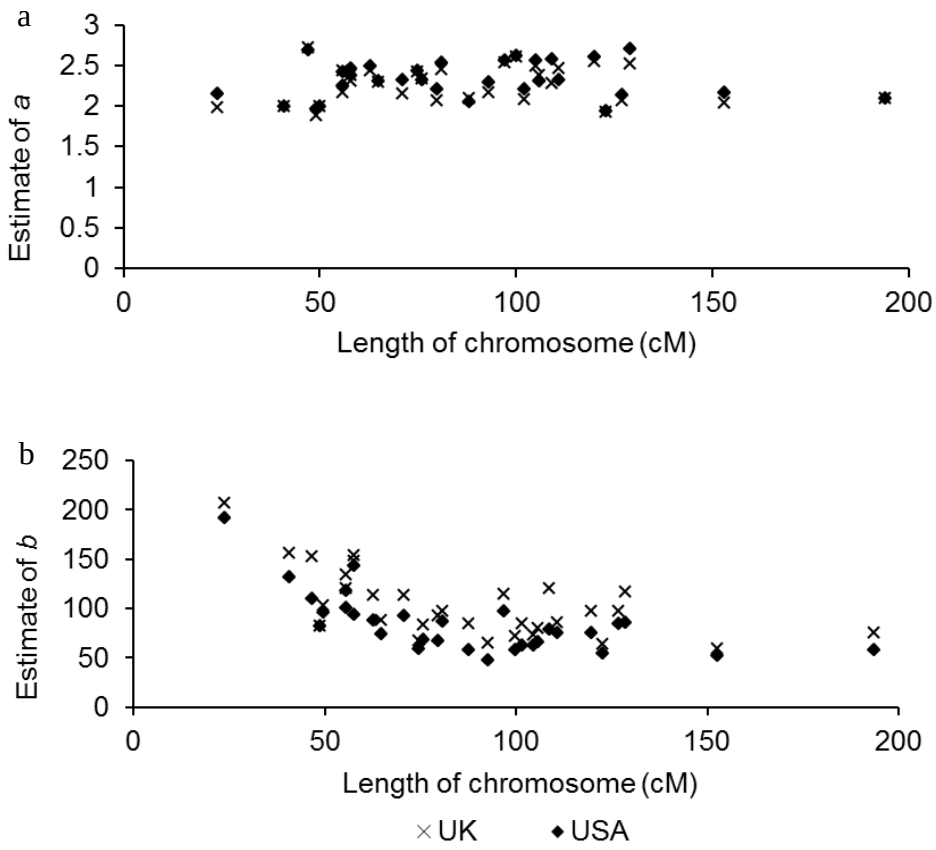

Figure 3-4 Parameter estimates from the modelling of Equation 3-3 by chromosome, plotted against chromosome length according to the equine linkage map of Swinburne *et al.* (2006) [22] a) Estimates of *a*; b) Estimates of *b*

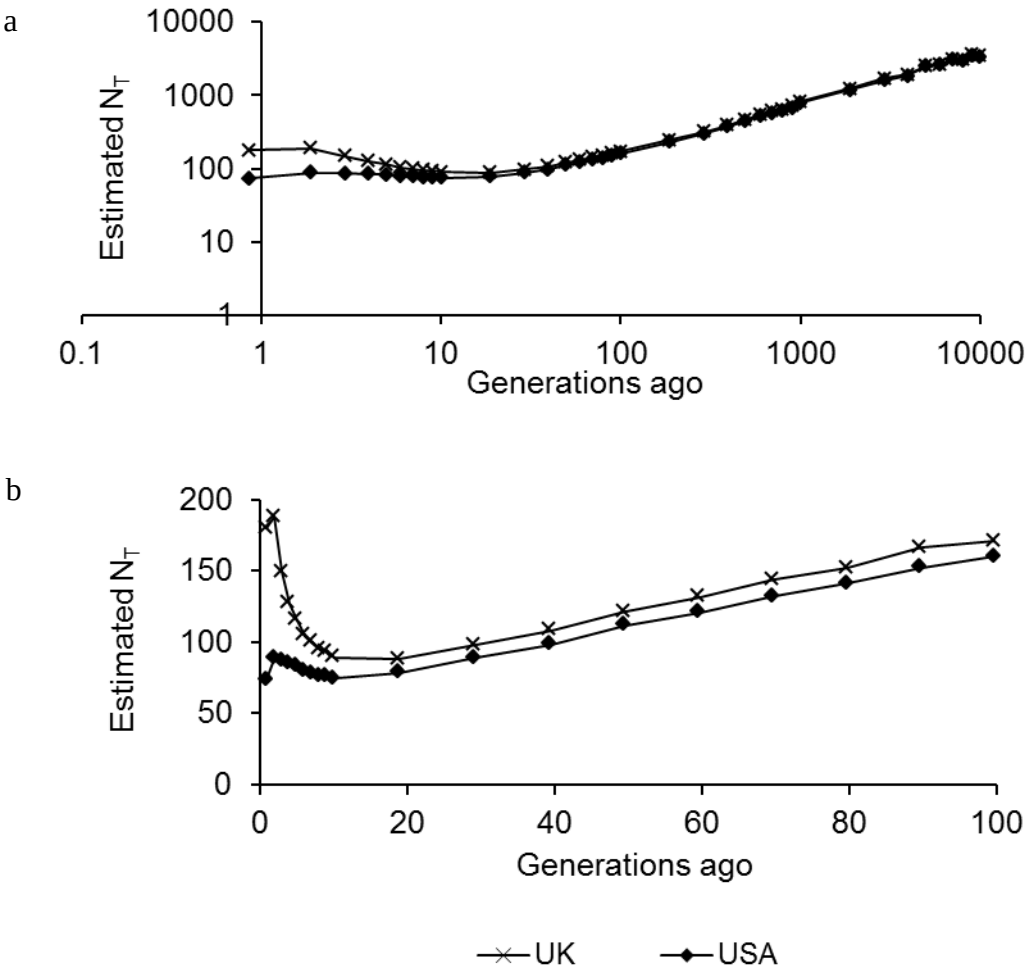

Figure 3-5 Average $\hat{N}_T$ plotted against generations in the past

a) Generations 1 to 10,000 on a log scale

b) Generations 1 to 100

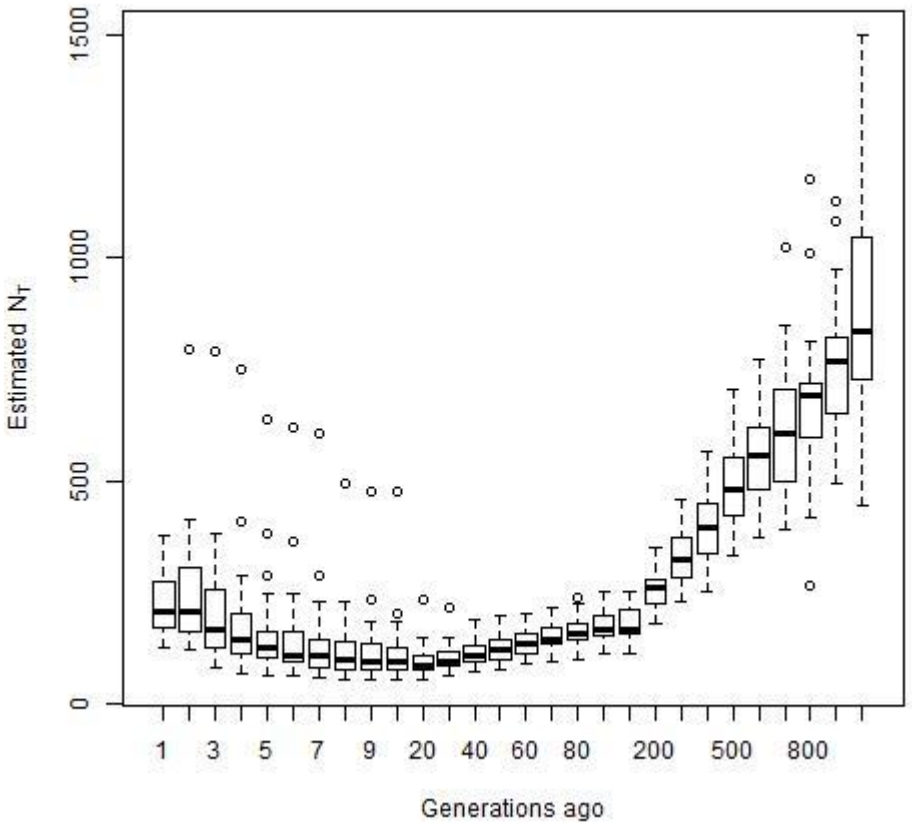

Figure 3-6 Boxplot representing $\hat{N}_T$ plotted against generations in the past (on a non-linear scale), truncated at 1,000 generations. Variation at each time point reflects variability in estimates within generation bins between the 31 autosomes (results for the UK dataset).

3.3.5 Sample size comparison

As shown in Figure 3-2, LD values were consistently higher in the US dataset compared to the UK sample. In the subsequent subset analysis, the level of LD in the subset of UK data closely followed that of the full UK dataset for marker pairs up to around 20Mb apart (Figure 3-7). However, at distances greater than this, LD levels in the subset were approximately midway between the full UK dataset and the US dataset (Figure 3-7).

The results of the non-linear regression modelling of the decline of LD are shown in Table 3-5. Estimates of *a* were relatively consistent across the three datasets, whilst estimates of *b* differed the most between the UK datasets and the US dataset. Estimates of ancestral effective population size for the UK subset tracked that of the full UK dataset until around generation seven (data not shown), at which point estimates for the UK subset fell between that of the full UK sample and the US sample (Figure 3-8).

Table 3-5 Parameter estimates from non-linear regression modelling of ECA1 data

	Estimate (SE)		
Parameter	**UK**	**UK Subset**	**US**
a	2.09 (2.13 x 10^{-3})	2.11 (2.13 x 10^{-3})	2.08 (2.22 x 10^{-3})
b	75.2 (0.08)	73.2 (0.08)	58.0 (0.07)

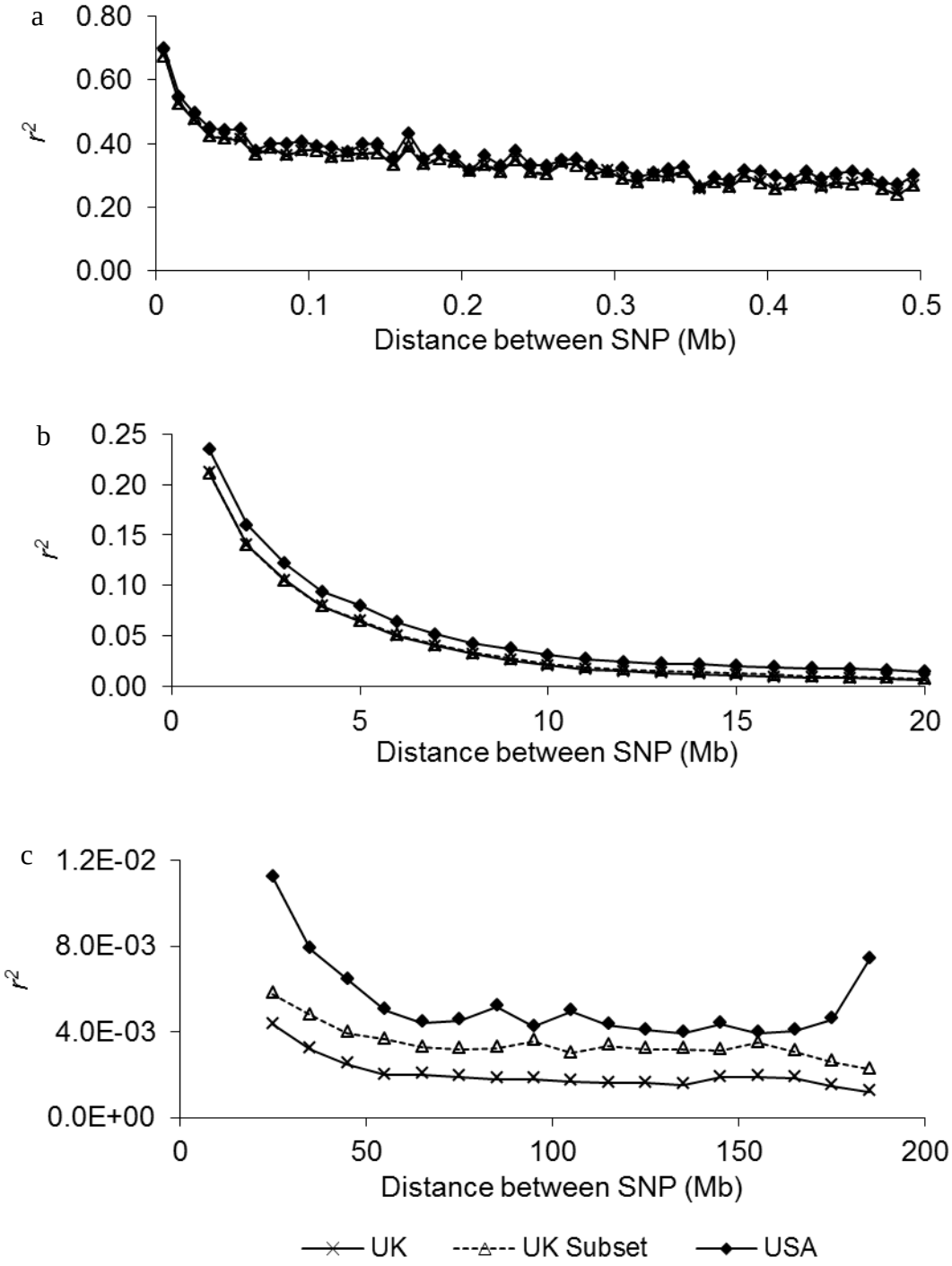

Figure 3-7 Average LD on ECA1 measured by r^2 and plotted against the median of the distance bin range

a) Distance range from 0 to 0.5Mb, in bins of 0.01Mb

b) Distance range from 0.5 to 20.5Mb, in bins of 1.0Mb

c) Distance range from 20.5 to 190Mb, in bins of 10.0Mb

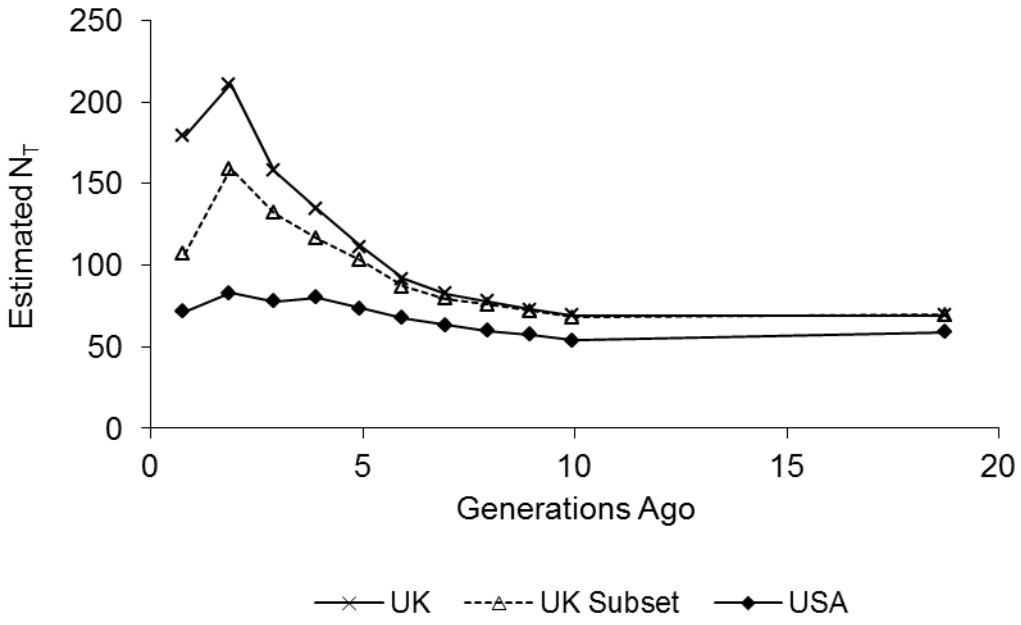

Figure 3-8 Average $\hat{N}_T$ plotted against generations in the past, truncated at 20 generations

3.4 Discussion

This study provides an overview of LD in the Thoroughbred using a high density SNP panel. Validation work by Khatkar *et al.* (2008) [128] on their cattle data suggests that my minimum sample size of 348 horses, is more than sufficient to obtain an unbiased picture of LD. The pattern of decline of LD with distance in these populations is consistent with that reported by Wade *et al.* (2009) [23] in a sample of 24 Thoroughbreds, with both data sets exhibiting a decrease in r^2 from around 0.6 to 0.2 when the distance between markers is increased to 0.5Mb. The LD observed is higher at short distances and more extensive than that observed in human populations [145]. Linkage disequilibrium declines more slowly in the Thoroughbred populations studied here than in the range of cattle populations studied by de Roos *et al.* (2008) [130], with average r^2 remaining above 0.3 for distances up to 185kb compared with a maximum distance of 35kb in the cattle data.

The mean value of r^2 between non-syntenic SNPs was 0.0018 and 0.0044 in the UK and US datasets, respectively, 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 values observed here are of a similar magnitude to that observed by Khatkar *et al.* (2008) [128] in a sample of over 1,500 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, whilst the larger non-syntenic LD value in the US sample here is likely to be due to a smaller sample size, in the Australian Holstein-Friesian cattle sample (n=1,546) of Khatkar *et al.* (2008) [128], it may reflect a lower N_e in this cattle population. The low LD seen between non-syntenic SNPs in both of the Thoroughbred populations suggests that the LD created by admixture during breed formation [125], 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 [129]. At distances greater than 100Mb, 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.

The LD observed in the US sample was slightly higher than that seen in the UK sample. The results of the sample size comparison suggest that, at distances <20Mb, the lower number of samples does not explain this difference. At long distances (>20Mb) and for the non-syntenic case, sample size appears to have some effect on estimates of LD but still does not fully account for the difference in LD between the UK and US samples. Whilst the remaining differences may be due to differences in demographic history, for example, greater inbreeding in the US Thoroughbred population, it is also possible that the results of the US analysis are biased by the presence of more half-sibs and other close relatives, with the average relationship of the US horses being greater than that of the UK. For this reason, the remainder of the discussion will focus on the results from the UK dataset which is likely more broadly representative of the wider Thoroughbred population.

By using Sved's (1971) [137] formula for the expectation of r^2, a non-linear regression model was fitted to the data in order 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 Equation 3-3 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). The estimate of *a* calculated here supports an alternative version of Sved's (1971) [137] equation, derived by Tenesa *et al.* (2007) [139], 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 Abasht *et al.* (2009) [146], Toosi *et al.* (2010) [134] and Zhao *et al.* (2005) [147] is appropriate. The impact of such model assumptions are explored in Chapter 4. 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 the results. A significant negative relationship between chromosome length (cM) and estimates of *b* from the non-linear model was observed, suggesting LD is higher in longer chromosomes. This contrasts with the findings of Tenesa *et al.* (2007) [139] who observed a positive relationship, but is in keeping with the observations in domestic livestock species of Khatkar *et al.* (2008) [128] and Muir *et al.* (2008) [148].

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, see Toosi *et al.* (2010) [134]. As an alternative, Figure 3-5 shows the results following the approach of Hayes *et al.* (2003) [138], that is, of 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 dip in N_e observed potentially representing any one of a number of scenarios including a founder event, an immigration event, a hybridisation event or any combination of these [119]. Therefore, it is useful to consider the results 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 [125]. 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) [134]. 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 results here 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 [149], that Thoroughbreds born in the 1960s were separated from seventeenth century founders by an average of 21.5 generations. Cunningham *et al.* (2001) [150] also found evidence for a population bottleneck at the time of breed formation.

The reliability of this method depends on the technical implementation (see Chapter 4) 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) [150] calculate the effective number of studbook founders of the Thoroughbred to be 28.2. Since this relies on calculating the long term contributions of the founders, quantitative genetic theory suggests that the N_e for this generation is twice this value if in HWE [151], 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) [149] 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 $1-\prod_{4}^{25}\left(1-{}^{1}\!\!\diagup\!\!{}_{2N_e}\right)$, with N_e estimated from Figure 3-5. The value obtained of 0.112 is remarkably close. Therefore, the 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 an

increase in N_e. However, the results of Binns *et al.* (2012) [152], published after this work had been completed, suggest an increase in inbreeding in the Thoroughbred since 1960. This apparent contradiction may be due to sample differences or, more likely, due to the technical limitations of the methods. The limitations of the methods used in this chapter are explored more fully in the next chapter.

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

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 has been shown to be inflated by $1/r^2$ when it is necessary to rely on marker-QTL LD, rather than on the QTL itself [153] 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 [130] and the consensus is that an average $r^2>0.3$ will permit reasonable sample sizes to be employed for GWAS [35, 128, 153]. In this dataset, markers 185kb 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 GWE and given the high variability of r^2 values at small distances, this is an underestimation of the actual number of SNPs needed. Indeed, in this study of 34,848 SNPs, just over half (54%) the adjacent SNP pairs exhibited r^2 values of at least 0.3 and this fell to 46% when SNP pairs with $r^2=1$ were excluded. In reality, because a causal mutation would be in LD with a marker on each side, this represents a good, though variable, level of coverage of the genome. With MAS also relying on close and consistent linkage between markers and QTL, the high LD observed here is promising. Genome-wide evaluation 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 [154]. Deterministic equations derived by Daetwyler *et al.* (2008) [61] 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 the current N_e estimate of ~180 (Figure 3-5) and assuming a random mating population, the M_e for the UK Thoroughbred population is ~1,500 [155]. Thus, the potential accuracy of GS in this population for a range of scenarios can now be predicted (see Table 3-6).

Table 3-6 Predictions of accuracy for genome-wide evaluation. Based on Daetwyler *et al.* (2008) [61] and assuming a disease prevalence of 25%.

		Sample Size		
N_e	Heritability	300	600	1200
100	0.1	0.15	0.21	0.29
	0.2	0.21	0.29	0.39
150	0.1	0.13	0.18	0.25
	0.2	0.18	0.25	0.34
200	0.1	0.11	0.16	0.22
	0.2	0.16	0.22	0.30

In summary, dense SNP genotype data was used to characterise LD and make inferences regarding ancestral N_e for two samples of Thoroughbred horses. In the population studied, LD extended for long distances, reaching base line levels at around 50Mb. From the decay in LD with distance, ancestral N_e was inferred and a decrease in N_e since the distant past reaching a minimum of ~90 20 generations ago, followed by an increase until the present time was observed. 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. A more detailed exploration of this approach is presented in the next chapter. The results from this chapter 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 50K SNP chip.

Chapter 4: The estimation of effective population size using linkage disequilibria: A methodological review

4.1 Introduction

Effective population size (N_e), which is defined as the number of individuals in a population with random selection and perfect random mating that would give rise to the same rate of inbreeding as observed in the actual breeding population [124], is an important parameter in a wide variety of biological fields. Whilst N_e in livestock populations can be estimated from pedigree information, or by using formulae which relate demographic parameters such as sex ratio and variance of reproductive success to N_e [124, 156], or predicted through expected long term genetic contributions [151], such approaches may not be appropriate in practice and cannot be applied in the case of natural populations where parameterisation is lacking.

In finite populations, genetic drift plays a crucial role in determining the amount and distribution of genetic variation within a population and through this process N_e influences genome structure. It is therefore possible that N_e for a population may be predicted from the genome structure using neutral and selection-free genetic markers in a random sample of the population. A range of summary statistics are available that enable such predictions to be made, for example, measures of heterozygosity, temporal changes in allele frequencies and measures of linkage disequilibrium (LD). Methodologies using these statistics have been reviewed by Wang (2005) [119]. This study focuses on a method by which N_e can be estimated from the LD between syntenic markers in the genome. Linkage disequilibrium describes the non-random association of alleles at different loci resulting from processes such as migration, selection and genetic drift in finite populations [119]. Whilst there are a number of measures of LD, the measure used extensively in this paper is r^2 [135].

Genetic models describing the relationship between N_e and LD appear in the literature from the 1960's onwards [135, 137, 141, 157-161]. In particular, Sved (1971) [137] and Sved & Feldman (1973) [161] derived a formula for the expectation of r^2 ($E[r^2]$) between a pair of markers as a function of N_e and recombination

frequency (*c*), validated by simulation for distances up to 10cM. Such work implied that the measurement of LD might provide information about N_e. However, Weir & Hill (1980) [162] showed that methods relating E[r^2] to N_e will also depend on sample size (*n*). Hill (1981) [141] explored the use of such equations in inferring N_e from LD, suggesting that LD at closely linked markers would reflect ancient population history, whilst LD between markers further apart would reflect more recent events. Hayes *et al.* (2003) [138] built on this idea by deriving a relationship between the lengths of homozygous segments (CSH) and generations in the past, under the assumption of constant linear growth.

In terms of empirical application and validation, feasibility was considered to be the limiting factor in early work [141]. However, this conclusion was made based on only 18 pairs of loci being available; subsequent developments in genotyping technology, such as dense single-nucleotide polymorphism (SNP) chips, have substantially increased marker density, and some of the limitations can now be overcome. Along with this new opportunity to exploit historical formulae has come the challenge of validating the theories, derived under a variety of assumptions, with empirically derived data. Since 2003, Hayes' formula [138], combined with Sved's (1971) [137] original formula for E[r^2] (together referred to here as the variable N_e method), has become an increasingly popular way of estimating past and present N_e in human and more often, in livestock populations [130-132, 139, 163, 164]. Due to concerns relating to accuracy and underlying assumptions, these authors and others investigating non-syntenic LD [165, 166], have implemented various changes to the formula for E[r^2] presented by Sved (1971) [137] in order to take account of, for example, sampling (as mentioned above) [132, 163] and mutation [131, 139, 164]. However, despite a continual effort being made to improve the fit of the model to empirically derived data, few studies have explored, in a comparative fashion, the importance of these modifications. Furthermore, the issue of the assumptions underlying these techniques has been raised [167].

This paper considers the wide range of approaches which previous authors have taken to estimate N_e from LD data and the impact the assumptions have on the estimates. The theoretical basis of each of the approaches is considered; these are

then tested firstly on simulated data and secondly on the sample of 817 UK Thoroughbreds sourced in the UK and genotyped for the Illumina Equine SNP50 Genotyping BeadChip. The approaches examined include methods for both variable and constant N_e. The results are considered in terms of potential applications of the general method.

4.2 Materials and methods

I will address the theoretical basis for the various models tested, and then describe their validation in simulated datasets and their application to Equine SNP50 genotype data. Models are tested first under the assumption of constant N_e by non-linear regression (constant N_e method) and subsequently under the assumption of linear growth in N_e as in Hayes *et al.* (2003) [138] (variable N_e method). Throughout this paper, *c* represents recombination frequency or fraction (no units) and *d* refers to linkage distance (in Morgans). In the case of the equine dataset, linkage distance between markers was calculated under the assumption of a genome-wide linear relationship such that 1cM≈1Mb.

4.2.1 Theory for constant effective population size

4.2.1.1 An expression for E[r²] and its interpretation

Hill & Robertson (1968) [135] introduced r^2, a squared correlation coefficient, as a normalised measure for the amount of disequilibrium between two loci where $r^2 = D^2\left[p_A(1-p_A)p_B(1-p_B)\right]^{-1}$, $D = p_{AB} - p_A p_B$, and p_A and p_B are the frequencies of alleles A and B. They recognised that the value of r^2 is almost entirely determined by the product of effective population size, N_e, recombination frequency, *c*, and generations since mutation, measured proportional to N_e. For loci initially in perfect disequilibrium, the limiting value of E[r^2] over time approaches $(4N_ec)^{-1}$ for large N_ec, and 1 for N_ec tending to 0. This led to an approximation of the form:

Equation 4-1 $E\left[r^2\right] = \left(1 + 4N_e f(c)\right)^{-1}$ [137]

In Hill & Robertson (1968) [135] and Sved (1971) [137], $f(c) = c$, with c assumed to be small. Sved (1971) [137] used simulation to show that Equation 4-1 holds with 'reasonable accuracy' assuming no selection or mutation.

However, this construction does not account for mutation. In the presence of mutation, only some of the loci will have reached such an asymptotic state at any given time, with the remainder having lower values of r^2, and therefore at c=0, E[r^2]<1. With mutations occurring at different time points on the same chromosome, the expected association between polymorphisms is incomplete even with physical linkage. With the same infinite sites model as used by Hill & Robertson (1968) [135], Ohta & Kimura (1969) [158, 159] used diffusion equations to calculate $\sigma_d^2 = E\left[D^2\right] / E\left[p_A(1-p_A)p_B(1-p_B)\right]$; whilst the ratio of expectations is not the same as the expectation of the ratio, they showed that $E\left[r^2\right] \approx \sigma_d^2$. Using this approximation, Ohta & Kimura (1971) [160] obtained an equilibrium expression for σ_d^2; in this formulation $\sigma_d^2 = 5/11$ when $N_e c$=0 (a value derived subsequently by Hill (1975) [157]), rather than 1, as implied in Equation 4-1. This expression may also be derived using coalescence theory [168]. The natural extension in Equation 4-1 is then:

Equation 4-2 $E\left[r^2\right] = \left(a + 4N_e f(c)\right)^{-1}$,

where, a=11/5=2.2 so that mutation is accounted for. In this context, Tenesa *et al.* (2007) [139] proposed a=2.

An important assumption in the derivation of Equation 4-1 was that c, was 'small'. However, the length of chromosomes covered by SNP markers often exceeds 100Mb, equivalent to c~0.43, assuming 1Mb≈1cM with Haldane's mapping function. Since such distances cannot be considered 'small', it may be more appropriate to return to the version of Equation 4-1 which was presented by Sved & Feldman (1973) [161]. Here, the inclusion of second order terms in N_e^{-1} and c, led to

$f(c) = c\left(1 - c/2\right)$ and N_e being replaced by (N_e – ½). A common interpretation is that $f(c)=d$ in Equation 4-1, where d is measured in Morgans and this may be questionable. The relationship of d with c can be described using Haldane's mapping function such that $d = -\frac{\ln(1-2c)}{2}$. From Taylor's approximation, for small c, $d \approx c$, but the second order approximation is $d \approx c(1+c)$, and it can be seen that the result of Sved & Feldman (1973) [161] is not a second order approximation to d and therefore provides no justification for setting $f(c)=d$. The frequent need to infer the linkage map distance between markers from their physical locations with only limited direct information on recombination rates, represents a further compromise in the determination of c for input to Equation 4-1. It has become common with cattle and human genomes to assume $d = 10^{-8}\delta$, where δ is the distance in base pairs; this approximation can also be used in the case of the equine genome (27.7M versus 2.68Gb).

4.2.1.2 Sampling effects

Up until this point it has been assumed that D is known exactly for any pair of loci. However, in reality D is estimated from a limited sample of individuals from the population. This additional sampling event can be expected to contribute to the LD observed and therefore any equation for E[r^2] should consist two components; the first being a function of N_e and c as in Equation 4-1 and Equation 4-2, and the second being a function of n, the number of diploid individuals sampled. From the work of Weir & Hill (1980) [162], it can be hypothesised that the necessary modification to formulae for E[r^2] depends on the approach taken to estimate r^2. In instances where the haplotypes of diploid individuals are known without error, the minimum term of adjustment, $(2n)^{-1}$, should be applied, where n is the number of individuals sampled. Alternatively, if coupling and repulsion double heterozygotes cannot be distinguished and r^2 is estimated using a composite coefficient, e.g. Burrows [169], the maximum term of adjustment, n^{-1}, is more appropriate. However, as analytical methods progress, we are presented with an ever increasing range of methods to predict r^2 from unphased genotype data, e.g. EM algorithms and phasing algorithms using multiple loci, and there has been no work done to determine the appropriate

term of adjustment for such methods, although it might be expected to reduce from $(n)^{-1}$ towards $(2n)^{-1}$. Here, observed r^2 values are adjusted by subtracting $(2n)^{-1}$ since in the case of the simulated data at least, haplotypes are known, and dense SNP typing is allowing progress towards this state. The extension to Equation 4-2 is thus:

Equation 4-3 $$E\left[r_{adj}^{\ 2}\right] = \left(a + 4N_e f(c)\right)^{-1},$$

where, $r_{adj}^{\ 2} = r^2 - (2n)^{-1}$

4.2.2 Extension to variable effective population size

The rearrangement of Equation 4-1 to solve for N_e, allows the prediction of N_e from LD data. This concept provides the basis of derivation of a formula by Hayes *et al.* (2003) [138] relating LD, or in their case CSH, to the N_e that existed t generations in the past, i.e. a function $N_T(t)$ to be estimated. This mapping from c to t was of the form $t = (2f(c_t))^{-1}$, therefore:

Equation 4-4 $$N_T(t) = \left(4 f(c_t)\right)^{-1}\left(E\left[r_{adj}^2 \mid c_t\right]^{-1} - a\right),$$

where, N_T is the effective population size t generations ago, and the expectation is conditional on the markers being the appropriate distance apart given t and the mapping function $f(c)$. A key assumption stated by Hayes *et al.* (2003) [138] is constant linear growth of N_e with t.

Equation 4-1 to Equation 4-4 allow both current and historical N_e to be estimated from dense marker data. Further, they provide an opportunity to assess the impact of simplifying assumptions made by various authors. I do this through analyses of simulated and real data.

4.2.3 Data sets investigated

4.2.3.1 Simulated data

SNP marker data was generated by forward simulation[6] of a population of random mating individuals with N_e=200, and the number of male and female parents each generation being ½N_e. In each generation, each individual was generated from a male and a female parent selected at random and with replacement, and the gametes were generated via independent recombination events. At generation 0, all individuals had identical genotypes; mutations were introduced at a rate of 10^{-8} (equivalent to 1 mutation per Morgan per generation). Recombination was modelled according to a Poisson distribution with a mean of 1 recombination per Morgan per generation. Haldane's mapping function (i.e. no interference) is implicit in the simulation. In generation 3,001 (by which time mutation-drift equilibrium had been established for >5N_e generations), a set of 750 individuals was generated by random mating as described above and subsequently subdivided into four sets of sample size (n) 50, 100, 200 and 400 individuals, representing 25%, 50%, 100% and 200% of the simulated N_e. Thirty chromosomes of one Morgan (1M) each were simulated using independent runs of the simulation program and these are treated as replicates. The number of segregating sites observed closely followed the theoretical expectation [170], with a mean percentage deviation across the four sample sizes of 1.3%. The expected heterozygosity over all loci was 0.149 and the mean heterozygosity observed when averaged over all loci for the 750 sampled individuals was 0.144. Further details of the simulation validation process can be found in A.ii.

For n=200, the following MAF thresholds were applied: 0.15, 0.10, 0.05 and 0.01, with SNPs below the threshold being excluded. Due to there being a greater number of SNPs for MAF thresholds of 0.05 and 0.01, a random subset was chosen in these cases to give similar numbers of SNPs for all MAF thresholds. For the remaining sample sizes, a MAF threshold of 0.10 was used. This resulted in, for example for

[6] Simulation program written by Ariel Liu and John Woolliams (both of The Roslin Institute).

n=50, an average of 1796 segregating sites remaining per chromosome. This is similar to the number of SNPs observed on equine chromosome 5 (which is 1M in length) at the same MAF threshold when a random sample of size 50 was drawn from the equine data set described below.

4.2.3.2 Equine data

The empirical dataset for this study comprised the population control dataset (*n*=817 Thoroughbreds from the UK) genotyped at the 50,707 SNPs that passed preliminary quality control (QC). Prior to analysis, SNPs genotyped in less than 95% of samples (21) and those that deviated significantly from Hardy-Weinberg equilibrium ($p<0.0001$) (173) were excluded [105, 106]. A further 13,372 SNPs were excluded based on a MAF threshold of 0.10. The decision to exclude these markers was made based on evidence that markers with low minor allele frequencies (MAF) can bias LD estimates [131, 133, 134] and that as a result of this Equation 4-1 may be a poor approximation when allele frequencies are close to zero [141, 143]. This study used only autosomal markers and of the 52,063 autosomal SNPs genotyped, 34,848 (66.9%) remained after filtering. A multi-dimensional scaling analysis showed no evidence of underlying population structure [105, 106]. The average distance between adjacent markers (±SD) was 64.05±86.84kb, with the distance between adjacent SNPs ranging from 1bp to 2849kb. Here, *d* was estimated by $d = 10^{-8}\delta$ and *c* was estimated from *d* according to Haldane's mapping function.

4.2.4 Data analysis

A number of analyses, based on the range of model interpretations described above, were carried out. Where appropriate, model combinations were chosen to reflect those most frequently used in existing literature.

4.2.4.1 Estimating linkage disequilibrium

The measurement of LD used throughout this study was the squared correlation coefficient between SNP pairs (r^2) [135]. In the case of the equine dataset, an EM algorithm [136] , written by Dr. Ricardo Pong-Wong, was implemented to estimate haplotype frequencies and r^2 was calculated (to four decimal places) for all syntenic marker pairs resulting in more than 20 million syntenic estimates. Individuals with a

missing genotype for a given marker were excluded when calculating LD for that marker. In the analysis of the equine dataset, marker pairs with less than 100bp between them were excluded in order to avoid gene conversion contributing to the breakdown of LD [35, 139, 142]. In the case of the simulated data, haplotypes were known and therefore r^2 values were calculated directly for marker pairs.

4.2.4.2 Estimating constant effective population size

Based on the range of model interpretations described above, a non-linear least squares approach to modelling was implemented within R [140]. The statistical model corresponding to Equation 4-1 was of the form:

$$y_i = \left(a + 4bf\left(c_i\right)\right)^{-1} + e_i,$$

where, y_i is the value of r^2 (or adjusted r^2) for SNP pair i, separated by $f(c_i)$, e_i represents the residual and, as appropriate, parameters a and b are estimated iteratively using least squares. Here b represents $\hat{N}_e$ when the models tested were $f(c) = c$ or $f(c) = d$, or $\hat{N}_e - 1/2$ when $f(c) = c\left(1 - c/2\right)$.

In the case of the equine dataset, models were applied to data for each autosome in turn and parameter estimates combined by meta-analysis in R [140] using a weighting method of DerSimonian & Laird (1986) [144] to account for potential variance between chromosomes (see A.i for further details). For the simulated datasets, models were applied to each chromosome and results were averaged, as systematic variation between replicates was not present. Model fit was explored by comparing plots of predicted r^2 with plots of observed values and by inspection of residual plots.

4.2.4.3 Estimating variable effective population size

To compute N_T, for the different models listed in Table 4-1, the number of prior generations, t, was nominated and a range for $f(c)$ appropriate to t was calculated as in Chapter 3 (see Table 3-4). This binning process was designed to ensure sufficient SNP pair comparisons within each bin to get an estimate of r^2 with acceptable precision. The mean distance and the mean r^2 between marker pairs in each bin were

then computed for insertion into the relevant equation from Table 4-1. This process was carried out for each chromosome in turn and also for markers pooled across chromosomes simultaneously, as suggested by Hayes *et al.* (2003) [138] to reduce the variability of estimates of N_T caused by finite population size.

Table 4-1 Description of formulae used in the estimation of variable effective population size

Model Reference in Results	Model[I]	*f*(*c*)
I	$N_T(t) = (4f(c))^{-1}\left((r^2)^{-1} - 1\right)$	c
II	$N_T(t) = (4f(c))^{-1}\left((r^2_{adj})^{-1} - 1\right)$	c
III	$N_T(t) = (4f(c))^{-1}\left((r^2)^{-1} - 1\right)$	d
IV	$N_T(t) = (4f(c))^{-1}\left((r^2_{adj})^{-1} - 1\right)$	d
V	$N_T(t) = (4f(c))^{-1}\left((r^2_{adj})^{-1} - 2.2\right)$	d
VI	$N_T(t) = (4f(c))^{-1}\left((r^2_{adj})^{-1} - 2.2\right)$	c
VII	$N_T(t) - 1/2 = (4f(c))^{-1}\left((r^2_{adj})^{-1} - 2.2\right)$	$c\left(1 - c/2\right)$

[I] **$N_T(t)$ is the effective population size t generations ago and $f(c) = (2t)^{-1}$ [138].**

4.3 Results

4.3.1 Simulated data

4.3.1.1 Impact of minor allele frequency threshold

Changes to the MAF threshold had a considerable impact on parameter estimates under the constant N_e scenario (shown in Table 4-2). For n=200, increasing the MAF threshold led to a decrease in the estimate of parameter a from more than 6 with a threshold of 0.01, down to 1.5 with a threshold of 0.15. Assuming a monotonic relationship between the threshold and a would result in values of 2.2 or 2.0 being obtained with a threshold between 0.05 and 0.10. Similarly, increasing the MAF threshold led to a decrease in $\hat{N}_e$. In nearly all cases, the MAF threshold which led to $\hat{N}_e$ closest to the simulated value was 0.05. With variable N_e models,

changes to the MAF threshold had the greatest impact on distant $\hat{N}_T$, i.e. small *c* (results not shown) since parameter *a* is a more dominant term when the N_e-dependent term is small.

4.3.1.2 Impact of adjustment for sample size

The bias in $\hat{N}_e$ due to sample size is demonstrated by the mean parameter estimates of *b* in Table 4-3. When calculated r^2 values were *not* adjusted for sampling effects, increasing the sample size resulted in an increase in $\hat{N}_e$ and, in nearly all cases, estimates closer to the simulated value of 200; this reflects the reduction in sampling error with increasing *n*. However, when adjustment was made, estimates of N_e were more stable and robust to changes in *n*. Assuming a variable N_e, the adjustment for finite sample size had the greatest impact on recent $\hat{N}_T$, i.e. for large *c*, where r^2 is smallest (Figure 4-1). Without adjustment, the values of r^2 are biased upward, resulting in a downward bias in $\hat{N}_T$, since $\hat{N}_T$ is related to the reciprocal. After subtraction of the error term the bias is reduced, but variation across estimates is increased (Figure 4-2); both effects are a result of the reciprocal transformation. Nevertheless, the adjustment reduces the predicted mean square error (MSE) at recent generations, at least when $f(c)=d$. When $f(c)=c$ or $c\left(1-c/2\right)$, recent $\hat{N}_T$ appears more sensitive to sample size than when $f(c)=d$. This causes recent $\hat{N}_T$ to become increasingly close to the simulated value with increasing *n*, but also to an overestimate of $\hat{N}_T$ whenever r^2 is adjusted (Figure 4-1).

Table 4-2 Estimates resulting from the non-linear least squares modelling of the simulated dataset. The estimates shown are in quadruplets arising from the different MAF thresholds used for inclusion of markers. Means and standard errors (SE) are calculated from 30 replicates and 'NA' denotes where *a* was fixed. The reciprocal of *a* is the value of r^2 at the intercept of the x-axis, that is the expectation of r^2 under complete linkage (*c*=0). Parameter *b* can be interpreted as an estimate of N_e assuming a constant population size over time, except for the case where $f(c) = c\left(1 - c/2\right)$ when it estimates $N_e - ½$.

Model	Parameter Estimates (MAF threshold[I]: 0.15; 0.10; 0.05; 0.01)			
	a		***b***	
	Mean	**SE**	**Mean**	**SE**
$f(c)=c$	1.51; 1.89; 2.80; 6.69	0.01; 0.02; 0.03; 0.14	155.5; 170.0; 191.7; 251.2	2.0; 2.1; 2.5; 2.9
$f(c)=d$	1.53; 1.91; 2.85; 6.87	0.01; 0.02; 0.03; 0.15	150.7; 164.2; 184.1; 238.1	2.1; 2.1; 2.5; 3.1
$f(c)=d,\ r^2_{adj}$	1.51; 1.88; 2.77; 6.30	0.01; 0.02; 0.03; 0.13	158.7; 175.5; 203.4; 300.9	2.2; 2.3; 2.8; 3.6
$f(c)=d,\ a=1$	NA		199.3; 262.3; 448.2; 2308.8	2.6; 3.1; 8.5; 105.3
$f(c)=d,\ a=2.2$	NA		126.5; 152.4; 217.4; 645.7	1.6; 1.8; 3.2; 15.8
$f(c) = c\left(1 - c/2\right)$	1.50; 1.87; 2.77; 6.58	0.01; 0.02; 0.03; 0.14	158.4; 173.3; 196.2; 259.5	2.0; 2.1; 2.5; 2.9

[I]**SNPs with a MAF below the threshold were excluded from the analyses.**

Table 4-3 Estimates resulting from the non-linear least squares modelling of the simulated dataset. A model and sample size comparison (MAF threshold of 0.10). The reciprocal of *a* is the value of r^2 at the intercept of the x-axis, that is the expectation of r^2 under complete linkage (*c*=0). Parameter *b* can be interpreted as an estimate of N_e assuming a constant population size over time, except for the case where $f(c)=c\left(1-c/2\right)$ when it estimates N_e – ½.

		Mean parameter estimates[I]							
		a[II]				*b*[III]			
Model	***n*=**	***50***	***100***	***200***	***400***	***50***	***100***	***200***	***400 [MSE[IV]]***
f(c)=c		2.11	1.95	1.89	1.86	136.7	160.6	170.0	176.2 [719]
f(c)=d		2.13	1.97	1.91	1.88	133.5	155.8	164.2	170.0 [1056]
f(c)=d, r^2_{adj}		1.96	1.90	1.88	1.87	178.2	178.5	175.5	175.6 [758]
f(c)=d, a=1		NA				241.4	257.7	262.3	266.8 [4825]
f(c)=d, a=2.2		NA				130.9	146.4	152.4	156.7 [1988]
f(c)=d, a=2.2, r^2_{adj}		NA				167.8	165.5	161.9	161.5 [1604]
$f(c)=c\left(1-c/2\right)$		2.10	1.93	1.87	1.84	139.0	163.5	173.3	179.7 [540]
$f(c)=c\left(1-c/2\right),$ r^2_{adj}		1.91	1.86	1.84	1.82	189.2	189.2	186.1	186.2 [339]
$f(c)=c\left(1-c/2\right),$ *a=2.2,* r^2_{adj}		NA				177.3	174.9	171.3	171.0 [940]

[I]Mean across 30 replicates.

[II]SE: 0.02.

[III]SE: 1.8-3.0, with the exception of when *a*=1, where SE: 3.1-3.9.

[IV]Predicted mean square error for when sample size was equal to 400, calculated as $\frac{\sum_{i=1}^{30}\left(\hat{N}_e-200\right)^2}{30}$ **, where *i* represents replicate number.**

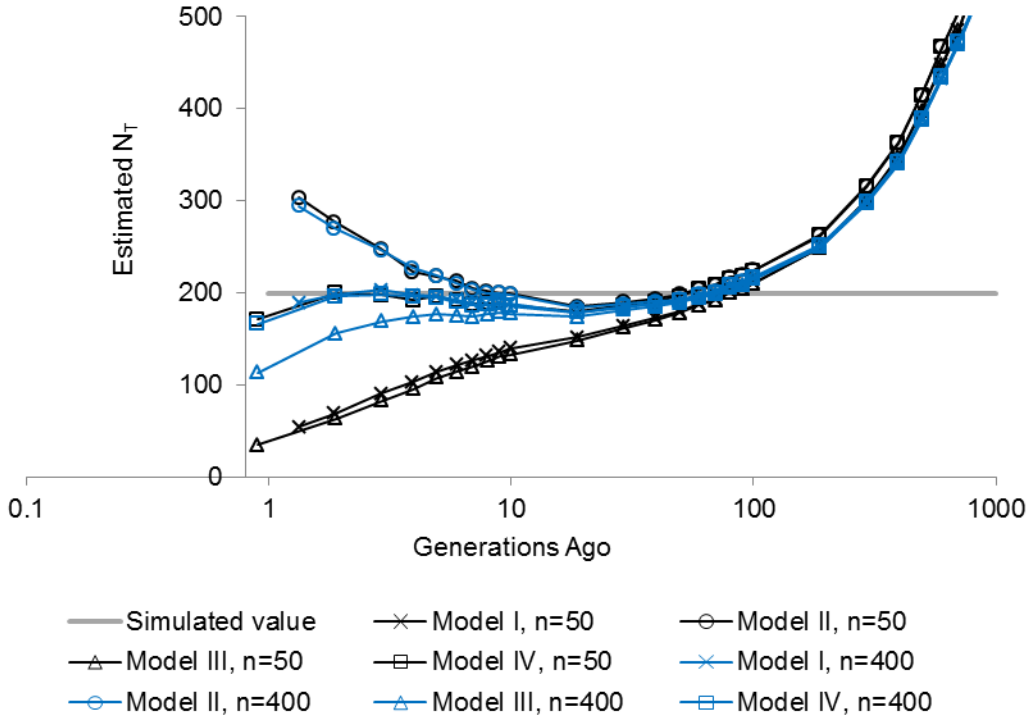

Figure 4-1 Simulated data: Average $\hat{N}_T$ (truncated at N_T=500) plotted against average generations in the past (on a logarithmic scale) truncated at 1,000. A comparison of sample size using models I, II, III and IV (MAF threshold of 0.10) (see Table 4-1 for specification of models).

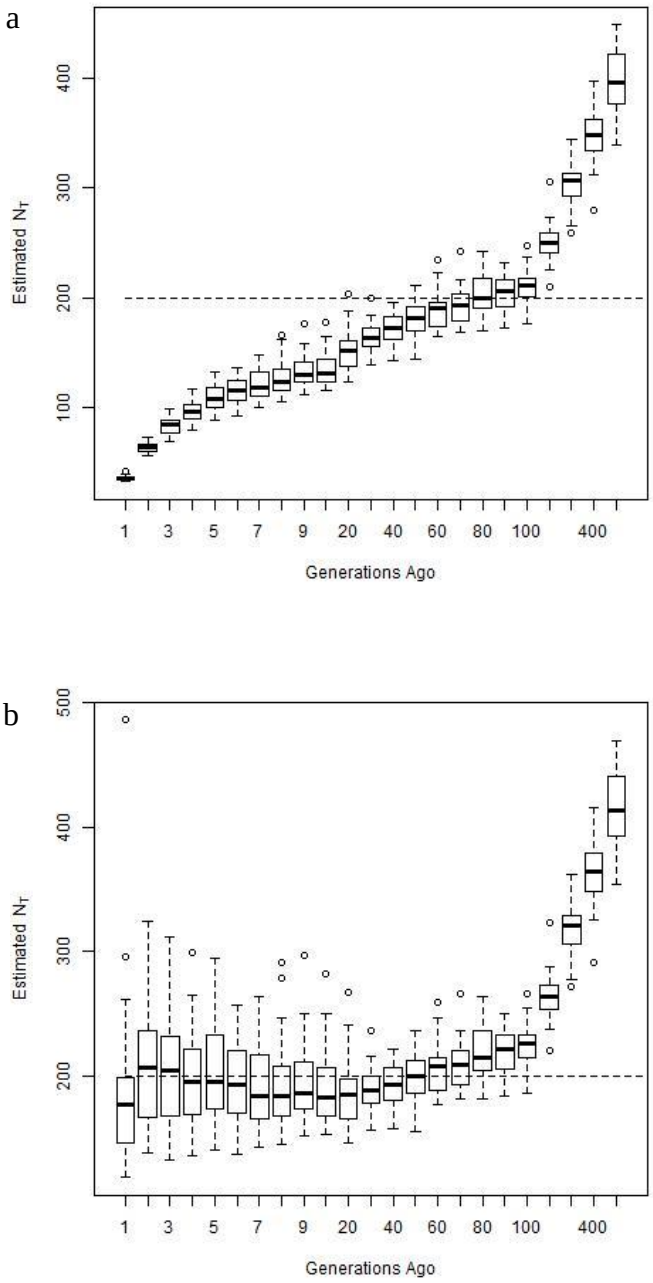

Figure 4-2 Simulated data: Boxplots of $\hat{N}_T$ plotted against generations in the past (on a non-linear scale), truncated at 500 generations (*f*(*c*)=*d*, MAF threshold of 0.10, *n*=50). Variation at each time point reflects variability across estimates within generation bins and across the 30 replicates. The dashed line represents the simulated N_e.

a) No adjustment of r^2 for sampling effects (Model III)

b) Adjustment of r^2 for sampling effects (Model IV)

4.3.1.3 Impact of form of *f*(*c*) and value of *a*

Restricting attention in Table 4-3 to the models with adjusted r^2, estimating a in the non-linear regression model resulted in the smallest MSE for N_e, and the use of $f(c) = c\left(1 - c/2\right)$ had consistently smaller MSE than $f(c)=d$. Altering the definition of $f(c)$ in the constant N_e model had relatively little impact on parameter estimates of a and b. When assuming a variable N_e, altering the value of a had the greatest impact on distant $\hat{N}_T$ which were subject to a continually increasing upward bias as t increased when a=1. This trend was reversed by setting a=2.2 (Figure 4-3), with, ultimately, negative values being obtained. The source of this discrepancy can be seen in Figure 4-5, which shows that neither model fitted the observed data convincingly. No value of a gave the most accurate $\hat{N}_T$, as judged by MSE, across all generations, but a=1 had the lowest MSE over generations 10 to 200. The form of $f(c)$ had little effect, except at the most recent generations (Figure 4-4).

4.3.1.4 Summary of models fitted to simulated data

Unless the finite sample size is accounted for the methods underestimate N_e, and when N_e is assumed to be variable, the underestimation is most severe for estimates of recent N_T. Changing the value of a to reflect mutation is also important as assuming a=1 results in overestimates of N_e which, in the case of the variable N_e model, are particularly severe for large t. When assuming a variable N_e, setting a=2.2 can lead to negative $\hat{N}_T$ at distant generations. When assuming a constant N_e, the problems with a can be avoided by estimating a. For constant N_e models, the lowest MSE was achieved using $f(c) = c\left(1 - c/2\right)$, whereas in contrast, for variable N_e, the combination which gave $\hat{N}_T$ closest to the simulation data, i.e. a growth rate of zero with N_e=200, across the widest range of generations and sample sizes, was $f(c)=d$ and a=2.2 with r^2 adjusted for sampling effects (Model V). Although this equation gave the most accurate $\hat{N}_T$, estimates were still below the true value (200) across all time points and for all sample sizes. Estimates of N_T were very sensitive to the model form for generations <7.

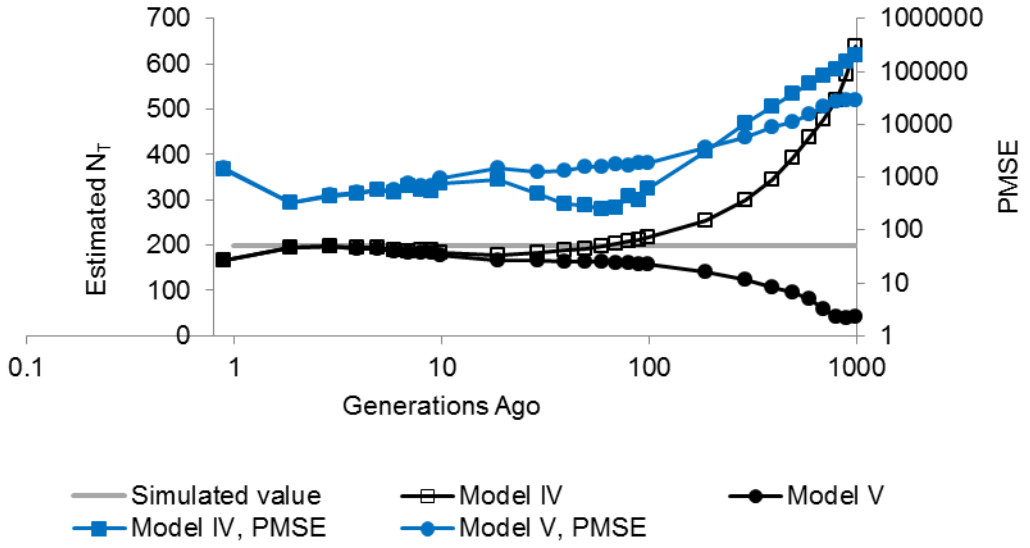

Figure 4-3 Simulated data: Average $\hat{N}_T$ plotted against average generations in the past (on a logarithmic scale) truncated at 1,000, with corresponding predicted mean square error (MSE). A comparison of models IV and V (MAF threshold of 0.10 and *n*=200) (see Table 4-1 for specification of models)

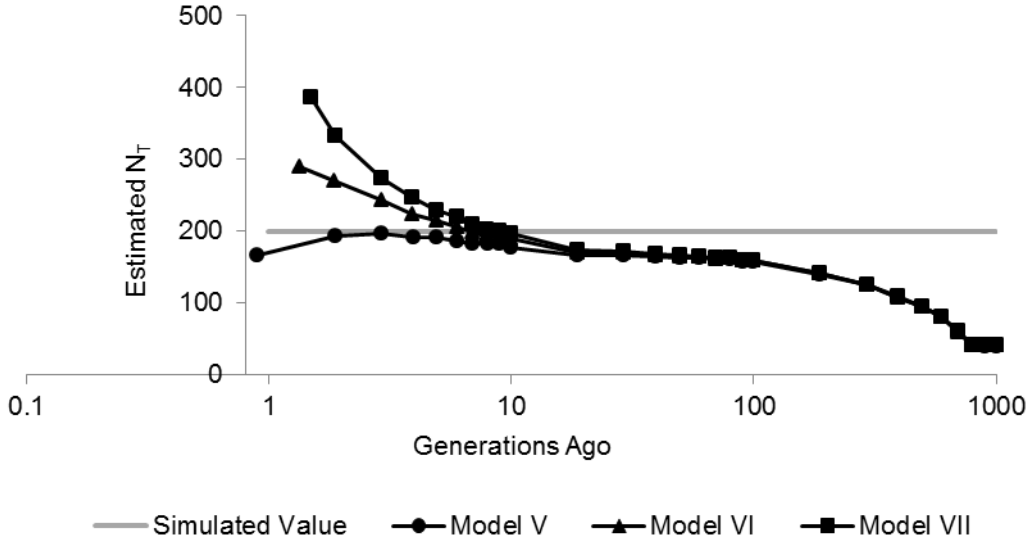

Figure 4-4 Simulated data: Average $\hat{N}_T$ plotted against average generations in the past (on a logarithmic scale), truncated at 1,000 generations. A comparison of models V, VI and VII (MAF threshold of 0.10 and *n*=200) (see Table 4-1 for specification of models).

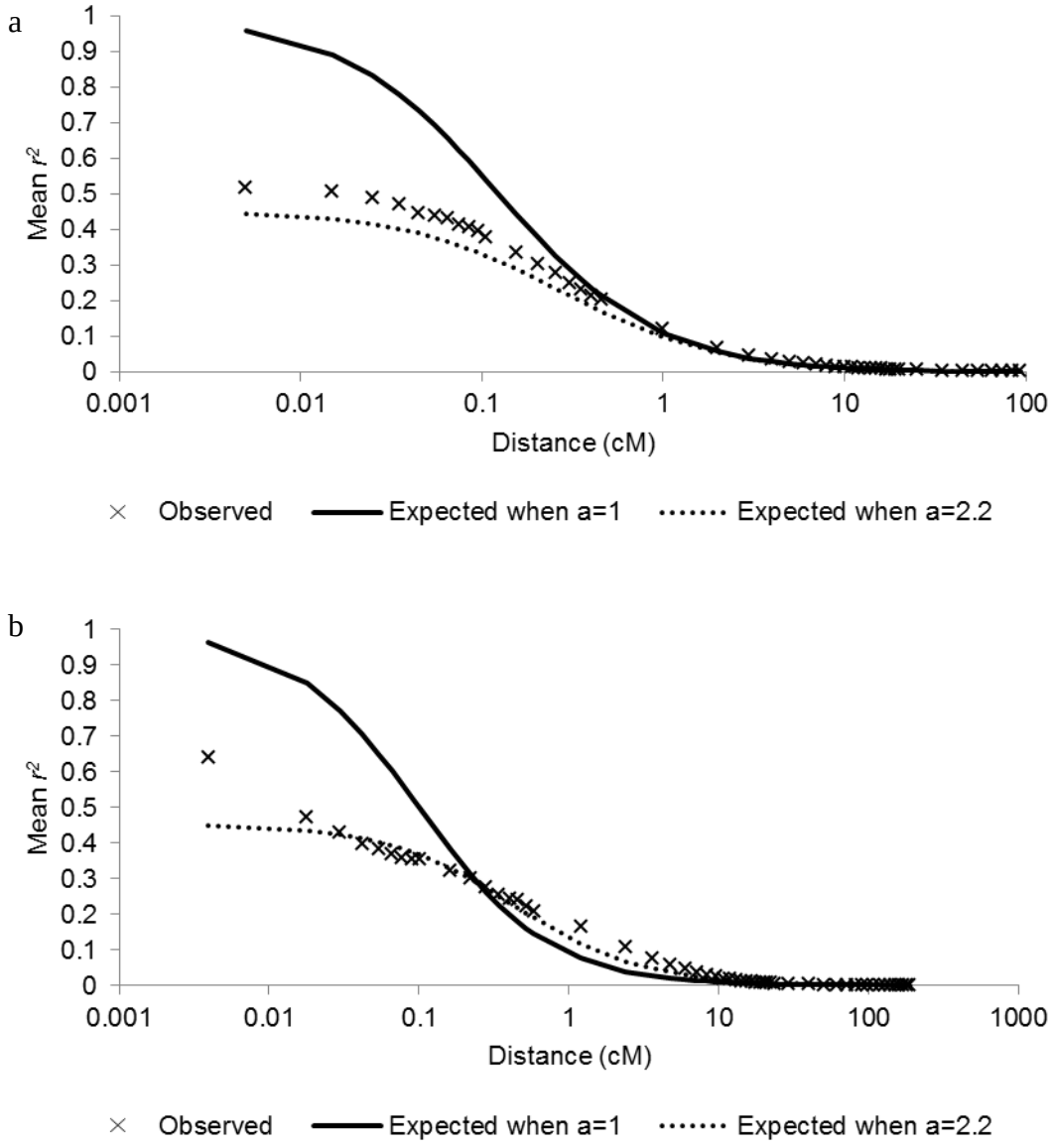

Figure 4-5 Predicted and observed r^2 plotted against mean marker distance (on a logarithmic scale).

a) Simulated data: Predicted r^2 was calculated using the true value of *b* (200) with *a* set to 1 or 2.2. Observed r^2 was calculated in the *n*=200 sample, with a MAF threshold of 0.10.

b) Equine data: Predicted r^2 was calculated using parameter estimates of *b* from models with *a* set to 1 or 2.2 and with *f*(*c*)=*d* (MAF threshold 0.10). For parameter estimates see Table 4-4.

4.3.2 Equine data

As shown in Table 4-4, the effect of changing the model underlying the constant N_e approach had relatively little impact on the parameter estimates, except for the case where a=1. Estimates for parameter a were in the range 2.17 to 2.25, whilst estimates of b ($\hat{N}_e$) ranged from 130 to 250. Figure 4-5 shows the fit of the non-linear regression model by comparison of the observed r^2 and the expected value based on estimated model parameters. Fixing a=1 as in Equation 4-1 provided a poor fit to the data, with observed r^2 being much less than expected at short distances and greater than expected at long distances. Fixing a=2.2 as in Equation 4-2 improved the fit to observed r^2 but led to a slight underestimation of r^2 at small distances. The pattern of ranking between the observed and the two models indicate that changing a alone will not provide a satisfactory fit. Using the best model from the simulated data ($f(c) = c\left(1 - c/2\right)$, r^2_{adj} and estimated a) resulted in $\hat{N}_e = 142$.

Analysis under the variable N_e scenario showed a similar pattern, irrespective of the estimation method, up to about 30 generations ago. The trend was for a decrease in $\hat{N}_T$ from a maximum (which varied with estimation method) at generations 1-2, to a minimum of $\hat{N}_T \approx 100$ at generation 20-30. This was followed by a continual increase in $\hat{N}_T$ with increasing generations into the past, reaching values of over 2,000 in the distant past, except where mutation was included in the model (a=2.2) in which case $\hat{N}_T$ declined after 1,000 generations, eventually becoming negative (Figure 4-6). The impact of alterations to the equation for N_T were consistent with those seen using the simulated data. The negative $\hat{N}_T$ indicate a problem with the model such that 2.2 is not the best estimate for a, i.e. Ohta & Kimura (1969) [158, 159] does not apply, and to address the theoretical issues discussed below, I used the average r^2 value for marker pairs less than 0.01Mb apart (r^2=0.64) to provide an empirical value of a (when c=0, $E\left[r^2\right] = 1/a$) which was calculated as 1.56, and replaced 2.2 in Model V (Figure 4-6).

Table 4-4 Estimates resulting from the non-linear least squares modelling of the equine dataset.A model comparison (MAF threshold of 0.10). The reciprocal of *a* is the value of r^2 at the intercept of the x-axis, that is the expectation of r^2 under complete linkage (*c*=0). Parameter *b* can be interpreted as an estimate of N_e assuming a constant population size over time, except for the case where $f(c)=c\left(1-c/2\right)$ when it estimates N_e – ½.

Model	Parameter Estimates			
	a		*b*	
	Mean[I]	95% C.I.	Mean[I]	95% C.I.
$f(c)=c$	2.20	[2.12,2.28]	134.9	[126.7,143.1]
$f(c)=d$	2.25	[2.18,2.33]	127.7	[119.6,135.9]
$f(c)=d,\ r^2_{adj}$	2.25	[2.17,2.32]	129.8	[121.5,138.1]
$f(c)=d,\ a=1$	NA		249.7	[230.1,269.4]
$f(c)=d,\ a=2.2$	NA		129.7	[120.6,138.9]
$f(c)=d,\ a=2.2,\ r^2_{adj}$	NA		131.5	[122.2,140.8]
$f(c)=c\left(1-c/2\right)$	2.17	[2.10,2.25]	138.7	[130.5,147.0]
$f(c)=c\left(1-c/2\right),\ r^2_{adj}$	2.17	[2.09,2.24]	141.0	[132.6,149.4]
$f(c)=c\left(1-c/2\right),\ a=2.2,\ r^2_{adj}$	NA		140.4	[131.1,149.6]

[I]**Mean across autosomes, as calculated by meta-analysis.**

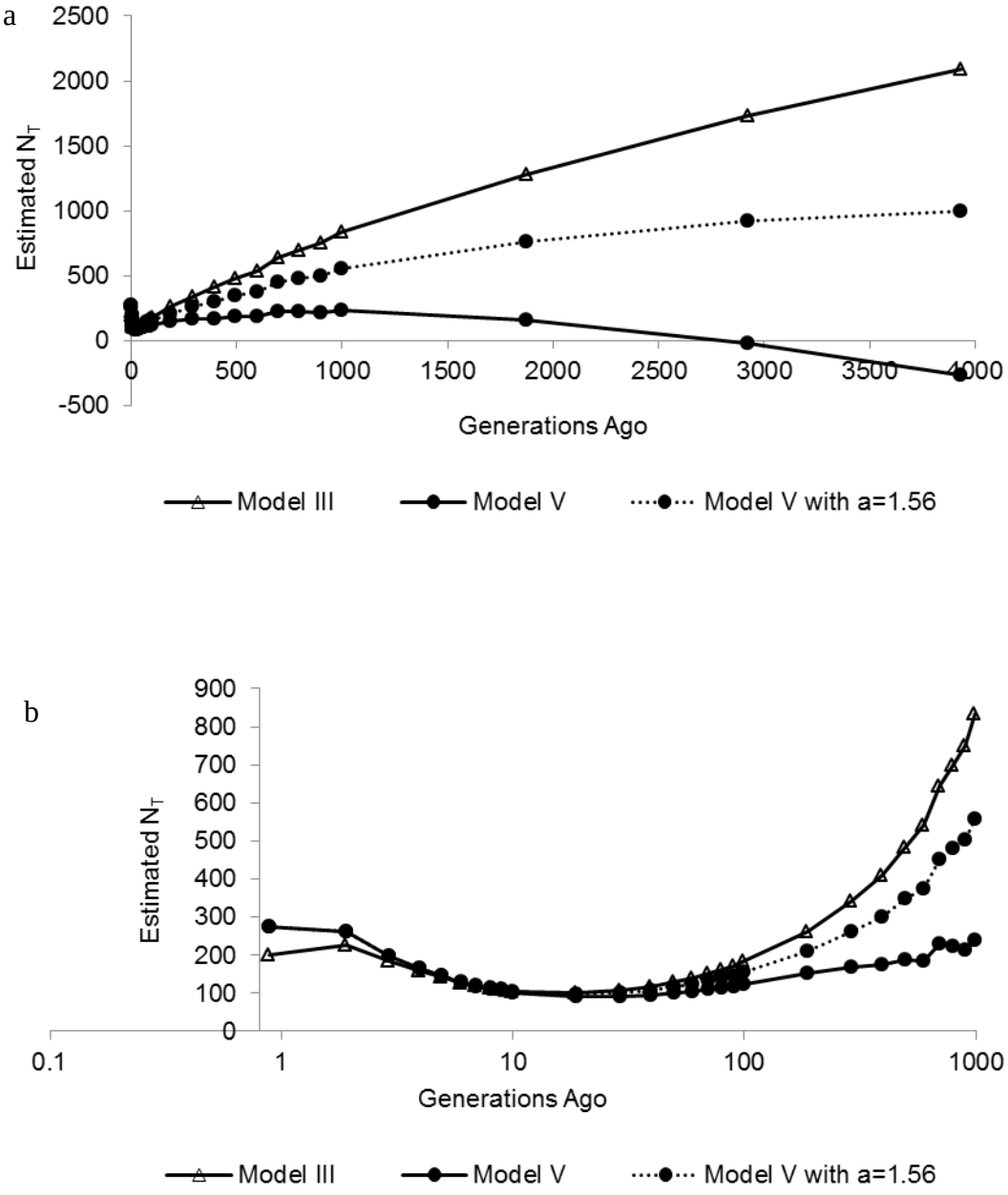

Figure 4-6 Equine data: Average $\hat{N}_T$ plotted against average generations in the past. A comparison of models III, V and V with *a*=1.56 (MAF threshold of 0.10) (see Table 4-1 for specification of models).

a) Generations in the past truncated at 4,000 generations

b) Generations in the past (on a logarithmic scale) truncated at 1,000 generations and $\hat{N}_T$ truncated at 500

4.4 Discussion

In this study the impact of a range of interpretations of Sved's (1971) [137] formula on estimated N_e have been investigated using simulated data and SNP50 genotype data of 817 UK Thoroughbred horses. The most accurate estimates of N_e, assuming it a constant, arose from adjusting the values of r^2 for finite sample size (assumed here as $(-2n)^{-1}$) and with a estimated from the data. Although the average estimates themselves were relatively robust to the form of $f(c)$ used, $f(c) = c\left(1 - c/2\right)$ yielded superior results and an increase in accuracy, as assessed by MSE. With a MAF threshold of between 0.05 and 0.10, the estimated value of a was closer to the theoretical value of 2.2 [160] than 1 [137], however, there was considerable sensitivity to the threshold used. When fitting models assuming variable N_e, the most robust models accounted for finite sample size and generally arose from using values of a that accounted for mutation, although between generations 10 to 200, values from using a=1 had lower MSE. Ignoring the effect of n had increasing effect as historical time decreased, whilst ignoring mutation had increasing effect as historical time increased. Similarly to the constant N_e model, $\hat{N}_T$ for middle and distant historical time were robust to the form of $f(c)$, however, the most recent $\hat{N}_T$ were increasingly sensitive to $f(c)$.

For both the constant and dynamic N_e models, the value of a^{-1} estimates $\lim_{c \to 0} E\left[r^2\right]$ for two SNPs within the set of markers, and critically the value obtained varied with the cut-off applied to MAF for the markers. It is plausible that this sensitivity will also be displayed with other criteria used to select the SNPs. With MAF≥0.10, estimates of a from the equine dataset appeared close to 2.2, as derived by Ohta & Kimura (1971) [160], and in the simulated dataset, close to the value of 2 chosen empirically by Tenesa *et al.*(2007) [139]. As the MAF threshold increases, variation in allele frequency is reduced and constraints on E[r^2] due to the impact of frequency differences on r^2 are eased [171], so increasing E[r^2] and so decreasing a. The impact of different MAF thresholds using actual SNP data from existing chips would likely be less, since the distribution of MAF is already biased upwards, resulting in proportionately less exclusions at a given MAF threshold. The practice of excluding

loci with extreme allele frequencies results in the preferential exclusion of the most recent mutations [172], and this is another form of ascertainment bias. Obtaining ~2.2 for the equine data (with MAF≥0.10) cannot be viewed as validation of theory since an historical bottleneck in the Thoroughbred is well documented [125, 150]. Critically, there is no clear asymptote such that by choosing particularly high MAF thresholds, an asymptotic value for *a* may be anticipated. More markers with lower MAF gave much greater values of *a* than expected from neutral theory, therefore the deviation from the theory and its associated assumptions will not be resolved by obtaining sequence data. Such a hypothesis can soon be tested, although inclusion of such loci will require large datasets. Using coalescence, McVean (2002) [168] obtains only approximate correspondence between $E\left[r^2\right] \approx \sigma_d^2$ when c=0 and with a MAF threshold of 0.10. Therefore, whilst Hudson (1985) [143] previously observed that models for $\lim_{c \to 0} E\left[r^2\right]$ with mutation [160] are sensitive to MAF thresholds, my results suggest ensuring MAF≥0.10 is too simplistic a solution for reconciling results with mutation models based on Sved (1971) [137]. Further, any change in the value of *a* or imposing an *ad hoc* MAF threshold to improve correspondence as *c* tends to 0, must also ensure that the equilibrium expression (Equation 4-2) is not compromised across the wider range. Table 4-2 shows that, of the MAF frequencies tested, whilst *a* was closest to theory when MAF≥0.10, N_e was closest to its true value when MAF≥0.05.

The original development of the theory for variable N_e (Hayes *et al.*(2003) [138] in Equation 4-4) was based on chromosomal segment homozygosity (CSH) and this was extended by Tenesa *et al.* (2007) [139] to cover E[r^2] since they have the same limiting forms, assuming constant N_e and no mutation [138, 139]. The acceptance of Equation 4-4 depends on a number of conceptual approximations. The first is that changes in N_e are linear with time, since this provides the theoretical underpinning of the mapping from time (measured in prior generations) to recombination fraction. The second is that the incorporation of mutation will alter the limiting forms of both CSH and E[r^2], and to similar degrees; whilst the former is based upon contiguous runs, the latter is a pairwise measure. Finally, is an assumption in implementations

including in Chapter 3, that c can be substituted by d or, more generally, $t=(2c)^{-1}$ by $t=(2f(c))^{-1}$.

When N_e is assumed constant there is no requirement to have a pre-determined value for $\lim_{c\to 0} E[r^2]$ since it can be estimated, and this provided the best results. However, the theory of variable N_e, which does require a value of a, cannot be disassociated from constant N_e since, at its heart, the development of the former relies on properties of the latter. This use of a pre-determined value a is problematical as there appears to be no uniformly best model over time. For example, the use of a = 1 fitted well from generations 10 to 100 but deviated more and more dramatically from the true N_e as historical time increased, whilst negative values were eventually obtained with a=2.2 (or 2). It is possible to use some external information upon which to calibrate the model; for example an estimate of recent N_e from demographic data. However, this option places an emphasis for calibration upon pairs of loci with large c, beyond the point where approximations in the derivations can be expected to hold. Alternatively, the data associated with very small values of c may be used to provide the best estimate of $\lim_{c\to 0} E[r^2]$, and hence a, as demonstrated in Figure 4-6. In principle, subject to sampling error, such an approach should avoid the appearance of negative estimates of N_e. Sacrificing N_e estimates beyond ~1,000-2,000 generations ago, corresponding to c<0.0005 – 0.00025, is unlikely to be a serious issue in the case of livestock or companion animals, as it is almost impossible to assign such estimates to a single specific population in any case. However, this approach distances procedure from theory.

The principle of sacrificing extreme historical N_e is further justified since the period over which N_e estimates are consistent with true values may be N_e dependent. For example, coalescence theory predicts the expected time for coalescence of a sample of gametes is less than $4N_e$ generations. Clearly a mutation cannot be older than the

coalescence time of the tree, and simulation of coalescent trees[7] (not shown) demonstrates that the probability of mutations older than $2N_e$ generations will be very small, and the probability of a pair of randomly sampled mutations older than $2N_e$ will be even smaller. Therefore, estimates of N_e more than ~$4N_e$ generations into the past must be questionable, corresponding to 800 generations in the simulations performed here. This horizon will be population dependent so, with simulations of human populations (N_e=10,000), McEvoy *et al.* (2011) [173] calculated representative N_e estimates for 200 to 5,000 generations in the past.

When fitting the model assuming constant N_e, the lowest mean square error was found with $f(c) = c\left(1 - c/2\right)$, not with $f(c)=d$, although d was found to be more robust with the variable N_e model. The former is a concave function of c whereas the latter is convex, thus the common practice of $f(c)=d$ is a modification to Sved (1971) [137] that moves contrary to theory and observation. This lack of consistency in results undermines the coherence of the variable N_e approach since it is derived from the constant model. An additional potential source of error when using $f(c)=d$ is the widely used assumption of a genome-wide linear relationship between genetic and physical distance; this has been shown to have an impact on more distant N_e estimates [131, 163].

More generally, the number of individuals required to obtain reliable estimates of r^2 and therefore of N_e, is a key issue. In most cases, increasing the sample size caused an increase in the accuracy of N_e estimates, with N_e being underestimated when the sample size was low. In the variable N_e case, sampling effects were most serious at recent generations, demonstrating the relative importance of sampling correction when estimating r^2 at relatively unlinked loci. Adjustment of observed values of r^2 to take account of sampling effects, led to constant N_e estimates being largely protected against the effects of n, although N_e estimates remained below the true N_e, supporting

[7] Simulations by John Woolliams (The Roslin Institute).

suggestions that the adjustment is inadequate [174]. This trend was also observed in the variable N_e case when $f(c)=d$. Somewhat unexpected was the failure of a sample size equal to or twice the true N_e to overcome the downward bias under the constant N_e method. This result is particularly surprising in light of the work of Waples (2006) [166], which showed that when using estimates of non-syntenic LD to calculate so-called 'contemporary' N_e estimates, sample sizes in excess of the true N_e led to its overestimation.

Despite the apparent inability of large sample sizes to completely ameliorate the downward bias in N_e estimates, increasing sample size did improve precision. The increase in precision with increasing sample size is not only intuitive but is predicted by Hill (1981) [141] in a formula for the coefficient of variation of $\hat{N}_e$. This equation also predicts a decrease in precision of $\hat{N}_T$ at recent generations, again leading myself and others [141] to question N_e estimates for the most recent generations. Whilst the massive increase in markers now available should theoretically improve accuracy across all time points, patterns of between-replicate variation do largely reflect concerns about recent $\hat{N}_T$, since estimates of N_e based on small sample sizes suffer from considerable between-replicate variation. The pooling of marker pairs across chromosomes is an important step in reducing the impact of this variation on final estimates. It is likely that the method of binning SNP pairs will also affect the relative precision of estimates at different time points. Here, marker pairs are binned by generation and as such, the number of marker pairs per bin and the variation in r^2 will be different from studies where marker pairs have been put in equally sized (by distance) bins. The optimal binning process is far from obvious and, although not examined here, may benefit from standardisation.

Whilst the constant N_e method has the potential to offer clues about a population's historical N_e, the frequent violation of the assumption of constant N_e in real populations limits its use as a predictor. Nevertheless, this study has demonstrated, through multiple comparisons, the potential impact of alternative expressions for $E[r^2]$ on estimates of N_e. I have also explored the effect of MAF thresholds and sample size on predicted N_e. In the case of the constant N_e method, a MAF threshold

between 0.05 and 0.10 seems preferable to impose on data prior to analysis. Furthermore, estimates of r^2 should be adjusted for sampling effects and $f(c)$ set equal to $c\left(1-c/2\right)$ with a estimated.

With respect to the variable N_e method implemented here, there is empirical evidence from the literature of approximate agreement between molecular and pedigree based estimates: I have previously compared my results to those from typical pedigree inbreeding coefficients (Chapter 3) and Uimari *et al.* (2011) [175] to actual pedigree; furthermore, analysis of human sequence data [176] confirmed results of Tenesa *et al.* (2007) [139]. However, the issues raised here cast doubt on the estimation of variable N_e, particularly for recent generations (T<10), even when corrected for finite sampling. Problems with T<10 should be anticipated since they rely on large c, whilst the theoretical development by Hayes *et al.* (2003) [138] relating CSH to N_e make use of approximations that become reasonable as c decreases towards 0.05 and T increases to 10. Further, the methodology is highly sensitive to the value of a, and in turn any MAF threshold used for screening, which will have considerable impact on distant past. To some extent this sensitivity can be overcome since the estimation of N_e can only be expected to be valid for $<4N_e$ generations, and this gives the opportunity for an estimate of a to be obtained. An approximation to the true shape over an intermediate period may be obtained, for example, preserving the local maxima or minima (although the assumption of linear N_e with time implies there are no such change points) and it appears that over middle ranges of historical time the methodology can give estimates of N_e that are within a factor of two; this in itself can be informative, for example, when comparing populations. Nevertheless, the theoretical basis for models of variable N_e based on $E[r^2]$ is unclear and has not been soundly established, especially in the form currently applied, e.g. $f(c)=d$, making outcomes unreliable. Further, this conclusion is based on simulations where N_e was held constant with random selection, random mating and discrete generations, and will strengthen with more complex population structure. Whilst the theoretical development here does not encompass such structure, the results of Toosi *et al.* (2010) [134] give some indication of what can be expected when Equation 4-1 is

used to estimate constant N_e in populations with a history of crossbreeding and admixture.

In summary, decision-making based on estimates of N_e calculated by the methods presented here, for example to inform conservation, should only be done with extreme caution. Alternative methodologies to predict N_e from SNP marker data have been developed and one of these is investigated in the next chapter of this thesis.

Chapter 5: The estimation of historical effective population size using a composite-likelihood methodology

5.1 Introduction

Linkage disequilibrium (LD) describes the non-random association of alleles at different loci and represents an important feature of genomes from many perspectives. The concept of using patterns of LD to provide insights into the evolutionary history of populations has already been introduced in Chapters 3 and 4. In these chapters, estimates of effective population size (N_e) were derived using deterministic equations based on the theoretical relationship between expected r^2 (E[r^2]), distance between markers and N_e. A potential disadvantage of this approach is its reliance on what is essentially a summary statistic (r^2) as this means that it is not using all of the available information from the data. Furthermore, r^2 is influenced by allele frequencies at the loci, being limited by both the minor allele frequencies (MAF) and the absolute difference in allele frequencies at the two loci [131, 153] [177]. In particular, it is clear that the full range of r^2 (0 to 1) can only be realised when the allele frequencies are the same at both loci. In order to address these limitations, an alternative approach based on estimating likelihoods under simple neutral models has been suggested [178-181]. This methodology involves the comparison of observed sample configurations to sampling distributions generated under a simple neutral model in order to infer properties of the population, such as the recombination parameter, $\rho=4N_ec$ (where c is the recombination rate and is referred to as r in Hudson (2001) [181]). The applicability of this approach was initially limited due to the intense computation required to generate both the sample probabilities and the maximum likelihood estimates of the parameter [181]. However, advances in computing and the ability to make sampling distributions available led Hudson (2001) [181] to re-visit this empirical approach.

The so-called composite-likelihood estimator (CLE) introduced by Hudson (2001) [181] is an *ad hoc* method for estimating the population recombination rate, ρ, on the basis of combining coalescent likelihoods over all pairwise comparisons of segregating sites. This approach uses two-locus sample distributions, generated by a

random-genealogies Monte Carlo method, to produce likelihoods for the estimation of ρ (and N_e if c is known). The correlation between markers on a chromosome assayed in a single sample, caused by linkage, means that marker pairs are not independent and therefore that the resulting expression is not a true likelihood, hence it is instead referred to as a composite likelihood [181]. There have been studies that demonstrate the potential utility of this and similar approaches, in some cases applying extensions to the methodology, for example, to incorporate variable recombination rates across the genome and to include gene conversion [142, 182-184]. However, these studies have focused primarily on the analysis of markers that are relatively close together, for example, those contained within a gene. Therefore, the methodology has only been tested for a limited range of ρ, typically up to around ρ=50.

One hypothesis is that, by taking into account both allele frequencies and sample size in generating sample probabilities, the CLE approach should be subject to less error than the approach used in Chapters 3 and 4 (referred to here as the syntenic LD method) when estimating N_e. In order to test this hypothesis, data simulated by the coalescent approach has been analysed using both deterministic expressions for E[r^2] as in Chapters 3 and 4, and by the CLE approach using sampling probabilities to estimate N_e. Data was simulated in order to represent typical sample genotype data for a livestock population. Additionally, the sensitivity of the CLE approach to the distance between markers is investigated.

5.2 Materials and methods

5.2.1 Data simulation

Single-nucleotide polymorphism (SNP) marker data was simulated using the *ms* program which assumes the standard coalescent approximation to the Wright-Fisher model [185]. The size of the simulated segment was equivalent to one Morgan (1M), achieved by setting the number of sites between which recombination can occur to 1,000,001 (equivalent to the number of base pairs in the locus) and the probability of crossing-over per generation between the ends of the locus to $1x10^{-6}$ (assuming a finite-sites uniform recombination model). An infinite-sites model of mutation is

assumed with the number of segregating sites set to 1,500 which represents a similar SNP density as is achieved with the 50K SNP chip in the horse. In instances where more than one marker was given the same location (due to rounding in *ms*), only the first marker was retained in the analysis. Simulations were run assuming a constant diploid population size of N_e=200 and N_e=1,000. Thirty independent samples of 50 haplotypes (T=50) (equivalent to 25 diploid individuals) were generated in each case.

5.2.2 Data analysis

5.2.2.1 Composite-likelihood estimator (CLE)

5.2.2.1.1 Derivation of sample probabilities

Analysis of the simulated data followed Hudson (2001) [181] and was based on a table of sample probabilities for all possible sample configurations for a pair of biallelic markers given 50 sampled haplotypes and for a set of ρ values ranging from 0 to 120. This table and similar tables for a range of sample sizes are available to download from http://home.uchicago.edu/~rhudson1; these probabilities were obtained empirically by generating a large number of independent two-locus genealogies using standard coalescent machinery [143, 181, 186].

Considering a pair of biallelic polymorphic sites, there are four possible haplotypes. The sample configuration can be denoted $\mathbf{n} = (n_{00}, n_{01}, n_{10}, n_{11})$, where n_{ij} is the number of sampled gametes that carry the allele A_i at locus A and B_j at locus B. The probability of a particular configuration, $\mathbf{n} = (i,j,k,l)$ can be denoted $q(i, j, k, l; \theta, \rho)$ where $\theta = 4N_e\mu$, assumed small, and $\rho = 4N_e c$. Hudson's tables are generated assuming that the ancestral allele is known, hence the total possible number of marginal allele frequencies (see Figure 5-1) for T=50 is 1,225, i.e. (49x50)/2. However, in this case it is assumed that the ancestral allele is unknown (referred to as 'a-d unspecified' by Hudson (2001) [181]), in which case the number of marginal allele frequencies possible is reduced to 325, i.e. (25x26)/2 because of equivalencies. New sample probabilities were calculated under this assumption by summing probabilities for equivalent sample configurations; for an example, see Figure 5-1. These probabilities were then used to calculate sample configuration probabilities conditional on marginal allele frequencies observed, i.e.

Equation 5-1 $$q_c(\mathbf{n}, \rho \mid \text{marginal allele frequency}) = \frac{q(\mathbf{n};\rho)}{\sum_m q(\mathbf{m};\rho)},$$

where the summation is over all possible sample configurations, ***m***, for a given marginal allele frequency, i.e. over the range of minor allele/minor allele haplotypes possible. For the case $\rho=\infty$, i.e. independent loci, conditional probabilities were calculated according to the hypergeometric distribution, such that:

Equation 5-2 $$q_c(\mathbf{n};\rho=\infty) = \frac{(i+j)!(k+l)!(i+k)!(j+l)!}{i!\,j!k!l!T!},$$

The logarithm of an approximation to the gamma function, based on the Lanczos approximation [187], was used to calculate these probabilities [188].

For each possible sample configuration, LD measured as the squared correlation coefficient between two-locus pairs (r^2) [135] and computed as:

Equation 5-3 $$r^2 = \frac{D^2}{p_{A_0} p_{A_1} p_{B_0} p_{B_1}},$$

where, $D = p_{A_0B_0} - p_{A_0} p_{B_0}$ and p_{A_0}, p_{A_1}, p_{B_0} and p_{B_1}, are the frequencies of alleles A_0, A_1, B_0, and B_1, respectively, was calculated. Subsequently, values of r^2, conditional on marginal allele frequency, were calculated for each ρ value in the set by multiplying the calculated r^2 by the conditional probability of the sample configuration. Finally, the expected r^2 conditional on the configuration ($E_c[r^2]$), was calculated by summing the conditional r^2 across all possible configurations for each given marginal allele frequency and for each ρ value in the set.

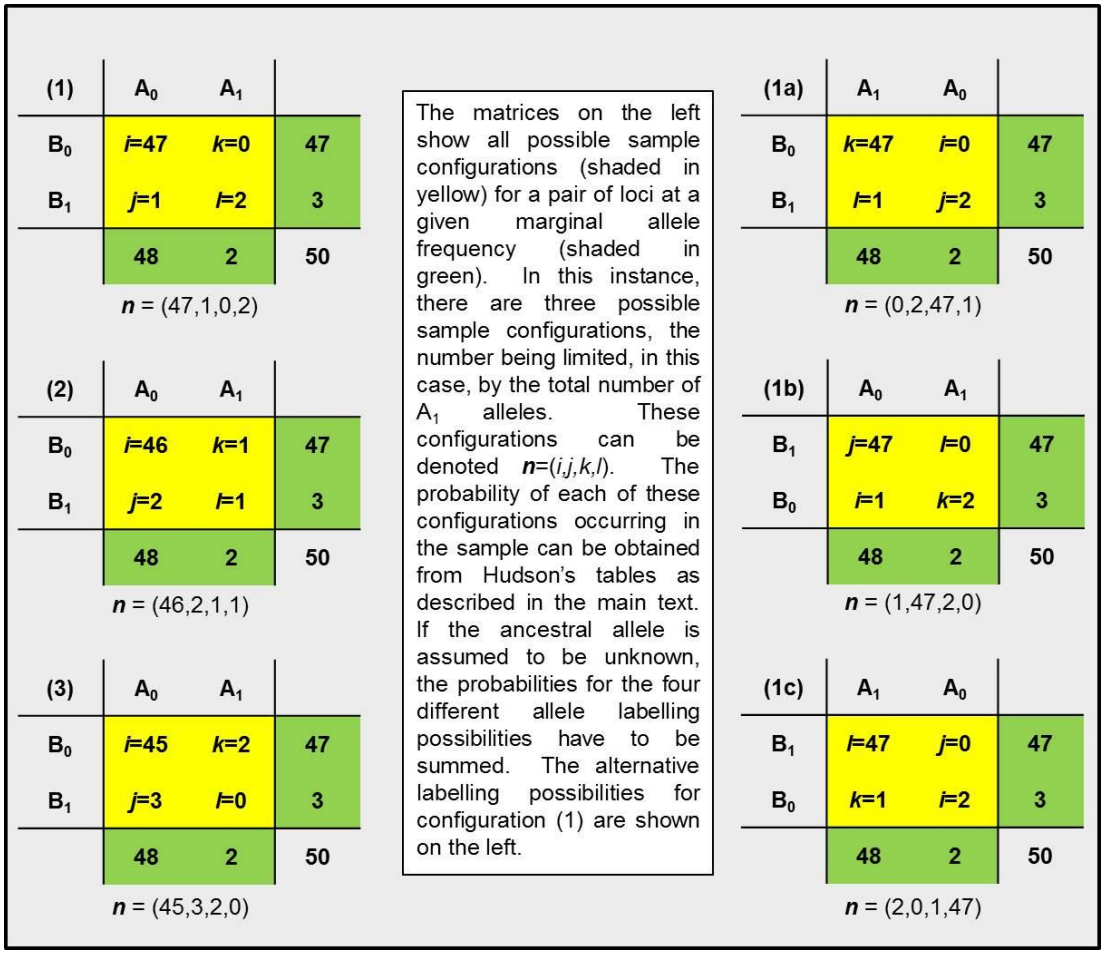

Figure 5-1 Sample configurations and marginal allele frequencies

5.2.2.1.2 Summary of linkage disequilibrium

To evaluate LD in the simulated data, r^2 was calculated for each pair of loci. First, the map distance between each pair of loci, d_i, was calculated as the product of the base pair distance and the recombination rate per base pair (10^{-6}) assuming a linear relationship such that 1Mb=1M. Pairs of loci were then pooled across the thirty replicates according to their marginal allele frequencies and the distance between loci in 0.001M intervals. The mean r^2 within each bin was then calculated. Pooling across repeats was necessary to ensure that sufficient numbers of comparisons were placed in each bin to get reliable estimates of r^2. Results are presented for pairs of loci with the following marginal allele frequencies (A_0,B_0): (5, 5), (5, 15), (5, 25), (15, 15), (15, 25) and (25, 25). The observed r^2 across the range of distances for each of the marginal allele frequencies was then compared to its expected value, $E_c[r^2]$, given different values of $\rho = 4 \times N_e \times d$.

5.2.2.1.3 Estimation of effective population size

An estimate of N_e was generated for each of the 30 replicates in turn. For a range of N_e values (from 0 to 10,000, with intervals of 10), ρ was calculated as $\rho = 4 \times N_e \times d$ and the value transformed by $1/(1+\rho)$. Transformation was necessary in order that interpolation was always done between two finite values; the set of ρ values in the table of conditional probabilities were similarly transformed. Sample configurations for each pair of loci were determined (described with reference to the minor allele and assuming the ancestral allele is unknown). The likelihood of observing the sample configuration in question, given the distance between the markers and therefore given ρ, was then calculated by linear interpolation from the table of conditional probabilities. Finally, an estimate of N_e for the entire sequence was obtained by combining the likelihoods from all pairwise comparisons:

Equation 5-4 $$L(\mathbf{n_1}, \mathbf{n_2}, ..., \mathbf{n_k}; N_e) \approx \prod_i^k q_c(\mathbf{n_i}; \rho_i)$$

A global estimate of N_e was taken as the value which had the highest composite log likelihood. This estimation procedure was then repeated for marker pairs binned by distance between them (in 0.02M intervals), with an N_e estimate generated for each bin.

5.2.2.2 Syntenic linkage disequilibrium (SLD)

5.2.2.2.1 Summary of linkage disequilibrium

Taking r^2 values previously calculated for all marker pairs, a MAF threshold of 0.05 was applied to the SNPs. Single-nucleotide polymorphism pairs were again binned according to the distance between loci (0.001M intervals), and the mean r^2 for each bin calculated. This observed r^2 was compared to its expectation under Equation 5-5 below, given N_e=200 or N_e=1,000, and a=1 or 2.

5.2.2.2.2 Estimation of effective population size

N_e estimates were derived in a deterministic manner from LD (r^2) using the approach described in Chapters 3 and 4, that is, based on the theoretical relationship between LD, distance between markers and N_e [137], such that:

Equation 5-5 $E\left[r_{adj}^2\right] = \left(a + 4N_e d\right)^{-1}$,

where, the reciprocal of *a* is the value of r_{adj}^2 at the intercept of the x-axis, *d* is the distance in Morgans between the SNPs, and $r_{adj}^2 = r^2 - \left(T\right)^{-1}$, where *T* is the number of haplotypes sampled (see Chapter 4). Estimates assuming, *a priori*, constant N_e were derived using the non-linear least squares approach to modelling described in 4.2.4.2. The statistical model took the form: $y_i = \left(a + 4bd_i\right)^{-1} + e_i$, where y_i is the value of adjusted r^2 for SNP pair *i* separated by d_i Morgans, e_i represents the residual, *a* was set to 1 or 2 or estimated iteratively alongside parameter *b* using least squares. The analyses were carried out firstly with a MAF threshold of 0.05 and then with a threshold of 0.10 applied to markers prior to analysis (see Chapter 4 for justification).

Marker pairs were then binned by distance between them (in 0.02M intervals). For each of the bins, the average r_{adj}^2 and average distance, d_i, between segregating sites were calculated. N_e was then calculated according to Equation 4-4 in Chapter 4, such that: $N_T\left(t\right) = \left(4d_t\right)^{-1}\left(E\left[r_{adj}^2 \mid d_t\right]^{-1} - a\right)$, where N_T is the effective population size *t* generations ago, $t = (2d_t)^{-1}$ [138], d_t is the distance between SNPs in Morgans and *a* was set to 1 or 2 (only results for a=2 are presented).

5.3 Results

5.3.1 Linkage disequilibrium

The observed decline of r^2 with increasing distance between markers was compared with the expectation both under Equation 5-5 and using sampling distributions for a selection of marginal allele frequencies (Figure 5-2). The curves of $E_c[r^2]$ vary with marginal allele frequencies and when marker pairs were partitioned according to their marginal allele frequencies the mean observed r^2 values approximately matched the expectation. The curves of observed r^2 were smoother when N_e=1,000, than when N_e=200. The curve of mean observed r^2 values calculated from all marker pairs simultaneously did not closely follow the curve of expectation based on Equation 5-5 with a=1. The E(r^2) based on Equation 5-5 with a=2 provided a better

fit to the observed data. This data is not shown here but resembled that shown in Figure 4-5, in Chapter 4.

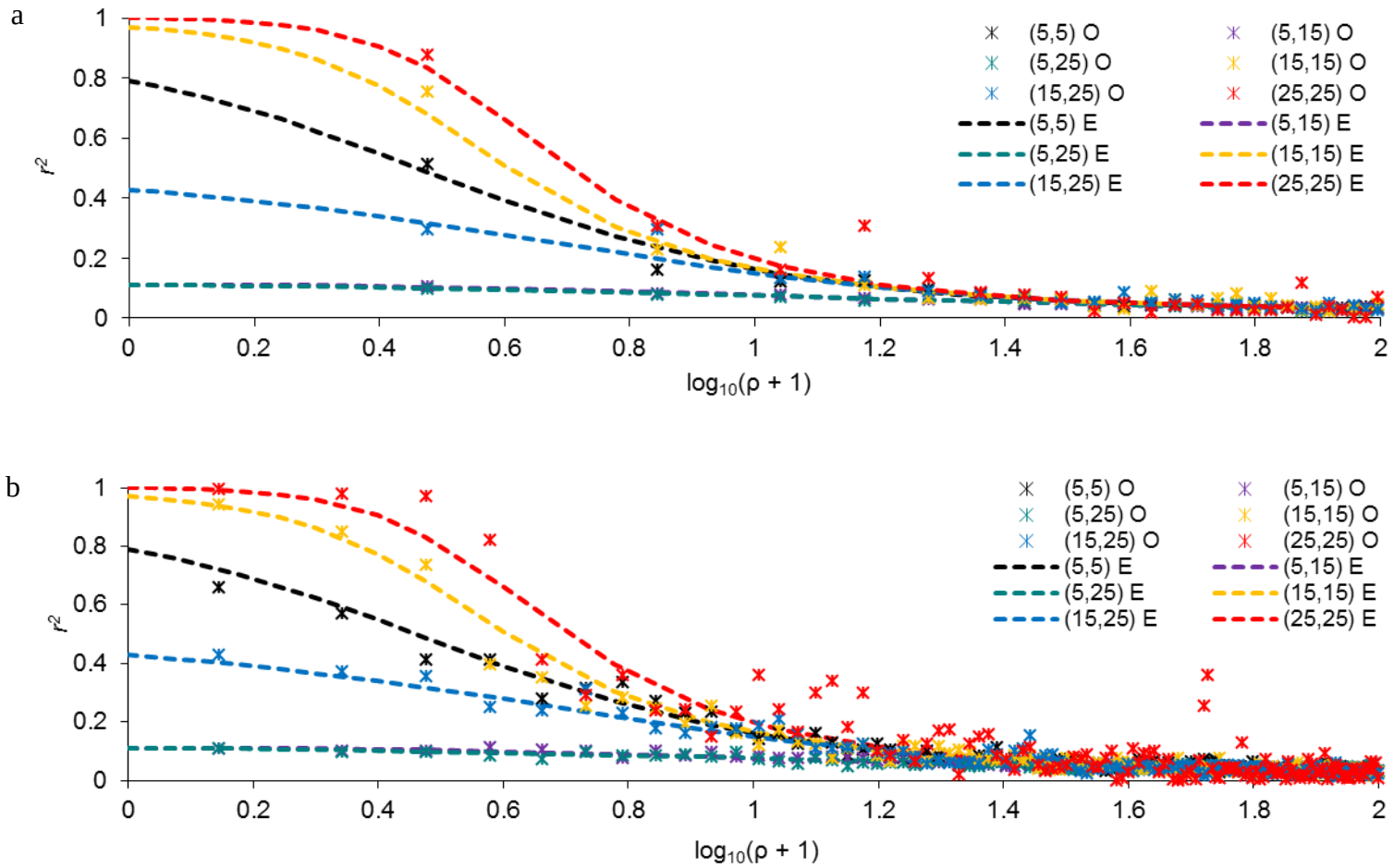

Figure 5-2 Observed r^2 (pooled across 30 replicates) (O) and $E_c[r^2]$ (E) for marker pairs with different marginal allele frequencies. a) N_e=1,000; b) N_e=200

5.3.2 Composite-likelihood estimator (CLE)

N_e estimates calculated by the CLE method are shown in Table 5-1 and Figure 5-3. The mean (±SD) N_e estimate for the case where the true N_e=1,000 was 977 (±151) and for the case where the true N_e=200 was 206 (±39). The maximum likelihood surfaces are shown in Figure 5-4 and curves were found to be distinctly asymmetric. A more peaked log likelihood curve was observed when true N_e=200 than when true N_e=1,000, indicating a narrower confidence interval.

When marker pairs were put into 0.02M bins according to the distance between them, the most accurate estimates of N_e were calculated from the closest markers. As the distance between markers increased, accuracy decreased and a positive bias was observed and this occurred at a faster rate when the true value of N_e=1,000 compared to when N_e=200 (Figure 5-5). The number of marker pairs per bin was similar across replicates but decreased linearly with increasing distance, making it a confounding factor in the relationship between accuracy of N_e estimates and the distance between marker pairs. Mean N_e estimates remained within 50% of the true value for markers up to 0.13M apart when true N_e=1,000 and for markers up to 0.23M apart when true N_e=200. The maximum N_e for which likelihoods were calculated was 10,000 and therefore, estimates of 10,000 would have been returned in all instances when the most likely N_e was ≥10,000. Using the harmonic mean eliminated the upward bias observed when using the arithmetic mean and the distance between markers was large, but did result in a slight underestimation of N_e in the case of the most distant markers (Figure 5-6).

Table 5-1 N_e estimates by CLE and SLD under various models

	N_e = 200		N_e = 1,000	
Method	**Mean [min, max]**	**MSE**	**Mean [min, max]**	**MSE**
CLE	206 [140, 330]	1530	977 [620, 1250]	22567
SLD: *a* estimated & MAF 0.05	204 [157, 282]	593	1001 [759, 1187]	7204
SLD: *a* estimated & MAF 0.10	152 [134, 236]	2287	876 [696, 991]	20534
SLD: *a* = 2 & MAF 0.05	264 [225, 339]	5045	1286 [1124, 1394]	87536
SLD: *a* = 2 & MAF 0.10	179 [147, 228]	790	905 [811, 1012]	11979

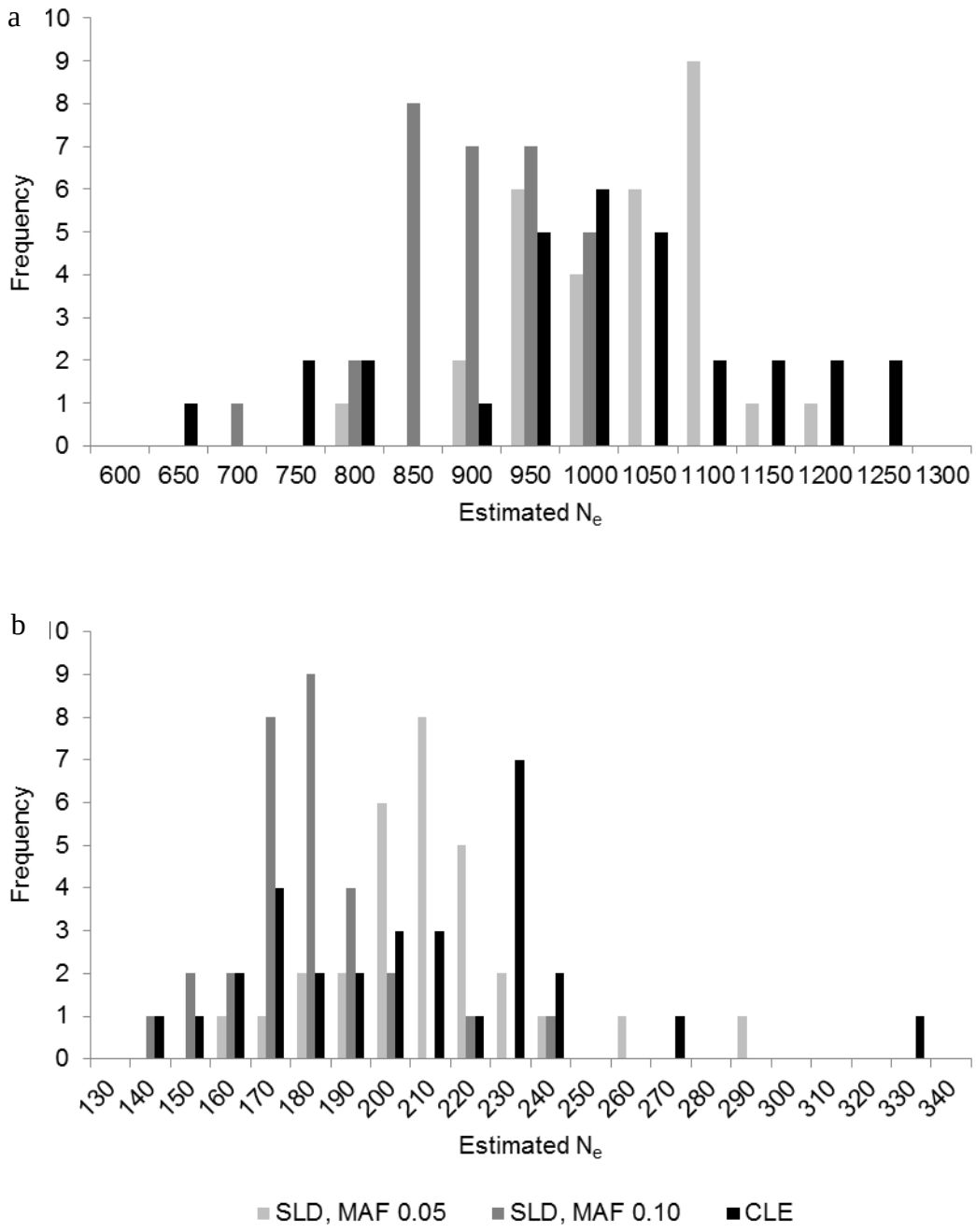

Figure 5-3 Histogram showing the distribution of N_e estimates by the CLE and SLD methods (with a=2 and with MAF thresholds 0.05 and 0.10).

a) N_e=1,000

b) N_e=200

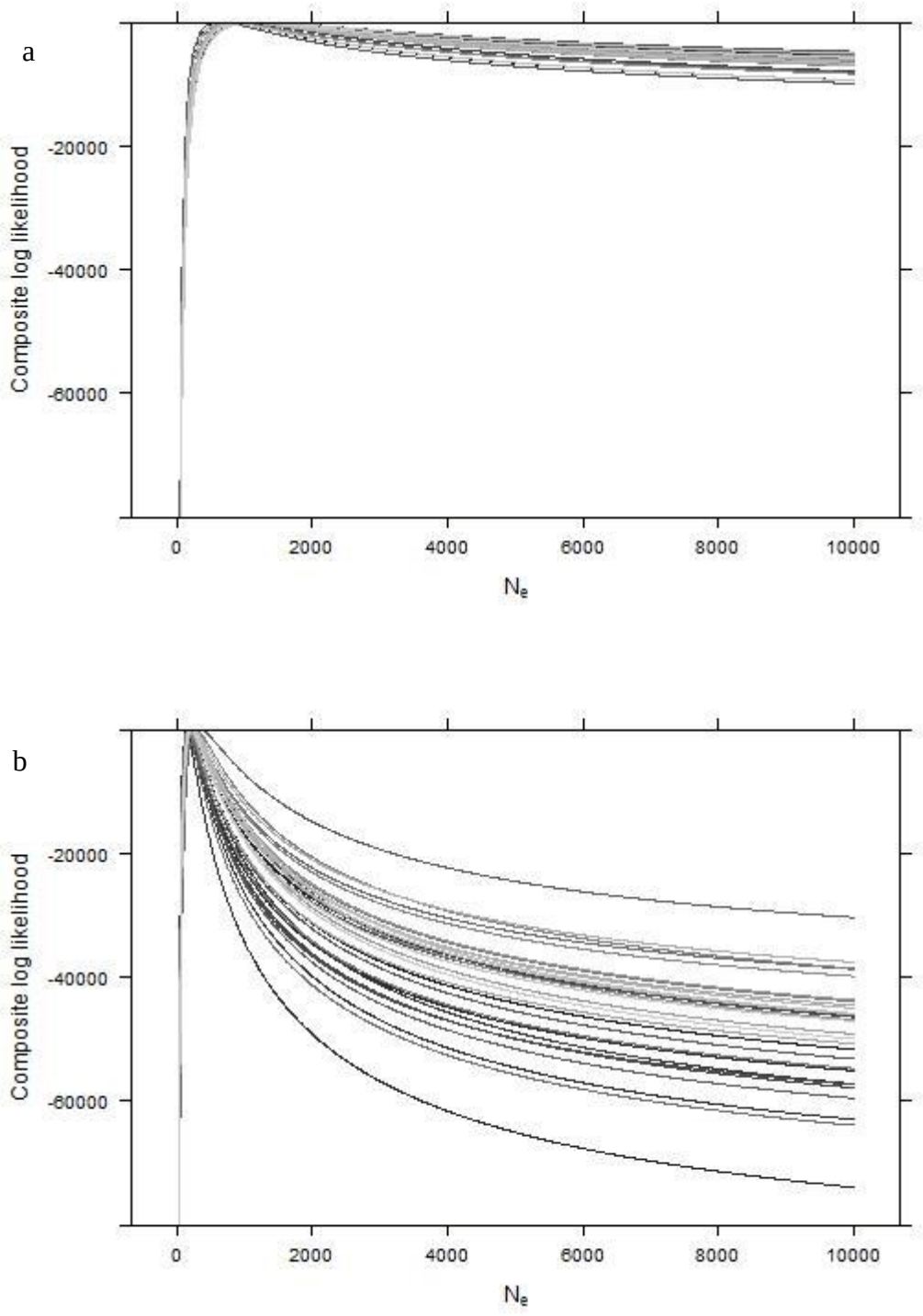

Figure 5-4 Composite log likelihood curves for replicates 1 to 30, scaled such that the maximum log likelihood for each curve is zero (y-axis truncated at -80,000).

a) N_e=1,000

b) N_e=200

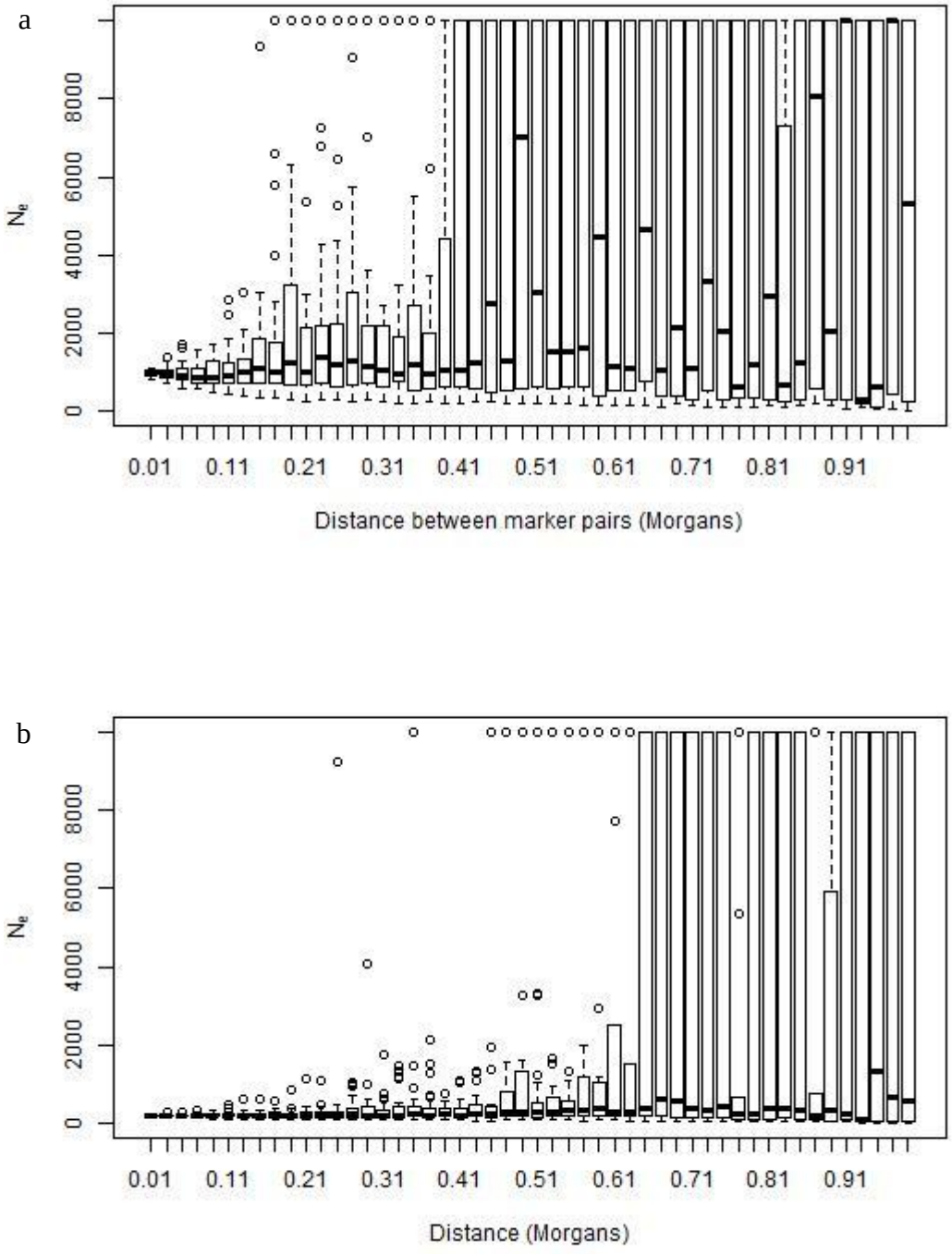

Figure 5-5 Estimates of N_e by CLE from marker pairs binned by distance. Variation at each distance reflects variability in estimates within distance bins across the 30 replicates.

a) N_e=1,000

b) N_e=200

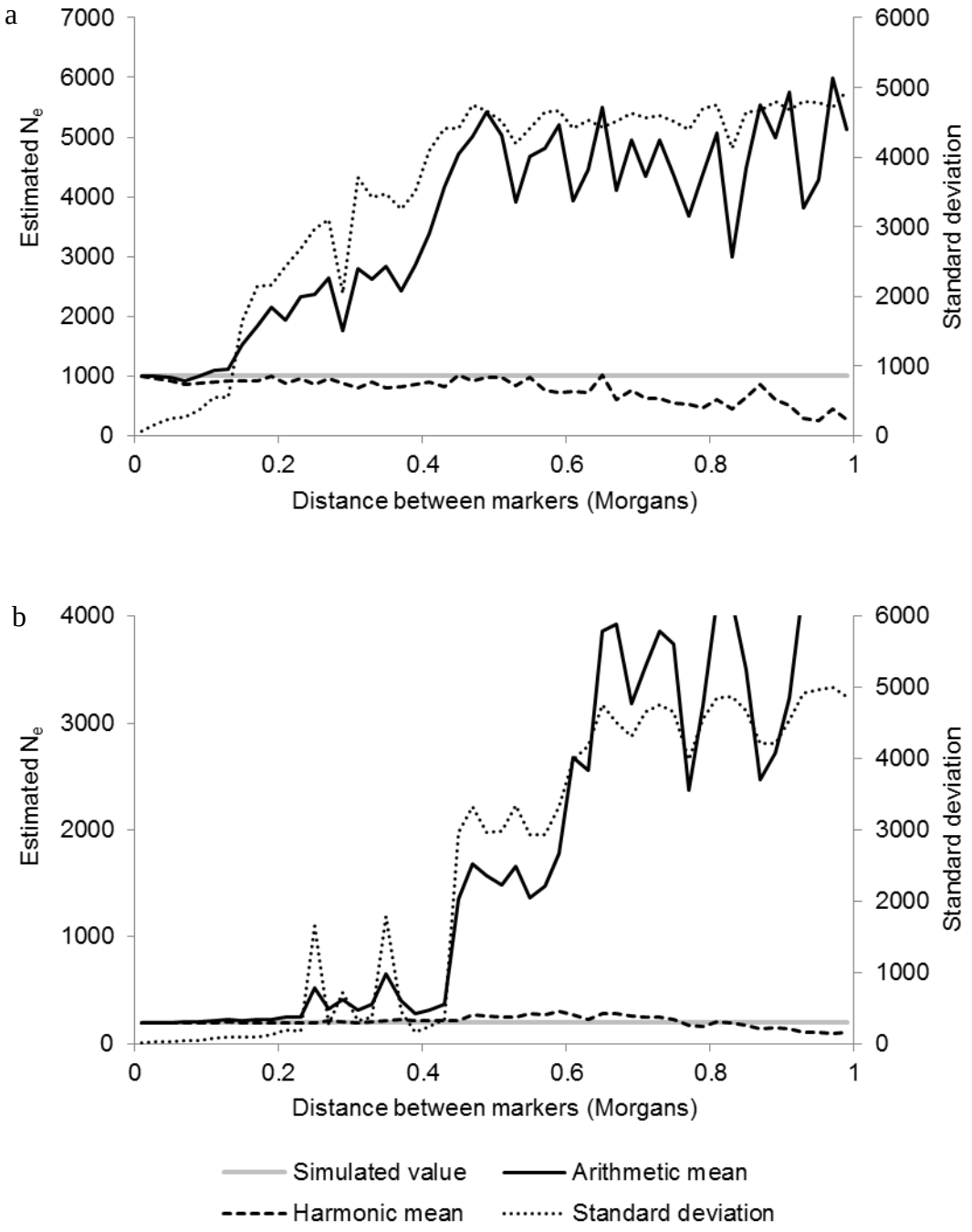

Figure 5-6 Mean and standard deviation of estimates of N_e by CLE from marker pairs binned by distance

a) N_e=1,000

b) N_e=200 (truncated at N_e=4,000)

5.3.3 Syntenic linkage disequilibrium (SLD)

N_e estimates calculated by SLD are shown in Table 5-1 and Figure 5-3. Regardless of the actual N_e, the best estimates of N_e, as judged by the smallest MSE, were obtained by estimating *a* and using a MAF threshold of 0.05. The best N_e estimate (±SD) for the case where the true N_e=1,000 was 1,001 (±86) and for the case where the true N_e=200 was 204 (±24). There was a moderate positive correlation between N_e estimates by CLE and SLD (r=0.5 for N_e=200; r=0.6 for N_e=1,000, p<0.01), as illustrated in Figure 5-7.

When marker pairs were put into 0.02M bins according to the distance between them (Figure 5-8), the most precise estimates of N_e were generally calculated from the closest markers, but for both N_e=1,000 and N_e=200, mean estimates were closest to the true N_e for markers in bin 1.4-1.6x10^5, with closer markers yielding underestimates. As the distance between markers increased, accuracy decreased with some large positive and negative estimates of N_e being observed. Mean N_e estimates remained within 50% of the true value for markers up to 0.21M apart when true N_e=1,000 and for markers up to 0.25M apart when true N_e=200. By replacing the negative N_e estimates with estimates of infinity and averaging across replicates using the harmonic mean, more consistent N_e estimates were generated over a longer distance (Figure 5-9).

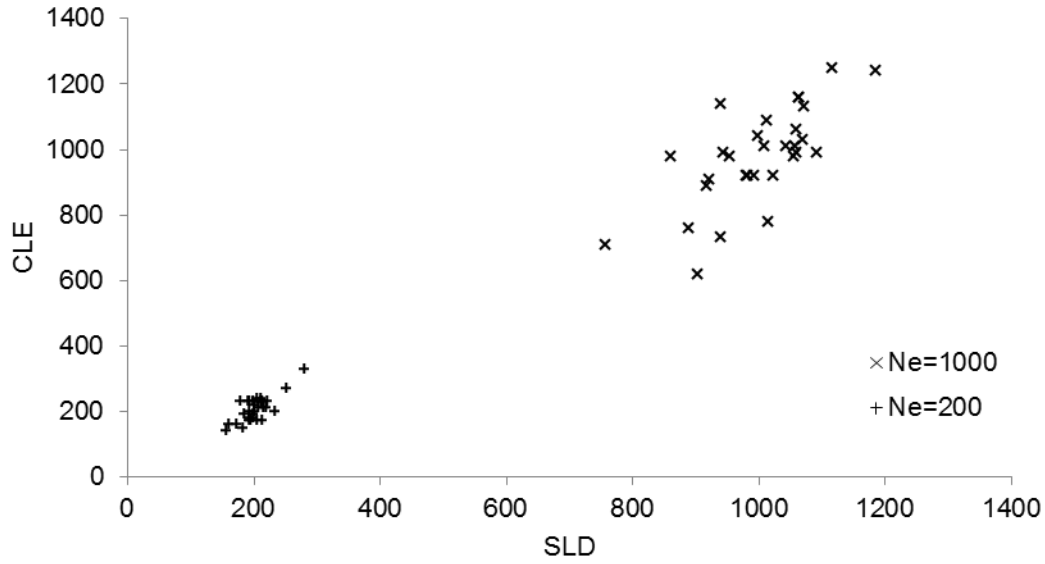

Figure 5-7 Relationship between estimates of N_e by CLE and SLD (MAF threshold of 0.05)

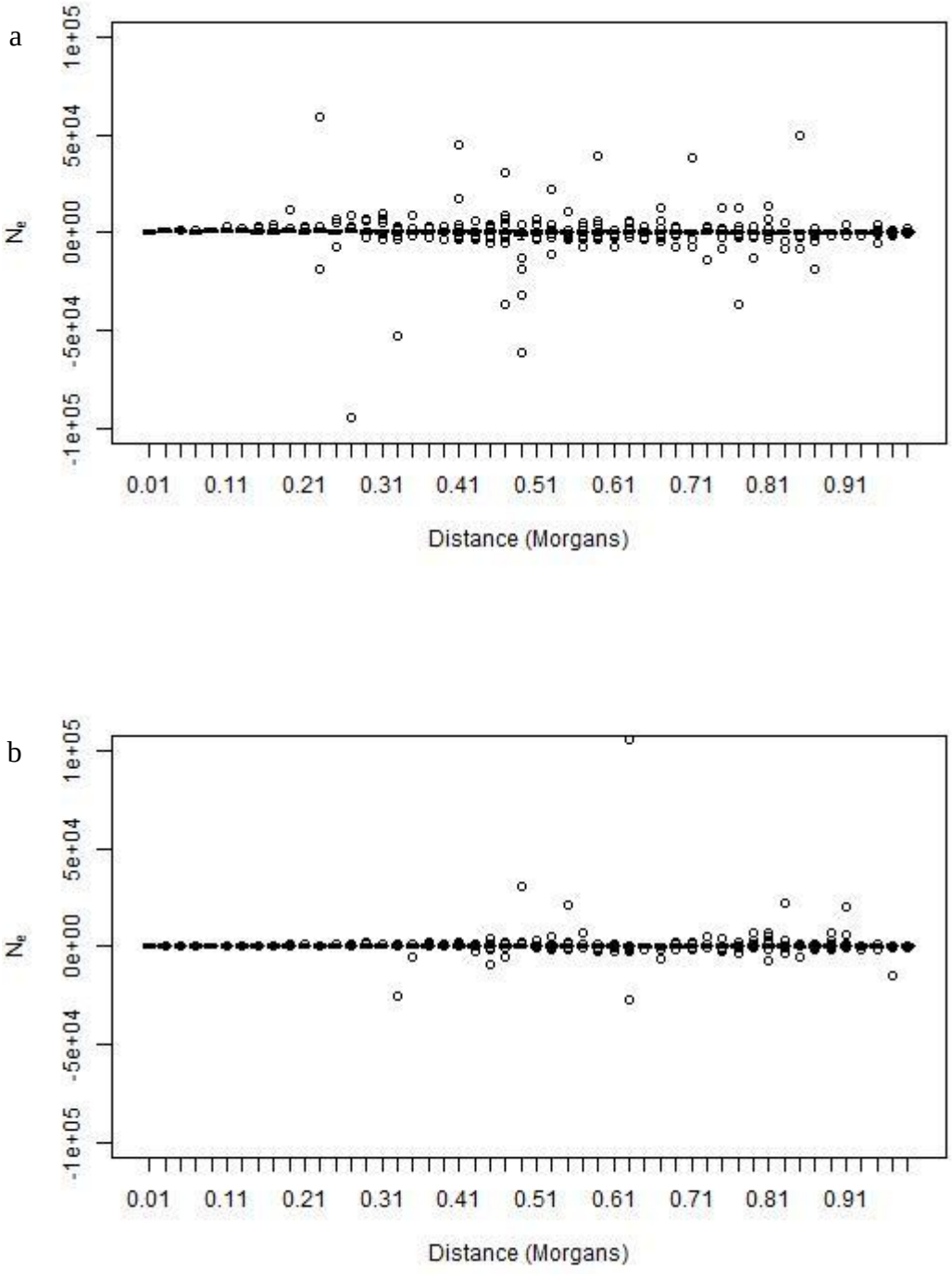

Figure 5-8 Estimates of N_e by SLD from marker pairs binned by distance. Variation at each distance reflects variability in estimates within distance bins and across the 30 replicates.

a) N_e=1,000

b) N_e=200

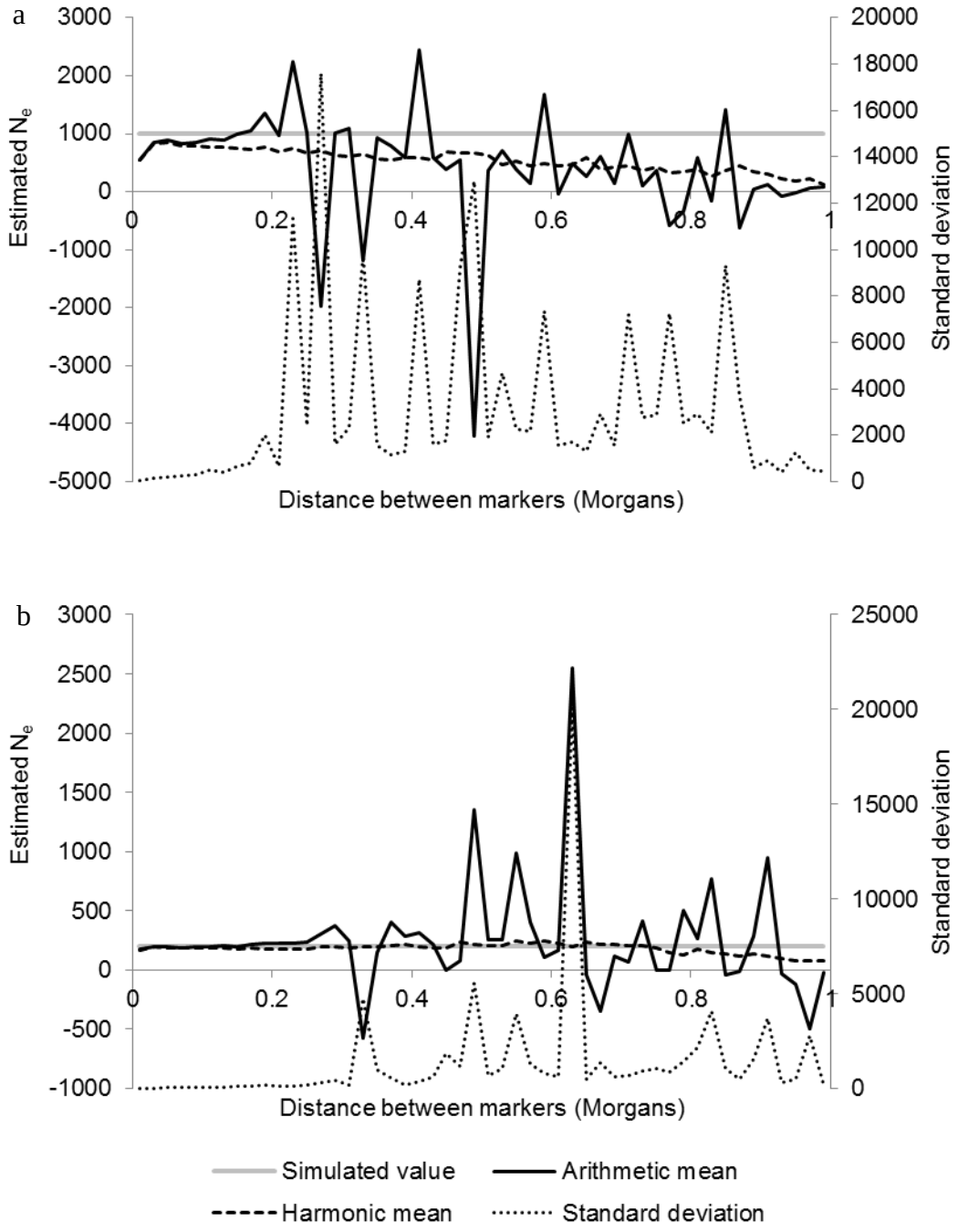

Figure 5-9 Mean and standard deviation of estimates of N_e by SLD from marker pairs binned by distance. For calculation of the harmonic mean, negative N_e estimates were replaced with estimates of infinity.

a) N_e=1,000

b) N_e=200

5.4 Discussion

To the author's knowledge, this is the first time the CLE method has been applied to markers on a chromosome-wide scale. Using data from all marker pairs, the CLE method produced two good mean N_e estimates. Asymmetric composite likelihood curves suggest that the method is less able to discriminate between large and very large N_e, than between small and very small N_e. Furthermore, the method appeared to more confidently estimate N_e when the true N_e was 200 than when it was 1,000, as evidenced by the more peaked likelihoods in Figure 5-4. Confidence intervals were not estimated because it has been suggested that the non-independence introduced into the calculation of coalescent likelihoods means that standard likelihood theory cannot be used to estimate uncertainty about point estimates [142, 182]. When CLE and SLD estimates are compared, while the mean estimates are similar, in two out of four of the scenarios tested, the SLD method was superior when judged according to the MSE of the estimates. This result is somewhat unexpected because according to Hudson (2001) [181], maximum likelihood methods should allow the most powerful analyses to be carried out on samples with multiple linked sites. Possible explanations for this finding are discussed below. A moderate positive correlation between N_e estimates by the CLE and SLD methods, indicates that they exploit some of the same information from the data.

Both the CLE and SLD methods showed considerable sensitivity to the distance between markers. When SNP pairs were binned according to the distance between them, SLD estimates remained within ±50% of the true value for slightly longer than CLE estimates. In the case of the CLE method, N_e estimates became increasingly upwardly biased and increasingly variable as the distance between the SNP pairs increased and this appears to be due to an increasing number of N_e estimates of 10,000. This is the maximum N_e that was tested and therefore these estimates could result from monotonically increasing likelihood curves, as described by Hudson (2001) [181], or may correspond to N_e estimates of greater than 10,000. Somewhat in contrast, SLD methods simply become extremely variable with increasing distance between SNPs and are in some cases negative because the mean r^2 value for the bin is less than the term of adjustment ($(T)^{-1}$). The rate at which the accuracy of N_e

estimates declines over distance depends on the actual N_e of the sample, such that when N_e = 200, estimates remain within ±50% for longer. Therefore, there appears to be less information contained in marker pairs that are further apart and in samples with a larger N_e. The number of marker pairs per bin acts as a confounding factor because of the linear relationship between distance and number of marker pairs in the bin. Further work that could be done to resolve this issue include: (i) fixing the number of marker pairs across the different distance classes; and (ii) investigating the impact of the number of marker pairs for a given distance between markers. However, for reasons discussed below, this method was not progressed further and therefore this was not done.

Using the harmonic mean to summarise estimates across replicates resulted in mean estimates closer to the simulated value in the case of both the CLE and SLD estimates. This approach resulted in N_e estimates becoming increasingly downwardly biased with increasing distance between marker pairs. The combination of N_e estimates by the harmonic mean has been used previously by Waples (2005) [189] when combining estimates from temporal samples and in reference to this work, Russell & Fewster (2009) [190] suggested that the reciprocal property ($1/N_e$), i.e. inbreeding rate, is a more direct driver of evolutionary processes than N_e itself. That said, the real problem is to understand on which scale different samples are additive, and therefore when it is natural to average.

The work presented here is a preliminary analysis and as such there are several areas within the methodology that would require further consideration should any further work be carried out. The CLE methodology described by Hudson (2001) [181] and followed here appears to implicitly assume small c and therefore, there are several issues which arise as a result of its application to long chromosome segments in this study. Using the genetic map distance, d_i, calculated as the product of the base pair distance and the recombination rate per base pair, in the calculation of ρ assumes an equivalence of map distance and recombination distance which is only valid for small c. Given the use here of marker pairs up to 1M apart, the transformation of d_i according to Haldane's mapping function before calculating ρ would seem more appropriate. The failure to do this here would be expected to result in the

underestimation of N_e when using markers that are far apart. An additional issue relating to the application of the method to long segments relates to the use of the downloaded tables of sample probabilities. These tables contain sample probabilities for ρ values ranging from 0 to 120. In this study, ρ values of up to 40,000 were used in calculating the CLE, placing considerable reliance on probabilities calculated via interpolation between probabilities for ρ=120 and ρ=∞. Because this was considered to be largely unsatisfactory, considerable time was spent by John Woolliams on generating sample probabilities for $120<\rho<\infty$, following Hudson (1983 & 1985) [143, 186]. However, this process revealed potential issues relating to the application of the coalescent theory itself.

Coalescent theory relies on the assumption that $n<<N_e$ [143] and therefore, the minimum sample size to N_e ratio of 1:4 used here may not be appropriate. Interestingly, having used Hudson's *ms* program to simulate a sample where $n>N_e$, Uleberg *et al.* (2011) [191] stated that they 'expected the genealogies to still resemble those in QTL mapping experiments involving unrelated individuals.' However, a more thorough evaluation of the properties of the sample would be needed to confirm the suitability of the *ms* program for simulating large samples from populations with small N_e. In addition, coalescent theory assumes small c [143] and so presumably cannot be expected to extend over a chromosome of 1M as simulated here; this may have contributed to the bias in N_e estimates observed at long distances. These issues have the potential to affect both the validity of my simulated data and the applicability of the sampled probabilities to my data. Further discussion around this point can be found in Woolliams & Corbin (2012) [192], but the fundamental question remains, can coalescent theory be used in the current context?

As pointed out above, the result that greater variation was seen in N_e estimates when the CLE method was used was unexpected. One potential explanation for this was that low MAF alleles added noise to the data because of their low information content. However, when a MAF threshold of 0.05 was implemented (as applied in the SLD method) and the analysis repeated, no gain in precision was seen (data not shown). This coincides with the results of McVean (2002) [183] who also found that the removal of rare variants had little effect on estimates of ρ. It is also possible

that there was either an undetected bug in the computational code, or an error in the interpretation of the procedure described by Hudson (2001) [181] to estimate N_e. The description of the CLE method provided by Hudson (2001) [181] is somewhat complex and therefore the implementation of the method was challenging. Considerable effort was put into generating the sample probabilities independently[8] in order to check the interpretation of Hudson's procedure, revealing several anomalies in Hudson's original work. For example, the method used here to calculate the 'a-d unspecified' sample probabilities did not include the adjustment by π, as described in Hudson's equation (6), because it was not clear what π represented.

In conclusion, using Hudson's CLE method in this case did not lead to a substantial increase in accuracy of N_e estimates over the non-linear regression based SLD method. Both methods suffered from a lack of power when linkage between markers was relatively loose, that is when the distance between markers was high. It seems likely that the CLE method as described by Hudson (2001) [181] can only usefully be applied over short distances. This, combined with the relative complexity of the CLE method when compared to the SLD method, meant that this approach was not pursued further in this thesis.

[8] Simulation program written by John Woolliams (The Roslin Institute).

Chapter 6: The identification of SNPs on the Equine SNP50 BeadChip with indeterminate positions

6.1 Introduction

During The Broad Institute's Equine Genome Sequencing Project, DNA from a single inbred Thoroughbred mare (Twilight) was sequenced to 6.8X coverage, producing a high quality draft assembly (for a full description of the assembly process see Wade *et al.* (2009) [23]). The sequence was generated using a whole genome shotgun (WGS) approach which involves the sequencing of many short fragments followed by their assembly, first into small groups or contigs, and then into larger sections known as supercontigs or scaffolds. The sequencing of the equine genome followed that of the human, mouse and dog and as such benefited from lessons learned. The assembly of the equine genome was done primarily using an enhanced version of the ARACHNE software algorithm [193] to align the many sequence reads, alongside molecular tools such as Fluorescence In-situ Hybridisation (FISH) which were used to assign remaining pieces of sequence [23]. Advantages of the WGS approach include its simplicity and ability to produce rapid early coverage [194]. In the case of simple genomes with few repeats WGS sequencing produces extremely high quality sequences. However, in the case of more complex genomes which contain numerous repeat sequences, the process of contig aggregation is much more difficult and therefore assembly errors are more common. Mammalian genomes show this complexity, with 46% of the assembled equine genome sequence identified as being repetitive [23]. It was estimated that chromosomal misassignment and local mis-ordering affected <0.3% of the initial assembly of the mouse genome sequence [194]. The equine genome is similarly complex and a similar proportion of errors would translate into 8Mb of incorrect sequence. Therefore, despite being sequenced to relatively high depth (6.8X), the sequence remains a draft and will likely continue to be updated for some time as these errors are identified. In the meantime, it serves as an invaluable resource for genomics studies.

A single-nucleotide polymorphism (SNP) library was compiled during the sequencing project. Whole genome shotgun sequencing is a particularly effective

method for detecting sequence variation in tandem with whole-genome assembly [195] and SNPs were identified both during the primary sequencing and subsequently through the re-sequencing of a number of horses selected on the basis of breed diversity. The correct positioning of these SNPs is dependent on the correct assembly of the contigs on which they reside and any assembly errors will be reflected in the positioning of these markers. The content of the Illumina Equine SNP50 BeadChip ("SNP chip") was derived from the resulting EquCab2.0 SNP collection and has proved to be a useful tool for equine geneticists worldwide.

The most common use of the SNP chip to date has been in the hunt for genes underlying both simple and complex diseases. The increased marker density offered has enabled both genome-wide association studies (GWAS) with combined linkage analyses and simple GWAS to be carried out with some success [65, 66, 196-198]. In such analyses whilst correct positioning of SNPs is clearly desirable, subsequent fine-mapping and re-sequencing work will likely bring to light any issues relating to SNP position. However, other genomic methodologies are much more sensitive to SNP order. The first of these, homozygosity mapping, can be particularly useful when attempting to find the gene variants responsible for rare recessive diseases for which there is often a paucity of samples available for analysis. The aim of this technique is to identify long stretches of homozygosity present only in the cases which may contain the causal mutation, the assumption being that this segment is inherited identical-by-descent from a common ancestor carrying the mutation. In this instance, incorrectly positioned SNPs may disrupt runs of homozygosity leading to a failure in their identification by algorithms designed to pick up runs of unusual length. Whilst single misplacements may be dealt with by allowing for some errors, when lengths of sequence contain more than one misplaced SNP, it becomes impossible to differentiate such errors from normal variation. A second technique which is heavily dependent on SNP order is haplotype and phase assignment. Increasingly, investigators are expecting genetically derived algorithms to assign phase to population data, allowing individuals in the population to be assigned haplotypes in place of the diplotypes which are returned from the SNP genotyping process. These haplotypes can then be used both in future analyses, such as

haplotype association tests, and in order to impute from low-density SNP panels containing so-called tag SNPs to higher densities. It has been suggested that incorrect SNP order could compromise the accuracy of haplotype analysis [199], specifically causing the inaccurate estimation of haplotype block size in the region with errors [200].

Whilst there are many molecular methods available to help refine genome sequence assemblies, it has been suggested that linkage disequilibrium (LD) analyses may be a useful addition to our armoury in resolving SNP order [201, 202]. Khatkar *et al.* (2010) [201] developed a locus ordering procedure based on LD (LODE) in order to position unassigned SNPs and scaffolds on the *Bos taurus* genome, with validation work showing that 96.7% of SNPs with a minor allele frequency (MAF) greater than 0.05 could be positioned with an accuracy of 99.9% and an average precision of ~1Mb; the same procedure was then used to check existing genome assembly. Bohmanova *et al.* (2010) [202] compared observed LD patterns to expectations under genetics' theory in order both to identify misplaced SNPs and to re-position them.

Previous studies of LD in the horse have shown LD to be moderate within-breed and to be particularly high in Thoroughbreds, being maintained above non-syntenic levels for up to 20Mb (Chapter 3 and [23]). In this chapter, LD is used to check genome assembly in preparation for work to identify QTL associated with OCD (Chapter 7) and work to select SNPs for low-density panels (Chapter 9). An outcome of this chapter is to identify regions of the genome which may benefit from further re-sequencing.

6.2 Materials and methods

The data for this study comprised all 1,201 available samples (n=853 Thoroughbreds from the UK and n=348 Thoroughbreds from the US), genotyped at the 50,707 SNPs that passed preliminary quality control (QC). PLINK [105, 106] was used to calculate the squared correlation based on genotypic allele counts; this is identical to the r^2 measure of LD when mating is at random, i.e. assuming genotypic frequencies are in Hardy-Weinberg equilibrium (HWE). However, to denote the distinction in

calculation method, r_g^2 will be used. Values of r_g^2 were calculated in a pair-wise fashion between all SNPs with a MAF≥0.01 (9,997 exclusions made), including those on different chromosomes. The average MAF of the remaining SNPs was 0.26 (±0.14 SD). Autosomal SNP pairs with r_g^2>0.25 were retained for further analysis.

A filter was applied to non-syntenic marker pairs in order to identify all such pairs with r_g^2>0.25. A filter was also applied to syntenic marker pairs, this time to identify all SNPs for which the maximum recorded r_g^2 value was with another SNP that was >10Mb away in at least one direction. Previous work has shown that less than 0.5% of SNP pairs separated by 10Mb can be expected to exhibit r^2>0.25 (Chapter 3). This initial analysis uses subjectively chosen thresholds to identify SNPs whose location may be incorrect whilst minimising false positives and is consistent with Bohmanova *et al.* (2010) [202]. A second step was taken to validate the regions identified through the visualisation of LD structure in Haploview [203] and through the examination of comparative maps of *Equus caballus* chromosomes (ECA) against *Homo sapiens* chromosomes (HSA) using the synteny option in ensemble (Ensembl (2012). *Equus caballus.* [Online] Available from: http://www.ensembl.org/Equus_caballus).

6.3 Results

6.3.1 SNPs assigned to the wrong chromosome

The test for non-syntenic r_g^2>0.25 identified two potential discrepancies. The first involved a single SNP, BIEC2-723147, whose position is given as 232,540 on ECA28 but which showed considerable LD with a group of SNPs on ECA10. This SNP had r_g^2>0.25 with a total of 63 SNPs on ECA10 (located between 19,507,420 and 33,005,513), and was in complete LD with two of these SNPs (BIEC2-111876 and BIEC2-111538). Further analysis in Haploview [203] revealed BIEC2-723147 had r_g^2≤0.002 with its current neighbours on ECA28 up to 1.1Mb away (Figure 6-1) and was in greater LD with SNPs in the region identified on ECA10. BIEC2-723147 is the second SNP on ECA28 on the SNP chip but in this analysis the first SNP was excluded due to monomorphism. No unusual pattern of synteny was revealed in the

region during comparative work; up to 30.6Mb on ECA28 corresponds to HSA12 and the remainder to HSA22.

The second discrepancy involved a cluster of SNPs mapped to ECA5 (14,185,421 – 15,408,157) which showed unusually high LD with many SNPs on ECA19 (6,860 – 4,114,838). Further analysis in Haploview [203] showed that the cluster of SNPs identified in the primary analysis in fact extended to a SNP at 15,914,852. White regions in the Haploview plot in Figure 6-2a, which are indicative of extremely low LD, clearly delineate these SNPs on ECA5 from the rest of the chromosome. This unusual LD structure led me to hypothesise that this group of SNPs currently mapped to ECA5 do not belong on this chromosome and may in fact belong on ECA19. The comparative analysis of this region offers further support for this hypothesis. Figure 6-2b shows the synteny between horse and human for ECA5. The centre chromosome represents ECA5, and the smaller chromosomes show syntenic regions with human sequence (in this case HSA1, 3 and 7). Blocks are coloured according to the human chromosome number and are connected with lines to indicate syntenic blocks with either the same orientation (black lines) or with the opposite orientation (brown lines). Figure 6-2b shows that, whilst the majority of ECA5 maps to HSA1, the genomic region identified in this analysis corresponds to HSA3. There is also a small section of ECA5 which shows synteny with HSA7. A similar evaluation of the synteny between ECA19 and humans showed a large part of ECA19, including the region with which the ECA5 SNP cluster showed LD, corresponds to HSA3 (data not shown).

HWE was examined as a guard against genotyping error; none of the SNPs identified in this part of the analysis deviated significantly from HWE ($p<0.001$), as calculated using Plink [105, 106] according to the exact test described by Wigginton *et al.* (2005) [204]. The average MAF of anomalous SNPs by chromosome can be found in Table 6-1.

Table 6-1 MAF of SNPs identified as having r_g^2>0.25 with SNP(s) on different chromosomes

Chromosome	Number of SNPs	MAF			
		Mean	SD	Min	Max
5	20	0.288	0.122	0.096	0.495
19	44	0.393	0.086	0.150	0.495
10	63	0.182	0.101	0.073	0.359
28	1	0.113	NA	NA	NA

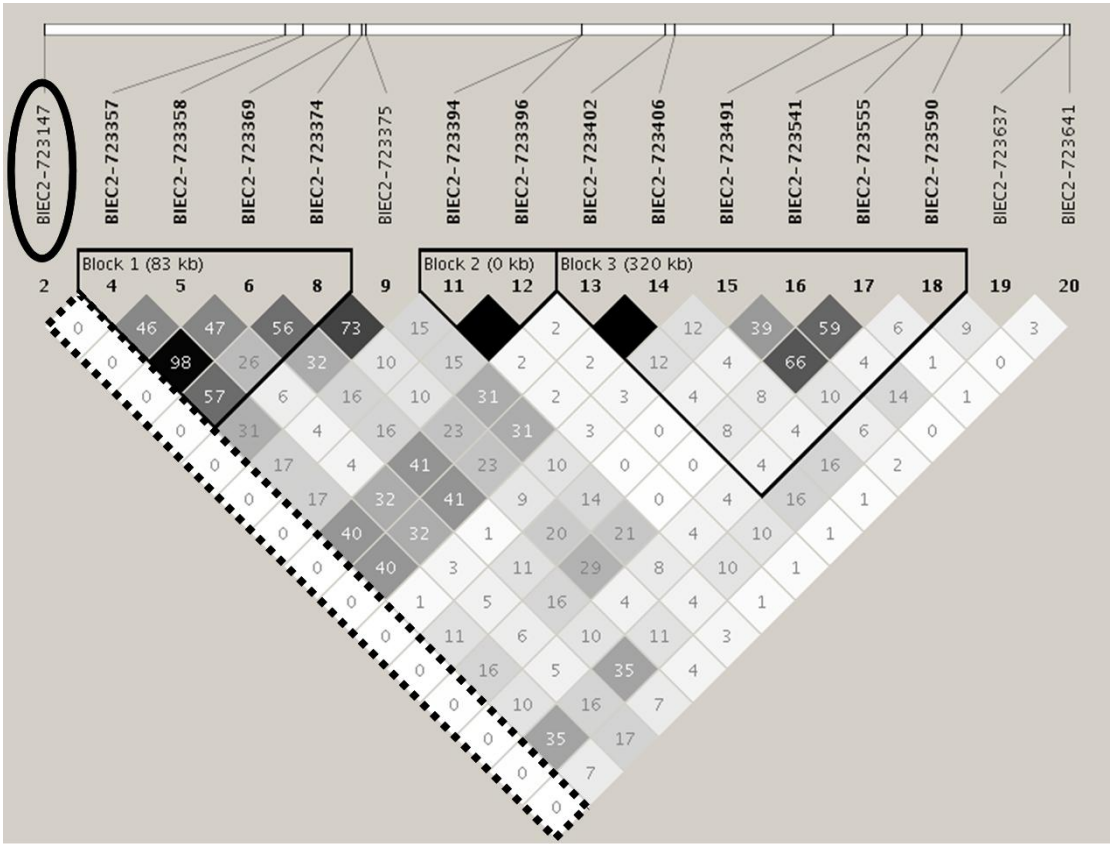

Figure 6-1 LD plot of ECA28: 0 – 1,342,640 with r^2x10^2 values shown. Anomalous SNP, BIEC2-723147, is circled (plot produced in Haploview [203]). SNPs 1, 3, 7, and 10 excluded during QC.

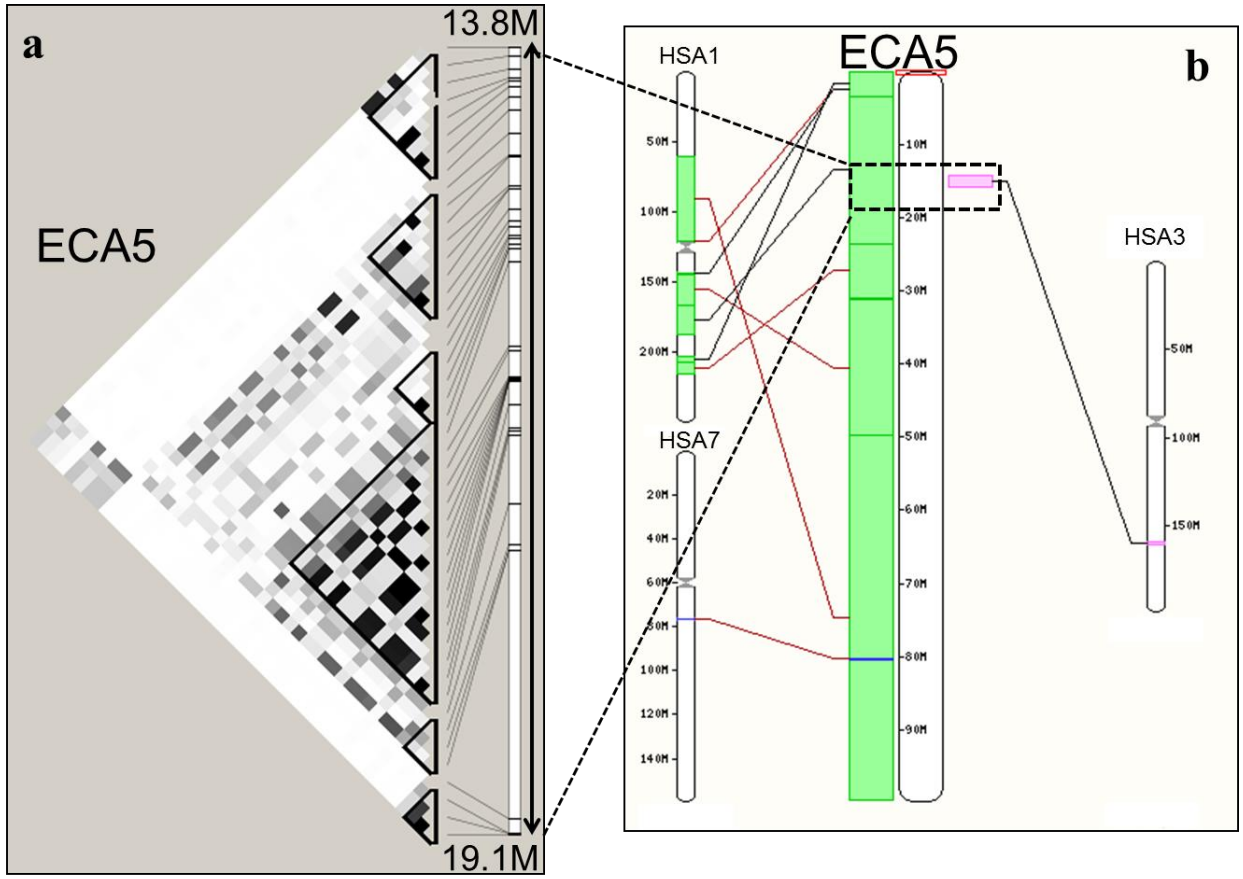

Figure 6-2 LD and synteny on ECA5

a) LD plot of ECA5: 13,827,917 – 17,091,141 (plot produced in Haploview [203]); b) Map of ECA5 showing synteny between horse ECA5 and human HSA1, 3 and 7 (http://www.ensembl.org/Equus_caballus). Dashed lines indicate correspondence between the two diagrams.

6.3.2 Within-chromosome discrepancies

A total of 39 SNP pairs met the criteria for unusual within chromosome LD. Most of the SNPs were on ECA5 (6), ECA7 (11) and ECA8 (6) (see Table 6-2 for full list). It should be noted that the SNPs identified in this analysis had a much lower than average MAF which may have led to biases in the estimation of r_g^2. None of the SNPs identified deviated significantly from HWE, calculated as described above. In the case of syntenic marker pairs, it is less straightforward to identify the limits of the discrepancies using Haploview [203]. However, the LD structure around anomalous SNPs was found to display potential irregularities. As shown for ECA7 and ECA8 in Figure 6-3 and Figure 6-4, respectively, some of the areas identified by this analysis showed a banding pattern of LD. In such cases, comparative analysis showed a complicated picture of chromosome synteny, exhibiting both intra- and inter-chromosomal rearrangements when compared to the human genome (Figure 6-3 and Figure 6-4).

Table 6-2 SNPs for which the maximum recorded LD value (in at least one direction) was r_g^2>0.25 and with a SNP >10Mb away

SNP1[I]					SNP 2[II]	
Name	ECA	Position (bp)	MAF	HWE *p*-value[III]	Name	Position (bp)
BIEC2-30677	1	72,451,788	0.02	0.04	BIEC2-25581	62,349,857
BIEC2-475014	2	41,857,074	0.11	0.65	BIEC2-477849	51,877,913
BIEC2-477849	2	51,877,913	0.08	0.15	BIEC2-474820	41,239,675
BIEC2-487672	2	71,430,607	0.11	0.66	BIEC2-493166	82,177,609
BIEC2-775088	3	27,995,999	0.06	0.79	BIEC2-779131	44,484,312
BIEC2-779131	3	44,484,312	0.05	1.00	BIEC2-774953	27,229,795
BIEC2-787064	3	64,658,125	0.11	0.17	BIEC2-781998	53,959,246
BIEC2-891591	5	10,750,326	0.05	0.24	BIEC2-898063	22,107,371
BIEC2-898063	5	22,107,371	0.06	0.00	BIEC2-891591	10,750,326
BIEC2-906263	5	39,300,989	0.17	0.22	BIEC2-908762	51,065,448
BIEC2-907143	5	41,918,001	0.09	0.72	BIEC2-909281	53,875,750
BIEC2-907163	5	42,334,840	0.10	0.64	BIEC2-909281	53,875,750
BIEC2-916364	5	69,429,005	0.15	0.29	BIEC2-910936	57,420,998
BIEC2-996254	7	38,197,047	0.12	0.41	BIEC2-998369	48,314,947
BIEC2-996263	7	38,199,984	0.33	0.40	BIEC2-1000225	53,386,780
BIEC2-996515	7	39,533,116	0.16	1.00	BIEC2-1000179	53,064,035
BIEC2-996542	7	39,651,243	0.30	0.27	BIEC2-998618	50,594,329
BIEC2-996583	7	40,036,320	0.10	0.76	BIEC2-998755	51,944,102
BIEC2-997395	7	47,105,730	0.08	0.70	BIEC2-1002897	57,211,137
BIEC2-998395	7	48,651,819	0.45	0.42	BIEC2-996153	37,517,778
BIEC2-998755	7	51,944,102	0.10	0.52	BIEC2-996583	40,036,320

SNP1[I]					SNP 2[II]	
Name	ECA	Position (bp)	MAF	HWE *p*-value[III]	Name	Position (bp)
BIEC2-1000179	7	53,064,035	0.34	0.17	BIEC2-996491	39,157,011
BIEC2-1000243	7	53,507,061	0.03	0.65	BIEC2-996624	40,420,239
BIEC2-1002897	7	57,211,137	0.07	0.83	BIEC2-997395	47,105,730
BIEC2-1036317	8	21,474,716	0.02	1.00	BIEC2-1043377	33,536,248
BIEC2-1037747	8	24,686,529	0.06	0.79	BIEC2-1046813	42,900,973
BIEC2-1041132	8	29,729,387	0.05	0.77	BIEC2-1046803	42,896,634
BIEC2-1045831	8	38,656,972	0.08	1.00	BIEC2-1039998	27,865,163
BIEC2-1046803	8	42,896,634	0.03	0.05	BIEC2-1041132	29,729,387
BIEC2-1047603	8	44,378,683	0.13	0.54	BIEC2-1040198	27,978,277
BIEC2-1088412	9	37,902,127	0.02	0.36	BIEC2-1092490	48,511,213
BIEC2-1092490	9	48,511,213	0.03	0.57	BIEC2-1088143	36,940,286
BIEC2-211410	13	13,096,070	0.05	0.74	BIEC2-202330	543,297
BIEC2-373784	17	21,903,878	0.01	1.00	BIEC2-375506	34,316,766
BIEC2-375506	17	34,316,766	0.02	0.47	BIEC2-373784	21,903,878
BIEC2-410605	18	34,575,897	0.03	0.58	BIEC2-408240	19,674,197
BIEC2-449520	19	59,892,096	0.06	0.12	BIEC2-444554	49,441,028
BIEC2-700515	27	3,763,777	0.03	1.00	BIEC2-707051	16,525,618
BIEC2-706894	27	16,393,229	0.03	1.00	BIEC2-700515	3,763,777

[I]SNPs identified during filtering procedure

[II]SNPs with which identified SNP(s) shared maximum r^2

[III]Calculated using Plink [105, 106] according to Wigginton *et al.* (2005) [204].

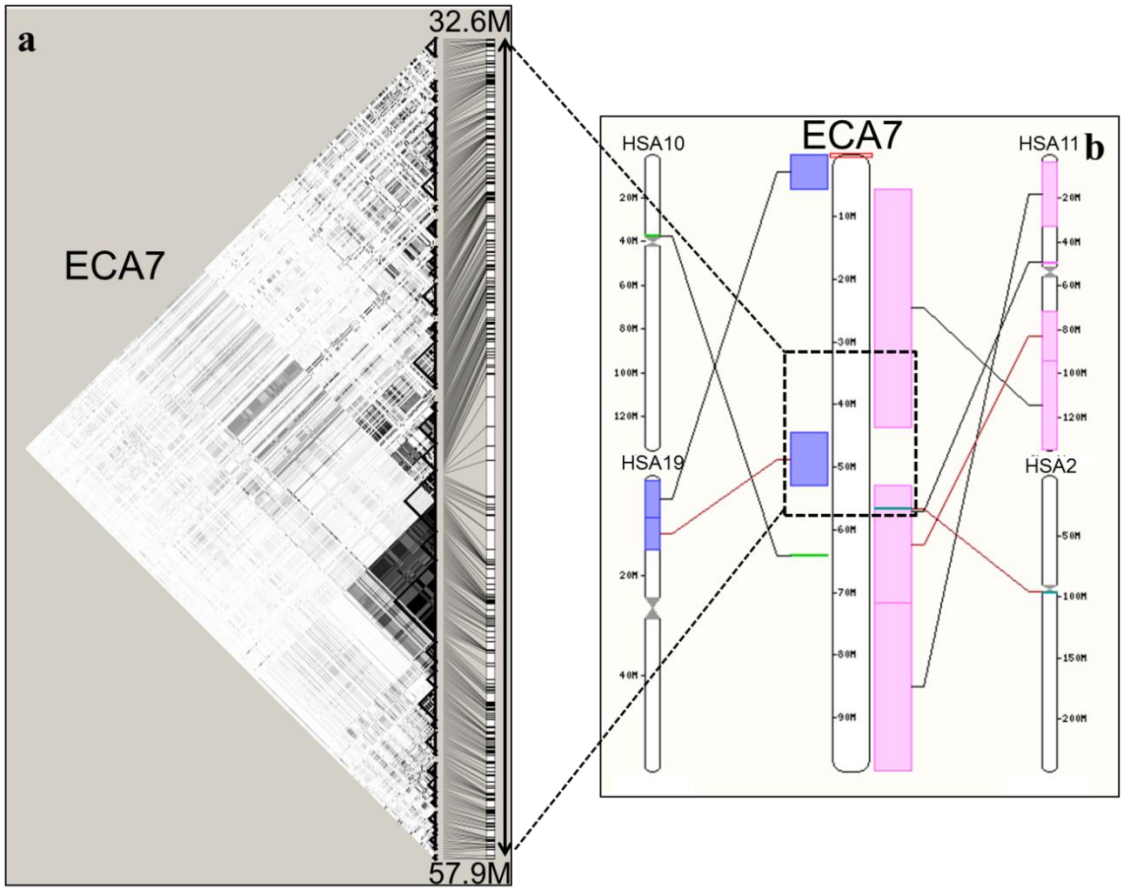

Figure 6-3 LD and synteny on ECA7. a) LD plot of ECA7: 32,554,745 – 57,862,186 (plot produced in Haploview [203]). This region contains SNPs from ECA7 listed in Table 6-2; b) Map of ECA7 showing synteny between horse ECA7 and human HSA2, 10, 11 and 19 (http://www.ensembl.org/Equus_caballus). Dashed lines indicate correspondence between the two diagrams.

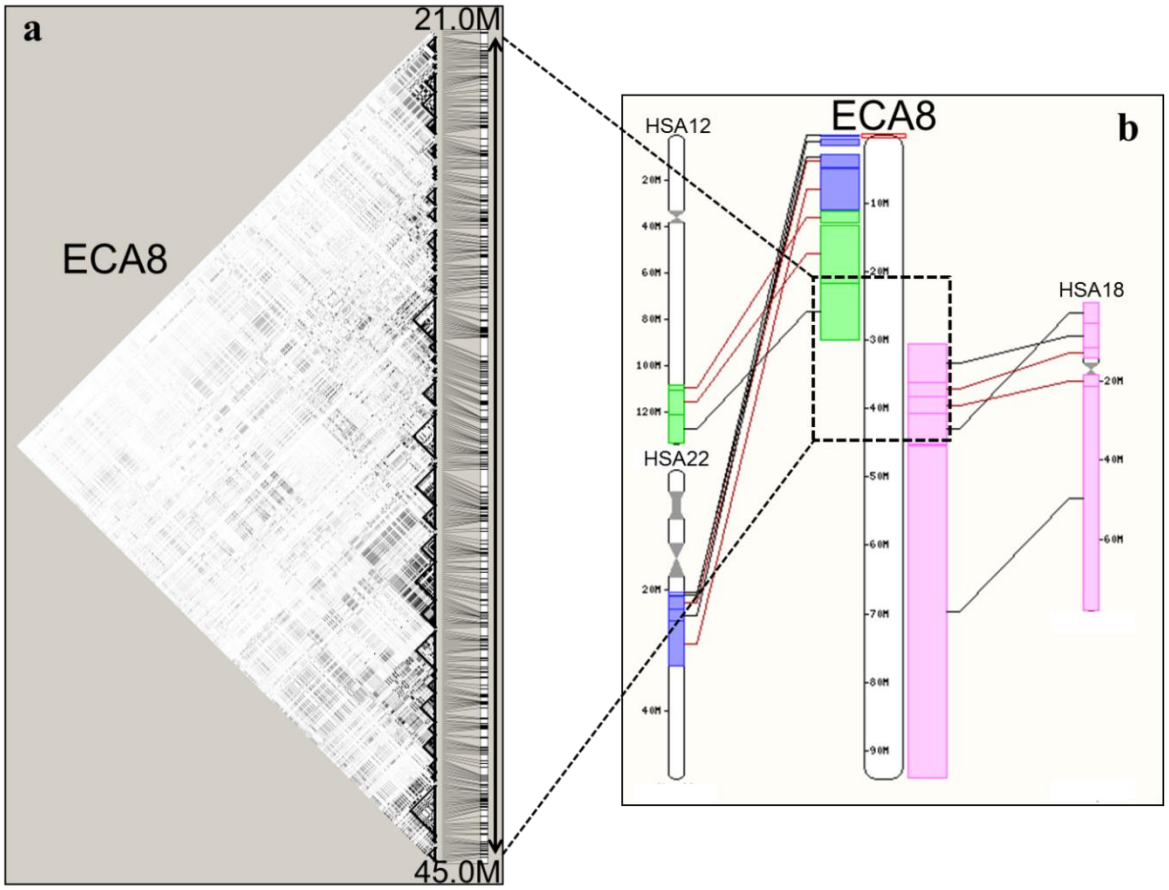

Figure 6-4 LD and synteny on ECA8. a) LD plot of ECA8: 21,048,394 – 44,982,527 (plot produced in Haploview, [203]). This region contains SNPs from ECA8 listed in Table 6-2; b) Map of ECA8 showing synteny between horse ECA7 and human HSA12, 18 and 22 (http://www.ensembl.org/Equus_caballus). Dashed lines indicate correspondence between the two diagrams.

6.4 Discussion

This work has demonstrated the use of LD analysis in the evaluation of sequence quality. The evaluation of across chromosome LD identified both a single SNP and a group of SNPs likely to represent sequence errors. Within-chromosome analyses revealed several SNPs which demonstrated higher LD with distant SNPs than with neighbouring SNPs. Further interrogation of the chromosome regions containing these SNPs revealed both unusual LD structure and complex patterns of synteny. Whilst these secondary findings provide some validation for the method used, there remains ambiguity about the cause of the genome structure observed.

In the case of non-syntenic SNPs, high LD between SNPs appears to be a good indication that there is an error in the genome assembly. My results suggest that, whilst looking for SNPs mapped to the wrong chromosome, the threshold used to identify erroneous SNPs (in this case r_g^2>0.25), is not crucial. Whilst the threshold used here did not pick up all the SNPs within the ambiguous region on ECA5, subsequent haplotype analysis helped to define the limits of the region (as illustrated in Figure 6-2). Therefore, provided at least one SNP in the region returns an r_g^2 value above the threshold, erroneous regions will be found. Furthermore, maintaining a relatively high threshold offers an advantage in terms of computing time. To see genuine LD across chromosomes is extremely unusual and, at the current time, there are only a few known (or suspected) causes of across chromosome LD [35]. Admixture or migration can cause spurious LD between unlinked markers but this quickly dissipates and in any case would be expected to affect multiple regions across the genome. Epistasis between loci has also been shown to have the potential to maintain non-syntenic LD at levels that would otherwise not be seen under free recombination [205]. However, the pattern of non-syntenic LD observed here does not seem to fit either of these scenarios.

Further investigation based on the synteny map shown in Figure 6-2b shows that, whilst the majority of ECA5 aligns to HSA1, there are two small regions that do not, one aligning to ECA7 and the other to ECA3. Interestingly, my analysis identified only the region aligning to ECA3 as being potentially erroneous. Zoo-FISH analysis has shown that many of the meta/submetacentric chromosomes (1-13) do indeed

exhibit such breakages in synteny conservation as a result of interchromsomal re-shuffling, including in some cases within-arm breakages [206]. Therefore, the apparent discrepancy in synteny shown cannot be considered confirmation of a sequencing error. Even the fact that in the zoo-FISH analysis ECA5 showed whole chromosome conservation with HSA1 [206] cannot be considered conclusive since segments smaller than 5-7Mb are usually difficult to detect through cross-species chromosome painting [207]. However, in this instance, confirmation was possible through discussions with a principal member of the Horse Genome Project who confirmed that the region I had identified on ECA5, had indeed been the subject of an assembly error which resulted in a section of the centromeric portion of ECA19 being incorrectly placed on ECA5 (C. M. Wade 2011, Pers. Comm.). Therefore, the independent discovery of this error shows the potential utility of this approach.

Being within the first 300kb of ECA28, the genomic region containing the single SNP identified on ECA28 as being potentially mapped to the wrong chromosome is at relatively high risk of being incorrectly assembled, with sequence coverage tending to be lower at the ends of chromosomes. The scarcity of markers in this region makes confirmation of the finding and further evaluation of the extent of the error problematic. The implication of this is that the potential discrepancy in the sequence will only be resolved with further sequencing.

In the case of SNPs showing unusual LD patterns within a chromosome, it is much more challenging to differentiate between likely sequence error and simply unusual LD. Suitable distance thresholds for the identification of erroneous SNPs are difficult to decide on but presumably should be based to some extent on expectation in order to identify unusual patterns of LD. However, unusual LD could be considered an expectation in itself. Whilst theoretically, LD decays with time and distance, such that $D_t = (1-r)^t D_0$, where D_0 is the extent of disequilibrium at some starting point and D_t is the extent of disequilibrium t generations later, this equation poorly represents the behaviour of LD over short distances [35] and it has been estimated from human data that, for relatively short distances (<5Mb), only 45% of variation in disequilibrium measures can be explained by physical distance [208].

Consequently, although a trend towards decreasing disequilibrium with increasing distance is observed (as shown in Chapters 3 and 4), significant variability exists as illustrated by the evaluation of LD in small regions [208, 209] and by looking at r^2 distributions produced from genome-wide data (Chapter 3, Figure 3-2). Factors (in addition to mutation and recombination) which have been shown to contribute to the extent and distribution of disequilibrium include demographic factors and natural selection (see Ardlie *et al.* (2002) [35] for a full discussion). In this instance, a relatively conservative distance threshold for unusually high LD of 10Mb was used; this allowed me to pick up the most extreme cases, whilst limiting the number of false positives (i.e. SNPs that are correctly positioned but demonstrate unusual patterns of LD) and ensuring a manageable number of SNPs for follow-up. Lowering the distance threshold would inevitably have led to more SNPs being identified, increasing the risk of finding false positives and leaving a long list of potential problem areas requiring further investigation.

Unlike the non-syntenic case, visualisation of the problem area does not seem to help define its limits. Furthermore, whilst both the banding patterns shown in the LD plots and the complex pattern of synteny (Figure 6-3 and Figure 6-4) for the areas of ECA7 and ECA8 identified in the syntenic analysis, support the finding of unusual LD, neither help us to resolve easily the source of the LD structure. Therefore, whilst incorrect sequence assembly is one explanation, it is also possible that relatively recent intrachromosomal rearrangements, known to occur within species, could be responsible. The correspondence between patterns of LD and synteny, as shown in Figure 6-3 and Figure 6-4 for ECA7 and ECA8 respectively, is in some cases as expected. For example, in Figure 6-3, a block of high LD corresponds approximately with a block of the chromosome marked on the synteny map as corresponding to HSA19. However, in other cases, more unexpected patterns can be seen, for example in the same figure, banding patterns show relatively high LD between SNPs in two non-adjacent regions that align to two different human chromosomes (HSA19 and HSA11). Overall, the interpretation of my results based on synteny between horse and human chromosomes is dependent on the human

genome sequence being correct, but, this seems a fair assumption to make given the extent of the work that has been done in compiling this sequence.

Given our inability to resolve the issue of causality, particularly in the case of unusual syntenic LD, how useful is this method in identifying misplaced SNPs? My results have demonstrated that LD can be used to identify SNPs, and therefore regions of sequence, which have been assigned to the wrong chromosome during the assembly process. These SNPs can then be excluded for analyses for which SNP order is important, e.g. homozygosity mapping. In the case of GWAS or similar QTL mapping studies, such SNPs need not be excluded from the analysis since association statistics are not generally affected by SNP order. However, it must be borne in mind that patterns of association observed in, for example, the commonly used Manhattan plot might be disrupted in areas containing incorrectly positioned markers. Furthermore, should any of the incorrectly positioned SNPs be identified as associated with the phenotype, assurance would be needed that the search for candidate genes and other functional genome components was conducted in the relevant region. In the long term, focussed re-sequencing of the regions concerned would presumably enable contigs to be confidently re-assigned to their correct position.

The lack of clarity in the syntenic case limits the usefulness of this approach in identifying SNPs that should be excluded from future analysis. Many of the SNPs identified as showing unusual patterns of LD had a low MAF and this result correlates with the findings of Khatkar *et al.* (2010) [201] who saw a dramatic decrease in the number of SNPs that could be positioned by LODE (their procedure for placing SNPs according to LD) when the MAF of SNP was <0.05. The impact of these SNPs on most analyses will likely be limited, with nearly one third of the SNPs in Table 6-2 having MAF <0.05, a commonly used exclusion threshold. However, remaining incorrectly positioned SNPs would still have the potential to disrupt analyses such as homozygosity mapping. The impact of incorrectly positioned SNPs during tag SNP selection can be limited by imposing a constraint on distance, such that SNPs cannot be picked as tags for other SNPs if they are greater than a defined distance away, regardless of the LD shown.

An alternative to simply identifying erroneous SNPs for exclusion might be to attempt to re-position SNPs using LD. However, in this case, the emphasis must be on the re-ordering of contigs and scaffolds, rather than the re-ordering of SNPs *per se*. Whilst SNPs and other markers allow the visualisation of sequence error, simply re-positioning SNP loci in isolation using LD would not address the sequence errors within the assembly itself and would leave SNPs disassociated from surrounding sequence. This represents an undesirable outcome when performing, for example, GWAS where regions surrounding associated SNPs are searched for candidate genes. For this reason, the re-positioning of SNPs according to LD, as suggested by Bohmanova *et al.* (2010) [202], would seem to have limited utility in the long term. Furthermore, patterns of LD are extremely variable, particularly over short distances, and therefore whilst it is possible to hypothesise the general location of SNPs using LD analyses such as those performed here, it seems unwise to attempt to definitively reposition them in this way. For example, since BIEC2-723147 (currently mapped to ECA28) exhibited complete LD with two SNPs on ECA10 which were separated by more than 2Mb containing 37 SNPs, re-positioning this particular SNP next to the SNP with which it exhibits the highest LD was neither possible nor desirable. It seems likely that only further molecular analysis will allow potential sequence errors to be completely resolved.

The methods used in this study are a step towards identifying and resolving errors in the equine genome sequence. The ability to prioritise specific regions of the genome for re-sequencing in this way may prove to be more important in species, such as the horse, for which resources are much more limited than in humans. With next generation sequencing making possible the sequencing of longer reads, it is likely that errors associated with incorrect contig assembly will be reduced in the future. In the meantime, this work serves as a reminder that the current version of the equine genome sequence remains a draft, and researchers should be aware that some SNPs on the Illumina Equine SNP50 BeadChip may be misplaced.

Chapter 7: A genome-wide association study of osteochondritis dissecans in the Thoroughbred

7.1 Introduction

Osteochondrosis (OC) is a disease of the locomotory system which affects the joints of many animals, most frequently being observed in pigs, horses and dogs. Osteochondrosis can be described as a focal disturbance of endochondral ossification [68] that occurs in young, growing individuals and as such has been classified as a developmental orthopaedic disease. A more detailed description of OC can be found in 1.3.1. Prevalence estimates for OC vary widely, ranging from 3% (stifle OC in Thoroughbreds [210]) to 70% (estimates for all joints in Dutch Warmbloods [69]). A large proportion of this variation is attributable to differences in the type and number of anatomical locations examined, differences in the specific manifestation of the disease considered and breed differences (see 1.3.4) [69, 77, 88, 211]. A recent prevalence estimate of 25% for the Thoroughbred [76] appears typical. This relatively high disease prevalence, along with the likely contribution of OC to the predominance of lameness as a cause of wastage in young horses [212, 213], makes OC a high priority for study.

Whilst there exists both experimental and anecdotal evidence of a genetic component to OC, the aetiopathogenesis of the disease is not fully understood [68]. The disease is considered multifactorial in origin with at least some evidence of both environmental factors, for example nutrition, and physiological factors, such as growth and body size, endocrine factors and conformation, which may themselves be mediated through genetics, playing a role in the condition [75, 76, 109]. Low to moderate estimates of heritability for OC across a range of breeds and disease manifestations [77, 80, 86-88] together with between breed differences in prevalence [76] indicate that genetic variability exists in disease susceptibility. Typical values for OC scored as a single binary trait (all joints combined) are 0.10 to 0.20 [77, 86] but heritability estimates of up to 0.5 have been reported for individual joints [214].

The search for markers to explain the proposed genetic variance in susceptibility to OC began several years ago, with the intention both of enhancing our understanding

of the condition and of enabling marker assisted selection (MAS). Early studies using primarily linkage based analyses (dependent on family data), to detect regions of the genome associated with OC in the horse have identified several putative quantitative trait loci (QTL) [97, 98]. As is typical for QTL discovered using this approach, their effects are generally large but their locations are imprecise. Whilst several of these QTL have undergone further refinement, few if any have been validated in independent data sets. Similar studies in pigs have revealed few [215] or no [216] QTL for OC. These results illustrate the difficulty in identifying truly associated regions for complex traits using linkage analysis.

The opportunity for QTL studies in horses has recently been advanced by the publication of the equine genome sequence [23] together with the release of the Illumina Equine SNP50 BeadChip, which has allowed the implementation of genome-wide association studies (GWAS). In contrast to linkage analysis, GWAS rely on samples of individuals, which may be unrelated, genotyped at medium to high density. It is expected that this approach will allow the identification of common variants which could not be found using the traditional linkage based approach [40]. I am aware of four GWAS for OC that have been carried out in three different horse breeds to date: Lampe (2009) [99] and Komm (2010) [100] (using the same data), Teyssèdre *et al.* (2011) [101] and Lykkjen *et al.* (2010) [102]. The number of QTL identified per study ranges from four [102] to 18 [99] with the range likely at least partly attributable both to differences in significance thresholds used and to differing phenotype definitions. Three putative correspondences between QTL have been described (see 1.3.5).

This study demonstrates the use of clinical observations as a source of data for use in genomic studies and is the first QTL mapping study for OC to be conducted in the Thoroughbred. A GWAS was performed on 348 samples using the Illumina Equine SNP50 BeadChip to identify loci associated with OCD in the Thoroughbred. In addition, QTL for OC previously identified in a Hanoverian Warmblood (HWB) population were tested for their effect in the current data set.

7.2 Materials and methods

The data for this study comprised the OC samples that were sourced in the US (*n*=348 Thoroughbreds from the US), genotyped at the 50,707 SNPs that passed preliminary quality control (QC).

7.2.1 Quality control

Quality control for this study was carried out in GenABEL [217]. Firstly, samples were checked for sex discrepancies (marker-based prediction of sex versus sample label) and intermediate X-chromosomal inbreeding ($0.2<F<0.8$), with exclusions being made on the basis of suspected sampling or genotyping errors. This process resulted in two exclusions due to sex discrepancy and 16 exclusions based on indeterminate sex as demonstrated by intermediate inbreeding, leaving 168 controls and 162 cases for further analysis. Further, 30 SNPs were excluded on the basis that they were positioned on the X chromosome but were likely autosomal.

For the GWAS (see below) the following thresholds were used for excluding data: minor allele frequency <0.05, >5% missing genotypes per SNP, >5% missing SNPs per sample and differential proportions of missing SNPs between cases and controls ($p<0.05$). No exclusions were made on the basis of Hardy-Weinberg equilibrium (HWE).

For construction of a kinship matrix (see below), a subset of markers meeting more stringent QC was chosen as recommended by Yang *et al.* (2011) [218] with exclusions made as follows: minor allele frequency <0.10, >0.5% missing genotypes per SNP, >1% missing genotypes per sample and HWE ($p<0.05$).

7.2.2 Mixed model analysis

Binary case/control phenotypes were adjusted for fixed and random effects using the following linear mixed model in ASReml [219]. A single categorical fixed effect was fitted which represents the division of samples into contemporary groups relating to the three most common reasons for surgery, other than OC, listed in Table 7-1 (angular limb deformities (ALD), fetlock chip(s) and sesamoid fracture(s)) and sex, resulting in $2^3 \times 2 = 16$ classes in total, 11 of which contained observations in

the final analysis (see 2.4 and A.iii for further details). A single random effect, animal, was fitted generating an individual animal model [220] in which the pedigree relationship matrix was replaced with a genomic kinship matrix in order to adjust for average allele sharing among sampled horses. Autosomal markers remaining after QC were used to generate the kinship matrix as follows: $f_{i,j} = \frac{1}{N}\sum_k \frac{(x_{i,k} - p_k)(x_{j,k} - p_k)}{(p_k(1-p_k))}$, where summation is across SNPs ($k=1,N$), x_{ik} is a genotype of the i^{th} horse at the k^{th} SNP coded as 0, ½, 1 and p_k is the frequency of the allele that is homozygous for the genotype coded as 1 [217]. On the diagonal, $f_{i,i} = 0.5(1+f_i)$, where f_i is the loss (or gain) of heterozygosity relative to the expectation. The kinship matrix describes the average relatedness between individuals and therefore controls for genetic stratification likely to be present in the sample. The transformation of the kinship matrix into a distance matrix followed by a multi-dimensional scaling (MDS) analysis [140, 221, 222], also allowed data to be inspected for the presence of outliers and substructure. Multi-dimensional scaling plots based on the first two principal components were considered with respect to farm of origin (see 2.5), sex and contemporary group. Following the implementation of the mixed model, a vector of approximately normally distributed ($N(0,1)$) residual errors replaced the binary (0,1) observation as the phenotype for testing in the GWAS.

Table 7-1 A description of conditions (other than OC) for which horses were treated. For further information, see 2.4.

Condition	No. of affected		
	Cases	Controls	Total
Angular limb deformity (ALD)	38	90	128
Fetlock chip(s)	36	71	107
Other chip(s)	3	3	6
Sesamoid fracture(s)	8	23	31
Other – bone related	4	1	5
Other – not bone related	7	3	10

7.2.3 Genome-wide association study

The GWAS was performed in GenABEL [217] using a score test for a Gaussian distributed trait and no covariates [223]. A genome-wide significance level was calculated by performing 10,000 permutations of the residual phenotypes against genotypes. Permutations were carried out within sex, and the 5% significance level empirically determined. Confirmation of the effects of SNPs found to be significant by this approach was carried out by fitting all such SNP genotypes (coded as 0, 1, 2) simultaneously as fixed effects in the original mixed model.

7.2.4 Evaluation of ECA3 haplotype blocks

A case/control haplotype association test (no covariates) with 10,000 permutations was performed in Haploview [203] for a 2Mb region containing a putative QTL identified in the single SNP GWAS described above. Haplotypes (determined by the default algorithm based on confidence intervals [224]) were tested for association with OCD status using a chi-squared test, with each haplotype configuration observed being tested in turn, generating a single test statistic per haplotype. This preliminary analysis was performed to supplement the single SNP GWAS after it had

been published. Unfortunately, time limitations meant that the next step of the analysis, which would have been incorporating the fixed effect of contemporary group, was not completed.

7.2.5 Testing previously published quantitative trait loci

Single-nucleotide polymorphisms selected to represent OC QTL detected in other studies were also tested by fitting them simultaneously as fixed effects in the mixed model. The QTL regions tested were based on GWAS results published in Lampe (2009) [99] and Komm (2010) [100]. These studies were performed on samples from HWB horses and it has been shown in a reference sample of more than 150,000 horses that the Thoroughbred contributes nearly 35% of this breed's genes [225]. Whilst these studies examined a range of OC phenotypes, I tested only QTL relevant to OC or OCD with fetlock and hock cases combined, as here (see Table A-4 for a list of QTL). Where SNP names or precise SNP locations were provided, the exact SNP was fitted in the mixed model with the exception of one case where the SNP was not typed in my sample, in this case the closest SNP was fitted in the mixed model (type A in Table A-4). In cases where only an approximate location was given, i.e. to the nearest 0.1Mb, current GWAS results for the region 1Mb upstream and 1Mb downstream were examined and the SNP with the smallest p-value fitted in the mixed model (type B in Table A-4). Finally, in cases where several SNPs within a region were listed as being significant, the same range was searched in the current GWAS analysis and the SNP with the smallest *p*-value fitted in the mixed model (type C and D in Table A-4). In order to assess their ability to enhance the model, all SNPs representing QTL were fitted simultaneously in the mixed model alongside contemporary group, SNPs found to be significant in the current GWAS and the genomic kinship matrix.

7.3 Results

7.3.1 Mixed model analysis

The genomic kinship matrix was calculated based on 30,554 autosomal SNPs that passed the stringent QC thresholds. The distribution of kinships between individuals in the sample is shown in Figure 7-1. Multidimensional scaling plots revealed no

obvious outliers or any genetic substructure relating to factors such as farm or contemporary group (data not shown). The fitting of a genomic relationship matrix (calculated by doubling the kinship matrix) in the mixed model resulted in an extremely small estimated genetic variance component ($<10^{-6}$) making it impossible to estimate trait heritability with any precision; estimates of random animal effects (estimated additive breeding values) were correspondingly small. The residuals generated for testing in the association study were influenced primarily by contemporary group and their distribution can be seen in Figure 7-2.

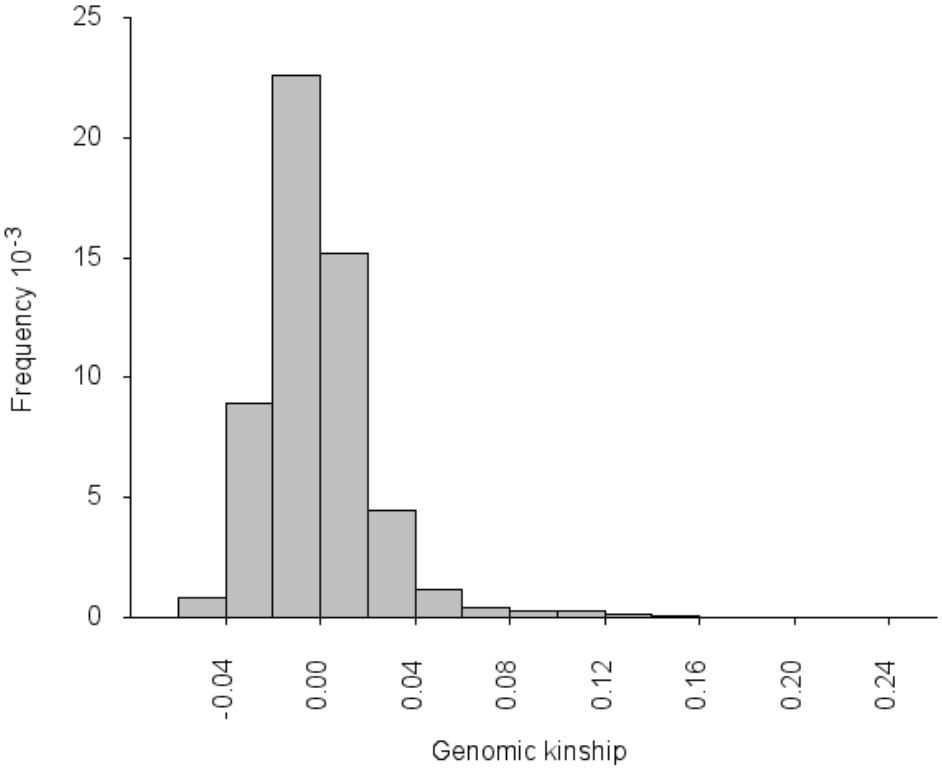

Figure 7-1 Distribution of genomic kinship between pairs of horses

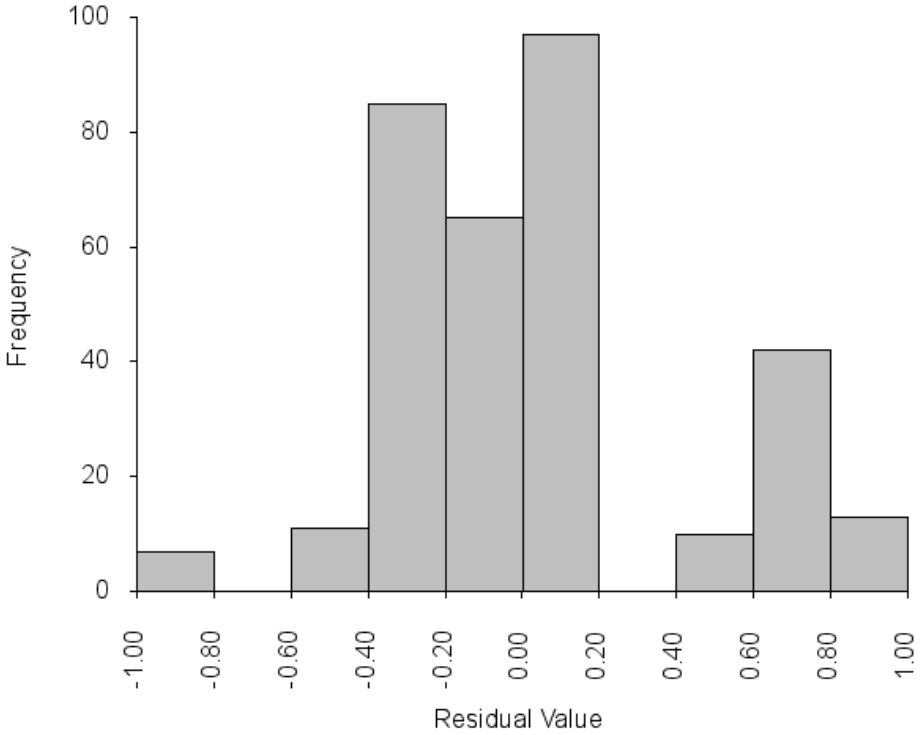

Figure 7-2 Distribution of residuals from mixed model analysis

7.3.2 Genome-wide association study

Following QC, 40,180 SNPs were tested for association; the mean minor allele frequency of remaining SNPs was 0.28 and the distribution of minor allele frequencies was approximately uniform. Based on empirical genome-wide significance ($p<2.91\text{x}10^{-6}$), a single SNP was found to be significantly associated with OCD as tested using residuals from the mixed model. This was SNP BIEC2-799865 located at 88,493,417bp on ECA3; this SNP has alleles C and T with a minor allele frequency (T) of 0.4 and conforms to a HWE genotype distribution (see Table 7-2 for genotype frequencies). Figure 7-3 shows a Manhattan plot of SNPs on ECA3. A haplotype block analysis of the region containing BIEC2-799865 revealed somewhat erratic linkage disequilibrium (LD) structure surrounding the SNP making the definition of an associated QTL region problematical (Figure 7-4). The apparent deviation from the expectation of decreasing LD with increasing distance between markers exhibited by BIEC2-799865 and its neighbours goes someway to explaining why this SNP stands apart from surrounding SNPs in Figure 7-3. With SNPs exhibiting r^2 [105, 106] with BIEC2-799865 of greater than 0.10 at distances up to 10Mb, I extended my search for other potentially associated SNPs within this range. A further four SNPs within 10Mb of BIEC2-799865 had $p<0.001$; two of these SNPs (BIEC2-799867 and BIEC2-799905) had r^2 of 0.45 – 0.55 with and were within 3 SNPs of BIEC2-799865 (see Figure 7-4), with the remainder being >4Mb away and having $r^2<0.10$. All four SNPs were located to the right of BIEC2-799865.

Fitting BIEC2-799865 as an additional covariate in the mixed model resulted in an estimated additive effect of -0.16 (±0.03), i.e. for every T allele an individual carries at the locus, that individual's probability of OCD is decreased by 0.16. This allows a crude estimate of the contribution of this SNP to the overall phenotypic variance to be made. Under the assumption of no dominance or interaction effects and using $V_A = 2p(1-p)\alpha^2$ [124] where p is allele frequency at the locus and α is the estimated SNP effect, BIEC2-799865 explains ~5% of the variance of OCD. The effect of BIEC2-799865 remained significant even when contemporary group was removed from the mixed model. Fitting the additional four SNPs with $p<0.001$ alongside contemporary group and BIEC2-799865 resulted in both BIEC2-799865

and one of the more distant SNPs (BIEC2-802230), having regression coefficients significantly different from zero.

7.3.3 Evaluation of ECA3 haplotype blocks

The 2Mb region surrounding BIEC2-799865 (1Mb up and down) contained 4 haplotype blocks as determined by the LD based algorithm of Gabriel *et al.* (2002) [224]; 24 of the 32 SNPs in the region were contained within one of these haplotype blocks (Figure 7-4). The mean number of haplotypes per block (with a frequency >0.01) was 4.5 (range 3 to 7) and the mean number of SNPs per block was 6.0 (range 4 to 10). A single haplotype in the fourth haplotype block (CCGATTAACC) had $p<0.10$, suggesting a possible association with OCD status.

7.3.4 Testing previously published quantitative trait loci

For each of the 24 QTL regions listed in Table A-4, a representative SNP was added to the mixed model containing contemporary group, BIEC2-799865 and the random effect of animal so that all SNPs were analysed simultaneously. This analysis resulted in only two of the 24 SNPs having a significant association with OCD. These SNPs were BIEC2-859811 on ECA4 (39,852,072bp), representing a QTL at 39.26Mb (Table A-4, QTL no. 8) [100] and BIEC2-410967 on ECA18 (36,772,271), representing a QTL between 36,408,881bp and 38,738,316bp [99] (Table A-4, QTL no. 16). BIEC2-799865 remained significant when fitted alongside the 24 QTL SNPs albeit with a slightly reduced size of effect (-0.11).

Table 7-2 Genotype frequencies of BIEC2-799865 and results of chi-square tests for association with OCD

	Genotype frequency			Total No. of Samples	*p-value from X^2 test*[I]
	C/C	C/T	T/T		
Controls	0.26	0.55	0.19	168	
Cases[II]	0.44	0.46	0.10	162	0.002
Hock cases	0.42	0.46	0.12	89	0.034
Stifle cases	0.48	0.44	0.08	50	0.008
Fetlock cases	0.44	0.46	0.10	41	0.062

[I]The chi-square tests compare each case category with the controls.

[II]Note, the number of cases is not equal to the sum of the cases in each joint location because some horses were affected in multiple joint locations.

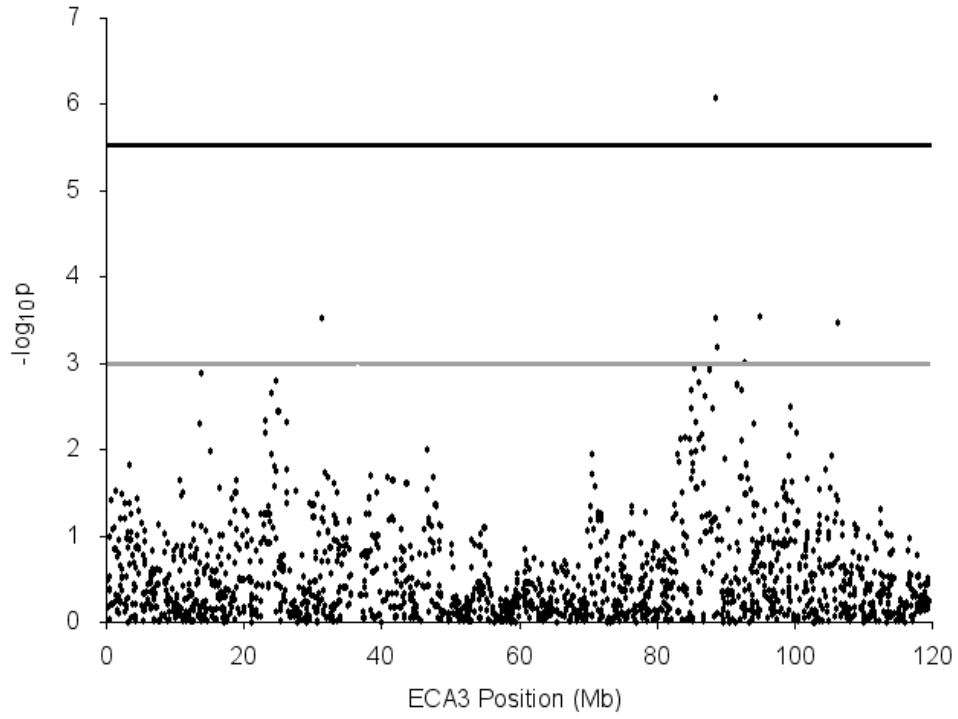

Figure 7-3 A Manhattan plot showing association results for ECA3. The black line represents the genome-wide significance level and the grey line represents the significance level used to identify surrounding SNPs with possible relevance.

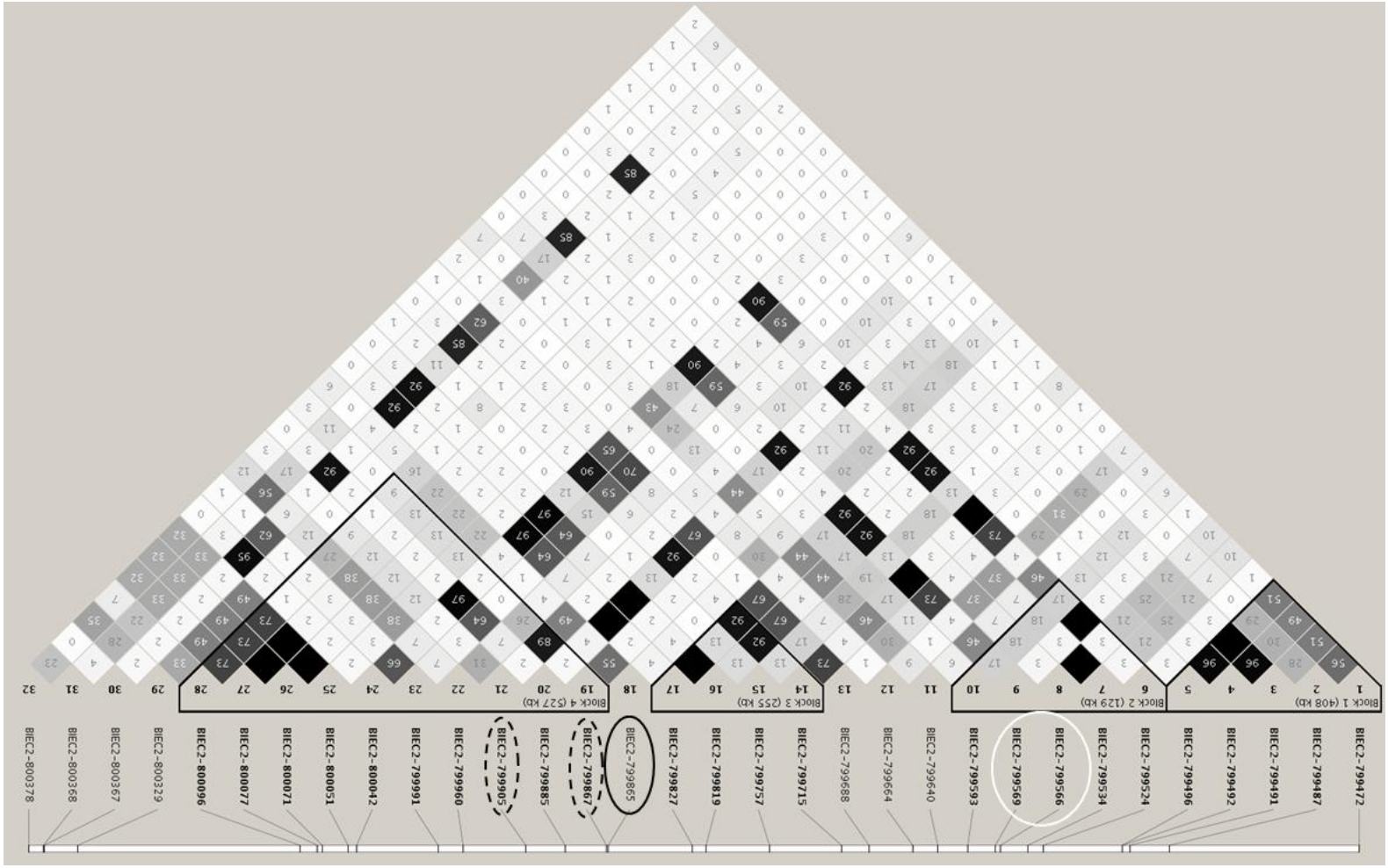

Figure 7-4 LD plot [203] of ECA3 region 1Mb either side of BIEC2-799865 (solid line, black circle). SNPs within the UGDH gene are indicated by a white circle. SNPs with a p<0.001 in the GWAS are indicated by a dashed circle. Marked haplotype blocks calculated according to Gabriel et al. (2002) [224].

7.4 Discussion

This GWAS in the Thoroughbred revealed a single SNP, BIEC2-799865 on ECA3, to be associated with OCD at a genome-wide level of significance when tested using the residuals from a mixed model analysis. Population genetics theory allows us to predict that, assuming the heritability for OCD is 0.15, this QTL accounts for ~34% of the genetic variation of the trait. However, effect estimates based on primary GWAS data have been shown to be upwardly biased, often to a large degree [226] and so a majority of the genetic variance underlying OCD remains to be captured. Two neighbouring SNPs showed an association with OCD which approached significance ($p<0.001$); the relatively lower MAF of these SNPs (0.27 and 0.25) compared to that of BIEC2-799865 (0.4) may explain their failure to reach genome-wide significance. The lack of LD around BIEC2-799865, means that the much sought after and characteristic GWAS peak is not observed in this case.

There are several reasons for considering haplotypes as well as individual SNPs when searching for variants associated with disease risk and Clark (2004) [227] discusses this in detail. Most importantly, haplotypes represent more accurately the units of inheritance and testing haplotype associations is an efficient way of exploring interactions between SNPs [92]. For instance, in cases where an associated SNP appears on several haplotypes, one of which is strongly associated with disease, haplotype analysis will result in a more significant result. Here, haplotype analysis was also intended to guide selection of a region for the investigation of candidate genes. Whilst in this analysis BIEC2-799865 was found to lie between haplotype blocks, four blocks were identified within 1Mb of the SNP. One of the seven haplotypes observed in the fourth block (containing the two aforementioned neighbouring SNPs) was associated with OCD status at a suggestive level of significance ($p<0.10$), but, the single SNP association tests for BIEC2-799865, BIEC2-799867 and BIEC2-799905 (under the same model) showed greater significance. Whilst this could be an indication that, in this instance, there was little to be gained by considering haplotypes, it could also be due to the fact that contemporary group was not fitted in the haplotype model; this would need to be done to confirm this result. Genotyping additional markers in the region may help to

resolve haplotype block structure in the immediate vicinity of BIEC2-799865 in the future. Whilst the implication of these results on the validity of the association is not clear, it does impact on my ability to precisely define a corresponding QTL region for further evaluation. For the purposes of candidate gene discovery I chose to examine the region 1Mb either side of the SNP.

The 2Mb window surrounding BIEC2-799865 contained 22 labelled genes, 21 of which are described as protein coding and one of which is labelled as a pseudo gene. Whilst according to the current annotation BIEC2-799865 lies between genes, LOC100064680 located at 88,494,283 - 88,511,285bp, contains (within an intron) BIEC2-799867; this SNP is adjacent to and most highly correlated with BIEC2-799865 and falls within haplotype block 4. This gene is described as being similar to basic kruppel like factor and studies in mice and *C.elegans* show orthologues to this gene, *kruppel-like factor 3 (basic)* (*KLF3*), to be involved in adipogenesis [228, 229]. More generally, KLFs have been described as DNA binding transcriptional regulators that play diverse roles during differentiation and development [230]. Whilst the likely function of *KLF3* does not preclude its relevance, there is no evidence of a direct role for this gene in OC. This was true of most of the genes located within the QTL region defined, with the exception of *UDP-glucose dehydrogenase* (*UGDH*).

The *UGDH* gene (located at 87,818,121 – 87,843,937bp) appears to function in the regulation of glycosaminoglycan (GAG) synthesis in cells lining the articular cartilage surface [231]. These GAGs are involved in extra-cellular matrix integrity, playing a crucial role in chondrogenesis, homeostasis and compressive resilience [231]. A potential link between GAGs and OC has been demonstrated by the observation of differential levels of GAGs in osteochondritic lesions versus healthy cartilage [232, 233]. However, the direction of causality is not clear and several other studies have observed no significant difference [234, 235]. Two SNPs located within introns of *UGDH* were not significantly associated with OCD ($0.05< p<0.10$) and nor were any of the haplotypes in block 1 (Figure 7-4). One of the SNPs did, however, show moderate LD (r^2=0.1–0.2) with BIEC2-799865 and the two neighbouring SNPs mentioned above; as before, the relatively lower MAF of this

SNP (0.34) may have prevented it from appearing above the background in terms of significance. The second SNP in *UGDH* had a MAF of 0.06 and therefore provides little information about either association or LD. Whilst the distance of the *UGDH* gene from BIEC2-799865 and the relatively low LD in the region lead me to question its relevance, there are likely to be many untyped variants in this region, some of which could plausibly have stronger LD with BIEC2-799865.

Four previous GWAS for OC in the horse have also identified QTL on ECA3 [99, 100, 102, 236]. The closest to BIEC2-799865 was presented recently in a study carried out in French Trotters and is located at 100-110Mb [101]. The relatively close proximity of the two QTL represents some correspondence between studies. However, with average LD at this distance (~12Mb) being $r^2<0.02$ (Chapter 3), it is also possible that these QTL represent two different underlying genetic variants.

Adding SNPs to represent previously identified QTL to the model (which included BIEC2-799865) resulted in two out of 24 SNPs tested having regression coefficients significantly different from zero ($p<0.05$) and therefore showing the potential to enhance the fit of the model. On ECA4, BIEC2-859811 (39,852,072bp) had a regression coefficient of -0.102 (±0.049). Komm (2010) [100] identified six candidate genes located between 37.1Mb and 44.7Mb. On ECA18, BIEC2-410967 (36,772,271bp) had an estimated effect size of -0.085 (±0.042). Lampe (2009) [99] identified three candidate genes in the vicinity of the QTL corresponding to this SNP. These apparent validations should however be viewed with caution since adjustments to the mixed model, for example the removal of BIEC2-799865, lead to different QTL being significant and I was therefore unable to unambiguously confirm any of the previous QTL in the current dataset.

There are several reasons for the poor correspondence between QTL studies of OC in the horse. Firstly, the QTL which have been identified to date may be false positives [42]. Alternatively, subsequent studies may have been underpowered to detect them. In this case, such results may be due to, for example, differences in phenotypic definition or population ancestry. Ideally, replication studies should involve precisely the same allele or haplotype, the same phenotype and the same genetic

model as the original signal [237]. In this study, by testing only the QTL regions associated with OC under the combined phenotype definition (hock and fetlock) used by Lampe (2009) [99] and Komm (2009) [100], the difference in phenotypic definition between the three studies was minimised.

Another reason for the lack of correspondence may be breed differences. Hamann *et al.* (2008) [225] estimated that 35% of the HWB genes came from Thoroughbred lines, but it is not known what the proportion was in the Komm (2010) [100] and Lampe (2009) [99] sample of 154 foals. Assuming the same QTL are controlling the genetic predisposition to OC in both breeds, differences in allele and haplotype frequencies between breeds will impact on the proportion of variance the QTL explain and therefore on our ability to detect them. Furthermore, with no standardised method either for reporting QTL, or for carrying out validation studies, the approach taken here to select SNPs for testing in the mixed model, was largely subjective and I may have missed more appropriate SNPs.

Despite being one of the largest GWAS of OC in horses performed to date, the principal limitation of this study remains a lack of power. This lack of power is evidenced by both the low number of genome-wide significant SNPs and the very small estimated genetic component. Whilst disappointing, the inability to estimate heritability in this sample is perhaps not surprising given the relatively large standard errors which accompany some of the heritability estimates for OC to date [77, 86]. Furthermore, my findings do not necessarily rule out a non-zero heritability, rather more data is needed to produce a reliable estimate.

The explanation for the apparent low power of this study is likely to be multifaceted. Firstly, since power is directly related to sample size, the relatively small number of horses genotyped for this study will have limited the number of identifiable QTL, as shown by power calculations of, for example, Wang *et al.* (2005) [48]. Secondly, phenotypic definition can play an important role in determining the power of GWAS of complex diseases. Optimal phenotypic definitions are those with strict inclusion criteria, with minimising genetic heterogeneity between cases being a useful way of increasing study power [42]. Unfortunately, OC represents a clinically complex

phenotype, affecting multiple joints and predilection sites within joints, as well as appearing in a variety of different forms. Just as prevalence and heritability estimates for OC have been affected by this problem, so we can expect QTL mapping studies to be. In this study, by considering exclusively those cases with fragments present (OCD), the genetic heterogeneity of the cases has been reduced and recommendations by van Grevenhof *et al.* (2009) [87] that flattened bone contours and fragments should be evaluated as statistically different disorders have been followed.

Several studies to date have considered further subdivision of OC cases by joint affected, resulting in different QTL being identified for each subgroup [97, 98]. This is appealing given the apparent low correlation among the occurrence of lesions of OC in different body locations [87, 95, 238] and the corresponding idea that OC is in fact a localised disease [68]. However, subdividing cases in this way represents a significant loss of power. Furthermore, testing several manifestations of the disease serves to exacerbate the already serious problem of multiple testing. For this reason and from a practical selection perspective, expressing OC as a single trait is more appealing, and should enable the identification of QTL controlling more generalised factors.

In this study, model complexity due to the presence of horses suffering from conditions other than OC in the cohort may have reduced the power of the association test. The uneven representation of cases and controls across the contemporary groups describing the presence or absence of ALD, fetlock chips and sesamoid fractures in the samples, represented a potential cause of bias in the sample and therefore had to be fitted in the model. In the event that none of these conditions are related to OC or have a hereditary component, the adjustment for contemporary group represents a loss of power through the reduction in the number of degrees of freedom of the model. However, in the case where one or more of these diseases has a hereditary component (of which there is some evidence [86, 88]), the exclusion of contemporary group from the model would result in severe confounding. Since the latter is by far the more serious case, I chose to fit contemporary group in the mixed model.

However, there is seemingly a trade-off to be made. Whilst the use of clinical data in this case added complexity and potentially noise to the data, it also gave me increased confidence in the phenotypic classifications of OCD. In this study, all of the cases underwent arthroscopy, the so-called 'gold standard' of diagnosis of cartilage defects [239] and so we can be confident of high specificity. All of the controls had OC ruled out through a comprehensive radiographic survey of predilection sites and the evaluation of radiographs by a specialist in the field (LRB) significantly reduced the chance of OC going undiagnosed.

In this GWAS, a single SNP associated with OCD in a sample of 330 Thoroughbreds was identified. This association requires validation in an independent dataset in order to rule out the possibility that it represents a false positive association. In the event that the SNP is validated, further fine-mapping and re-sequencing of the region will be needed in order to elucidate the causal mutation behind this association. The likely issue of poor power to detect QTL in this study illustrates both the complexity of OC and the challenge faced by members of the equine genetics community in collecting and genotyping sufficiently large samples for effective GWAS to be carried out. This complexity and the seemingly low heritability of OC suggest genome-wide evaluation might be a more efficacious approach to selection than MAS and this hypothesis is explored in the following chapter.

Chapter 8: The use of genome-wide evaluation in the prediction of risk for osteochondritis dissecans in the Thoroughbred

8.1 Introduction

The genetic improvement of complex traits in livestock has traditionally been achieved through the selection of animals for breeding based on the performance of the individual, its contemporaries and its relatives. Through this approach, formalised in Best Linear Unbiased Prediction (BLUP) [220], huge gains have been made, particularly with respect to production traits such as milk volume per cow [240]. However, improvements in genotyping technology have paved the way for more informed breeding decisions based on the actual gene variants an animal is carrying rather than an expectation of what gene variants it is likely to be carrying. Initially, research was focused on identifying specific markers (or genes) which were associated with traits of interest and could then be used in marker (or gene) assisted selection (MAS/GAS). Whilst this approach has often proved to be successful when applied to simple traits, such as monogenic diseases, its success in the arena of complex traits has been much more limited. The principal reason for this limited success is that, in the case of complex phenotypes, there appear to be few genetic variants which explain a sufficiently large proportion of the total genetic variance of the trait to be useful for selection [92].

The concept of genome-wide evaluation (GWE) was first introduced in the context of selection by Visscher & Haley (1999) [51]. In many ways, GWE represents a reversion to the black box approach of BLUP, focusing on the average effect of all gene variants, rather than attempting to identify individual gene variants of significance. Meuwissen *et al.* (2001) [52] then showed that, with sufficiently dense markers and provided phenotypes had been available, it was possible to estimate an individual's breeding value from its genotype alone and so the concept of genomic selection (GS) was borne. New single-nucleotide polymorphism (SNP) genotyping technology has since delivered the required marker density in horses and other livestock species. The methodology involves first simultaneously estimating all SNP effects in a training population for which both phenotype and genotype data are

available. Essentially a multiple-linear regression problem, this step is described as a 'small *n*, large *p*' problem, with the number of SNP effects to be estimated (*p*) always outnumbering (and in many cases considerably so) the number of observations from which estimations can be made (*n*). The development of the optimal statistical solution to this problem has been the focus of many studies over the last ten years and yet remains largely open to debate. Commonly used statistical models and approaches include multiple linear regression (typically with some kind of pre-selection of SNPs) [52], Bayesian procedures [52] and semi-parametric specifications [56, 62]. The second stage of the process involves the estimation of additive genetic values or genomic estimated breeding values (GEBV) in a test population (usually the selection candidates) for which only genotype information is available, through the application of the 'SNP key' developed in the training population.

The accuracy of the GWE methodology is dependent on characteristics of the trait, such as heritability, properties of the genome, such as the extent of linkage disequilibrium (LD), and the coverage of the genome offered by the markers [61]. A key benefit of GS is that, whilst the accuracy of BLUP estimated breeding values is theoretically limited when based on individuals and ancestors, only being able to approach one with progeny or inbreeding, GEBV accuracy is not [241]. However, the advantages of GS over traditional approaches to selection extend beyond simply generating highly accurate breeding values. The availability of GEBV from birth offers the potential for greatly reduced generation intervals, perhaps even velogenetics [242]. The implementation of GS also eliminates the need for phenotyping every generation, which is particularly useful in the case of traits which are difficult or expensive to measure. Furthermore, in the case of sex specific traits, e.g. milk production, GEBV can be produced to equal accuracy for all animals.

Whilst the benefits of GS have the potential to apply to all livestock populations, its implementation has been recognised as being particularly useful in the dairy cattle industry. There are several factors that contribute to making implementation of GS a much more realistic risk-free prospect in this industry compared to in the more fragmented industries of the beef cattle and sheep sectors, these include: (1) the

structure of the industry, which consists of several large breeding companies who produce the elite breeding stock; (2) the domination of the industry by a single breed of small effective population size (Holstein); (3) the large proportion of costs and long generation intervals associated with traditional progeny testing schemes; and (4) the easy identification of cost savings for genotyping investments. The equine industry, at least within the UK and notwithstanding the economic domination of the Thoroughbred, is more akin to the beef and sheep industries and therefore will likely suffer similar challenges in the implementation of GS. These challenges include the need to produce GEBV across a large number of breeds (and crossbreds) and the coordination of the scheme across a number of disparate groups. However, since little genetic information is currently on offer to breeders, even relatively low accuracy GEBV would represent considerable progress and, in principle, should increase genetic gain.

In this study, the concept of GWE is applied to osteochondritis dissecans (OCD) expressed as a binary trait in the Thoroughbred. The aim is to exploit the methodology in order to derive GEBV for the genetic risk of OCD in horses. In addition, a trait was simulated by adding SNP effects to the available genotype data and then analysed in order to explore the power and other properties of the available data.

8.2 Materials and methods

8.2.1 Samples

The data for this study comprised the OC samples that were sourced in the US (n=348 Thoroughbreds from the US), genotyped at the 50,707 SNPs that passed preliminary quality control (QC). As in Chapter 7, horses were classified into contemporary groups relating to the three most common reasons for surgery, other than OC (angular limb deformities, fetlock chip(s) and sesamoid fracture(s)) and sex, resulting in 2^3 x 2 = 16 classes in total, 11 of which contained observations in the final analysis (see 2.4 and A.iii for further details). Contemporary group was subsequently fitted in models as a categorical fixed effect.

8.2.2 Quality control

Samples were checked for sex discrepancies (marker-based prediction of sex versus sample label) and intermediate X-chromosomal inbreeding ($0.2<F<0.8$), with exclusions being made on the basis of suspected sampling or genotyping errors. Two samples were excluded due to sex discrepancy and 16 based on indeterminate sex as demonstrated by intermediate inbreeding, leaving 168 controls and 162 cases for further analysis. Further, 30 SNPs were excluded on the basis that they were positioned on the X chromosome but were likely autosomal.

Only autosomal SNPs were used in the analyses. To be eligible for selection as a QTL in the simulation, SNPs had to be polymorphic and genotyped in at least 95% of samples; there was no restriction on minor allele frequency (MAF). Prior to the estimation of SNP effects (both for simulated and OCD phenotypes), the following thresholds were used for excluding SNPs: MAF <0.05 and >5% missing genotypes per SNP. No exclusions were made on the basis of Hardy-Weinberg equilibrium (HWE). In the OCD dataset, additional exclusions were made on the basis of differential proportions of missing SNPs between cases and controls ($p<0.05$). In the simulated dataset, SNPs which were assigned as QTL (see below) were also removed prior to genomic analysis. Finally, pairwise LD between remaining SNPs was assessed (calculated using an EM algorithm as in Chapter 3) and in cases where $r^2=1$, one of the pair excluded. Following these exclusions, 56% of the original SNP set remained to be used in the genome-wide evaluation.

8.2.3 Simulation of additive genetic merit and phenotypic performance

A trait with a target heritability of 0.20, representing a typical OC estimate, was simulated[9] using the genotypes from the sample described above. The target number of QTL was 1,000 and was chosen by rounding up from the number of independent

[9] The simulation program used was adapted from a program written by Hans Daetwyler (then of The Roslin Institute, now of the Victorian Department of Primary Industries).

segments in the genome, such that there would be at least as many QTL as independent segments in the genome (see 8.4). Each of the 48,387 SNPs were chosen as QTL in independent Bernoulli trials with a probability of 0.021 (1,000/48,387), resulting in 984 QTL. Each SNP had an equal chance of being selected, regardless of MAF, resulting in the distribution of MAF of the QTL following that of the SNP chip. SNPs designated as QTL were not used in the subsequent estimation of GEBV. True allele substitution effects (α_i) were sampled from $N(0,1)$. True genetic values for the 330 individuals were calculated as the sum of the average allele effects, calculated as $(1-p_i)\alpha_i$ for A_1 and $-p_i\alpha_i$ for A_2 (where p_i is the allele frequency of A_1) across all QTL. True genetic values were then scaled to approximate a $N(0,0.2)$ distribution. Residual effects for each animal were obtained by sampling from a $N(0,0.8)$ distribution. Finally, phenotypic records were simulated by adding residual effects to the simulated true genetic values. The final genetic and residual (environmental) variances were 0.20 and 0.73, respectively, giving a heritability of 0.21 on the continuous (liability) scale. Samples were then ordered according to their simulated phenotype. The bottom 50% of the distribution (n=165) were labelled as controls and assigned a binary phenotype value of 1. The top 50% of the distribution (n=165) were labelled as cases and assigned a binary phenotype value of 2. This designation of samples as cases/controls led to a dichotomous trait with 50% prevalence and heritability on the observed scale of 0.13.

8.2.4 Bayesian estimation of SNP effects

The program described in this section was written by Ricardo Pong-Wong. Marker effects were estimated using a BayesB type method similar to that described by Meuwissen *et al.* (2001) [52]. The model applied was: $\mathbf{y} = \boldsymbol{\mu}\mathbf{1} + \mathbf{Xb} + \sum_{\mathbf{i=1}}^{\mathbf{m}} \mathbf{g_i}\alpha_i + \mathbf{e}$, where $\mathbf{y}$ is the vector of phenotypes; $\mathbf{b}$ contains the fixed effect of contemporary group and $\mathbf{X}$ is its incidence matrix (in OCD dataset only); α_i is the allelic substitution effect for SNP i; $\mathbf{g_i}$ is the vector of genotypes (0, 1 or 2 for homozygote for the minor allele, heterozygote and homozygote for the major allele, respectively)

for SNP *i* and **e** is the vector of residuals distributed $N(0,\sigma_e^2)$. Since not all SNPs are expected to affect the trait, the allelic substitution effect, α_i, for each SNP was sampled from a mixture distribution with probability π of having an effect on the trait and with probability (1-π) of not having an effect on the trait. The SNP effects were then sampled from $N(0,\sigma_{snp}^2)$ such that the analysis might be referred to by some as a BayesC implementation [243].

The implementation of the model was done using Gibbs sampling. The parameter σ_e^2 was estimated in the analysis using a scaled inverse chi-squared distribution with $\upsilon = 2$ degrees of freedom (*df*) and scale parameter $\sigma^2 = 1$. The parameter σ_{snp}^2 was estimated in the analysis using a scaled inverse chi-squared distribution with $\upsilon = 2$ *df* and scale parameter $\sigma^2 = 0.001$. This represents a weak informative prior with a wide distribution. Two analyses were performed. In the first, π (the proportion of SNPs with an effect) was estimated during the analysis using a flat prior (BayesCπ). In the second, π was assumed to be known (BayesC) such that for the simulated data, π was set to twice the number of simulated QTL divided by the average number of SNPs used in the Bayesian analysis (see 8.4), calculated as 0.0693, and, in the case of the OCD data, π was set arbitrarily to 0.10.

For each analysis, a Monte Carlo Markov chain (MCMC) was run. The first 10,000 cycles were discarded as burn-in. This was followed by 30,000 realisations, with 50 cycles between each consecutive realisation. The posterior mean was used as the estimate for each parameter of interest. GEBV were generated for every individual in the training and reference samples by the summation across all markers of the product of the mean estimated substitution effect of the marker and the genotype value (0, 1 or 2). The estimated total genetic variance explained by the SNPs was calculated as the sum of the variance of the GEBV and the average predicted error variance.

8.2.5 Genomic BLUP procedure

Hayes *et al.* (2009) [244] state that the replacement of the average relationship matrix derived from pedigree by the realised (genomic) relationship matrix in BLUP

of breeding values is the equivalent to the GWE methodology where the effects of QTL contributing to variation in the trait are assumed to be normally distributed with constant variance. Using the same SNP markers that were used for the Bayesian estimation of marker effects, marker derived relationships were calculated using a program written by Ricardo Pong-Wong. Relationships were calculated as: $f_{i,j} = \frac{1}{N}\sum_k \frac{(x_{i,k} - 2p_k)(x_{j,k} - 2p_k)}{2(p_k(1-p_k))}$, where, summation is across SNPs ($k=1, N$), x_{ik} is a genotype of the i^{th} horse at the k^{th} SNP coded as 0, 1, 2 and p_k is the frequency of the allele that is homozygous for the genotype coded as 2 [217]. On the diagonal, $f_{i,i} = \frac{E_i - O_i}{E_i}$, where E_i is expected heterozygosity adjusted for T, the total number of alleles ($2N$), and calculated as: $E_i = \sum_{k=1}^{N} 1 - 2p_k q_k \left(\frac{T_k}{T_k - 1}\right)$, and O_i is observed heterozygosity. A single trait animal model was then fitted in ASReml [219] applying the model: $\mathbf{y} = \boldsymbol{\mu}\mathbf{1} + \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$, where $\mathbf{y}$ is the vector of phenotypes; $\mathbf{b}$ contains the fixed effect of contemporary group (in the OCD dataset only) and $\mathbf{X}$ is its incidence matrix; $\mathbf{u}$ is the vector of animal effects and $\mathbf{Z}$ is its incidence matrix and $\mathbf{e}$ is the vector of residuals distributed $N(0, \sigma_e^2)$. The pedigree based numerator relationship matrix was replaced with the marker-based genomic relationship matrix in the model. GEBV were generated for all individuals by solving the above linear mixed model equation using REML (restricted maximum likelihood). This approach is referred to as GBLUP.

8.2.6 Cross-validation procedure

Cross-validation (CV) is a robust, non-parametric technique for model evaluation [245]. The method consists of splitting the data into a number of training or reference datasets and validation or test datasets. Model parameters can then be estimated in the reference (training) datasets and used to predict observations in the test (validation) datasets. A five-fold CV procedure was used to assess accuracy of predicted GEBV. In the case of the simulated phenotype, cases and controls were randomly allocated to one of five subsets, such that each subset contained 33 cases and 33 controls. In the case of the OCD phenotype, samples were randomly

allocated to one of five subsets in such a way that both the number of cases and controls, and the number of samples in each contemporary group were approximately equal in each subset. Phenotypes from each subset in turn were masked (set to missing) and the remaining subsets used to predict parameters of the model in order to generate GEBV for all samples.

8.2.7 Evaluation of the accuracy of genomic estimated breeding values

The standard approach to assessing genomic evaluation models is to calculate the correlation between the true and estimated breeding values, denoted here as $r_{g\hat{g}}$, and referred to as the accuracy of prediction. The value of $r_{g\hat{g}}$ was calculated directly for the simulated dataset. However, the true breeding values for horses in the OCD dataset were unknown and therefore two alternative approaches were taken to assess the accuracy of the GEBV. The fixed effect, contemporary group, was a nuisance factor (fitted only to avoid confounding and not for its predictive ability) and therefore its contribution to the estimated phenotypes had to be discounted. For the first approach, this was achieved by obtaining a 'corrected true phenotype', calculated for each horse by subtracting the mean and fixed effect of contemporary group estimated in the model (either Bayesian or GBLUP) from its true phenotype. The corrected true phenotype therefore consisted of the genetic and residual components combined. The correlation between this corrected true phenotype and the GEBV was then calculated to give a proxy measure for $r_{g\hat{g}}$, denoted here as $r_{p\hat{p}}$. This proxy measure of accuracy could then be compared to expected accuracies according to the equations of Daetwyler *et al.* (2008) [61] (see 8.4).

The second approach taken to evaluate GEBV accuracy, involved the use of a permutation procedure to assess the significance of the contribution of the GEBV to phenotype prediction. Initially, the correlation between the true phenotype and the predicted phenotype was calculated in each test set, this is denoted $r_{y\hat{y}}$ and is referred to as the predictive ability of the model. Subsequently, GEBV were randomised amongst individuals of each test set and the correlation calculated again.

This was repeated 1,000 times in order to generate the null distribution of $r_{y\hat{y}}$ values, enabling the empirical determination of a 5% significance level in order to assess the contribution of the GEBV to the total predictive ability of the model. $r_{y\hat{y}}$ was also calculated for the simulated dataset as a means of evaluating the impact of using proxies such as $r_{p\hat{p}}$ and $r_{y\hat{y}}$ in place of the correlation between true and estimated additive genetic values ($r_{g\hat{g}}$) in the OCD dataset.

The assessment statistics described above ($r_{g\hat{g}}$, $r_{p\hat{p}}$ and $r_{y\hat{y}}$) were calculated for each CV test set in turn. Test sets were then combined and the same assessment statistics calculated to give overall performance. Assessment statistics were also calculated in reference sets for comparative purposes. For the remainder of this chapter, $r_{g\hat{g}}$ will be referred to as the 'accuracy of prediction' and $r_{y\hat{y}}$, as the 'predictive ability' of the model.

8.3 Results

8.3.1 Osteochondritis dissecans trait results

8.3.1.1 Bayesian estimation of SNP effects

The mean predictive ability of the BayesCπ model ($r_{y\hat{y}}$) was 0.73±0.01 and 0.60±0.02 in the reference and the test datasets, respectively. Predictive abilities in the test sets ranged from 0.58 to 0.64 (Table 8-1) and a comparison of these calculated values with the distribution of $r_{y\hat{y}}$ calculated from the permutation of GEBV, showed the improvement in fit derived from the GEBV not to be significant (at the 5% level) in any of the test sets. Figure 8-1 shows the distribution of $r_{y\hat{y}}$ that resulted from the permutation for Test Set 1. The improvement was also not significant when the test sets were combined. Similarly, when the results were assessed by $r_{p\hat{p}}$, none of the test sets exhibited a significant positive correlation between corrected true phenotype and GEBV. Using BayesCπ, an average of 0.6% of all SNPs were fitted in the model. A paired two-sample t-Test revealed that fixing the proportion of SNPs with an effect in the model to 10% did not significantly

improve the fit of the model (comparing $r_{y\hat{y}}$) in the test sets (Table 8-2), but did increase the mean correlation between the observed and the estimated phenotypes in the reference sets (data not shown). The mean estimated total genetic variance explained by the SNPs was 0.021±0.003 and 0.075±0.002, for the BayesCπ and BayesC models, respectively.

8.3.1.2 Genomic BLUP Procedure

The genomic BLUP estimate for the genetic variance component was fixed near the boundary in all CV sets, implying a very small genetic variance and resulting in GEBV in the range 10^{-7}. Therefore, estimated phenotypes were almost entirely determined by the fixed effect (contemporary group) and the permutation procedure showed no additional benefit could be gained by including GEBV (data not shown). When the results were assessed by $r_{p\hat{p}}$, none of the test sets exhibited a significant positive correlation between corrected true phenotype and GEBV. The relative performance within each test set was similar to that seen in the Bayesian analysis (Figure 8-2).

Table 8-1 OCD: BayesCπ

Test set	Correlation between corrected true phenotypes and GEBV		Correlation between predicted and observed phenotypes	
	$r_{p\hat{p}}$	***p*-value**	$r_{y\hat{y}}$	**5% sig. level**
1	0.15	0.23	0.593	0.596
2	0.03	0.82	0.598	0.610
3	0.03	0.80	0.592	0.606
4	-0.01	0.91	0.639	0.656
5	0.00	0.97	0.575	0.587
Overall	0.04	0.50	0.599	0.603

Table 8-2 OCD: BayesC (π=0.10)

Test set	Correlation between corrected true phenotypes and GEBV		Correlation between predicted and observed phenotypes	
	$r_{p\hat{p}}$	***p*-value**	$r_{y\hat{y}}$	**5% sig. level**
1	0.18	0.14	0.604	0.609
2	0.03	0.79	0.590	0.621
3	-0.01	0.95	0.561	0.608
4	0.00	0.97	0.617	0.659
5	-0.03	0.84	0.549	0.584
Overall	0.04	0.46	0.584	0.595

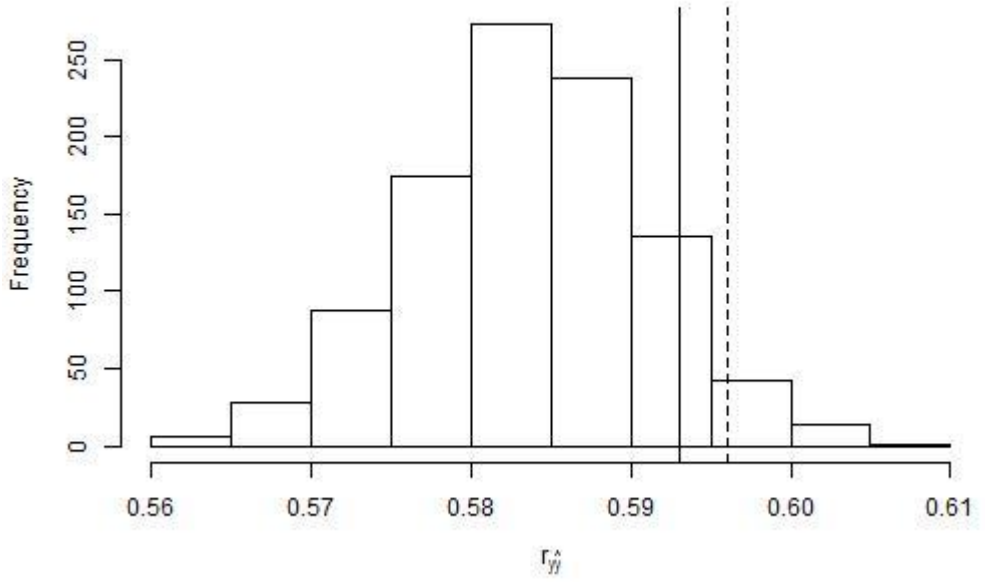

Figure 8-1 The distribution of correlations between predicted and true phenotypes ($r_{y\hat{y}}$) following the permutation of GEBV amongst individuals. Data for Test Set 1 (BayesCπ). The solid vertical line indicates the true correlation obtained and the dashed line the 5% significance threshold.

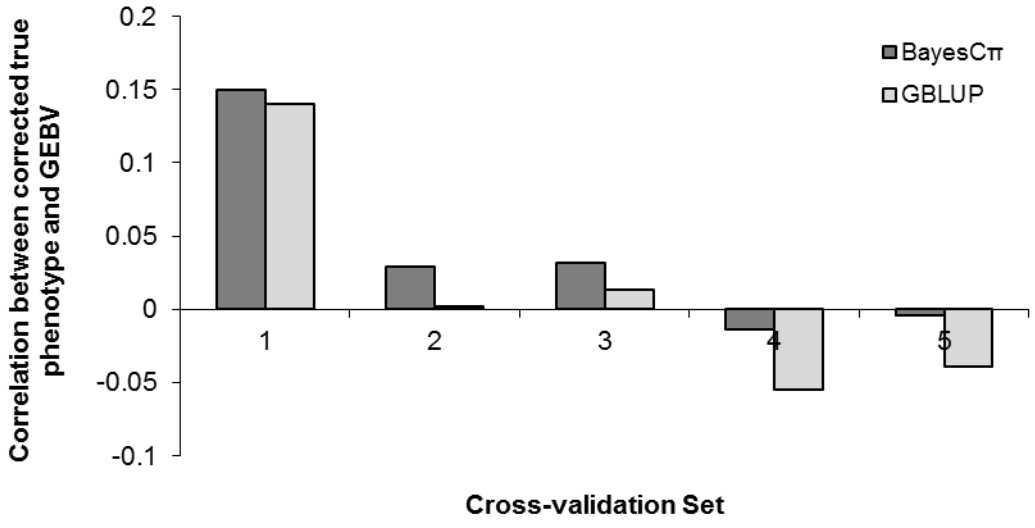

Figure 8-2 Correlation between corrected true phenotypes and genomic estimated breeding values ($r_{p\hat{p}}$) for BayesCπ and for GBLUP methods

8.3.2 Simulated trait

8.3.2.1 Bayesian estimation of SNP effects

The mean predictive ability of the BayesCπ model ($r_{y\hat{y}}$) was 0.93±0.02 and 0.16±0.07 in the reference and the test datasets, respectively. The predictive ability in the test sets ranged from 0.059 to 0.26, but the correlation was significant only in test set 5 (Table 8-3). A predictive ability, $r_{y\hat{y}}$, of 0.16 (p<0.005) was calculated when all test set data was combined. When the accuracy of prediction was calculated as the correlation between true and estimated breeding values ($r_{g\hat{g}}$), three CV sets showed a significant positive correlation and when estimates were combined across all test sets, a significant positive correlation of 0.19 (p<0.005) was calculated. On average, 2.4% of all SNPs were fitted in the BayesCπ model. Fixing the proportion of SNPs with an effect in the model to 6.93% (BayesC) did not significantly improve the accuracy of the GEBV (based on results of a paired two-sample t-Test) (Table 8-4) but did slightly increase the correlation between the true and the estimated phenotypes in the reference sets (data not shown). The mean estimated genetic variance was 0.08±0.01 and 0.11±0.007, for the BayesCπ and BayesC models, respectively.

8.3.2.2 Genomic BLUP

The mean predictive ability of the GBLUP model ($r_{y\hat{y}}$) was 0.92±0.04 and 0.13±0.08 in the reference and the test datasets, respectively. The predictive ability in the test sets ranged from 0.023 to 0.24, but was not significant (p<0.05) in any of the individual test sets (Table 8-5). A predictive ability, $r_{y\hat{y}}$, of 0.11 (p< 0.05) was calculated when all test set data was combined. When the accuracy of prediction was calculated as the correlation between true and estimated breeding values ($r_{g\hat{g}}$), three of the CV sets showed a significant positive correlation and when estimates were combined across all test sets, a significant positive correlation of 0.16 (p<0.005) was calculated. The mean heritability estimate on the observed scale was 0.29±0.12.

Table 8-3 Simulation: BayesCπ

Test set	Correlation between true and predicted phenotypes		Correlation between true and predicted breeding values	
	$r_{y\hat{y}}$	***p*-value[I]**	$r_{g\hat{g}}$	***p*-value[I]**
1	0.13	0.28	0.17	0.16
2	0.18	0.15	0.26	**0.03**
3	0.18	0.16	0.37	<**0.01**
4	0.06	0.64	-0.11	0.36
5	0.26	**0.04**	0.30	**0.01**
Overall	0.16	<**0.01**	0.19	<**0.01**

[I]**Significant results (*p*<0.05) shown in bold**

Table 8-4 Simulation: BayesC (π=0.693)

Test set	Correlation between true and predicted phenotypes		Correlation between true and predicted breeding values	
	$r_{y\hat{y}}$	***p*-value[I]**	$r_{g\hat{g}}$	***p*-value[I]**
1	0.19	0.12	0.19	0.12
2	0.17	0.17	0.26	**0.03**
3	0.17	0.16	0.36	<**0.01**
4	0.06	0.63	-0.12	0.35
5	0.24	**0.05**	0.30	**0.01**
Overall	0.16	<**0.01**	0.20	<**0.001**

[I]**Significant results (*p*<0.05) shown in bold**

Table 8-5 Simulation: GBLUP

Test set	Correlation between true and predicted phenotypes		Correlation between true and predicted breeding values		Heritability estimate[I] (standard error)
	$r_{y\hat{y}}$	***p*-value[II]**	$r_{g\hat{g}}$	***p*-value[II]**	
1	0.09	0.50	0.17	0.17	0.17 (0.14)
2	0.14	0.25	0.25	**0.04**	0.38 (0.18)
3	0.16	0.20	0.36	**<0.01**	0.25 (0.17)
4	0.02	0.86	-0.12	0.34	0.45 (0.17)
5	0.24	0.06	0.29	**0.02**	0.22 (0.17)
Overall	0.11	**0.04**	0.16	**<0.01**	n/a

[I]on the observed scale

[II]Significant results ($p<0.05$) shown in bold

8.4 Discussion

Results from the OCD dataset suggest that the amount of genetic variance underlying OCD susceptibility that is tagged by the markers in this sample is extremely small. The combination of the low estimated genetic variance and the low number of SNPs given non-zero substitution effects in the Bayesian model resulted in GEBV of the order 10^{-2} to 10^{-3} and estimated phenotypes that were dominated by the fixed effect (contemporary group). Hence, the improvement of fit to the model provided by the GEBV was minimal and shown by permutation not to be significant in any of the CV sets. The predictive ability of the model, as assessed by $r_{p\hat{p}}$, was also very low with no significant positive correlations observed in the test sets. The results of the genomic BLUP analysis of the OCD dataset were generally in line with results from the Bayesian method. The genetic component in the model was bounded which could be due to either a small genetic variance, or to data structure causing large errors, and subsequent GEBV were of the order 10^{-6} to 10^{-7}. Data was simulated in order to better understand the factors which may have led to this poor predictive ability.

The parameters of the simulated dataset were chosen to be representative of a complex disease scenario similar to OCD where heritability is low and there are a large number of QTL affecting disease status, relative to the genome structure. In order to make the best use of the available genotypes, disease prevalence was set to 50% rather than the predicted 25% prevalence in Thoroughbreds of OCD [76]. The random nature of the simulation resulted in a realised heritability (on the liability scale) of 0.21, just over the target heritability of 0.20. Of note, was the consistent over-estimation of heritability by the genomic BLUP procedure in all CV subsets. However, standard errors were large and three out of five estimates lay within a single standard error of the true observable heritability, 0.13. Further checking of the heritability was carried out by fitting raw simulated phenotypes for all 330 animals as a continuous trait in the animal model. Whilst the heritability was still overestimated, a log likelihood ratio test showed the estimate not to be significantly different from the simulated value of 0.21. The reason for the consistent over-

estimation of heritability is not clear, but may have to do with assumptions underlying the derivation of the genomic relationship matrix (see [59, 246, 247]).

In all cases, regardless of the method, the correlation between the true and estimated phenotypes ($r_{y\hat{y}}$) for the simulated data were positive. Predictive abilities ($r_{y\hat{y}}$) were higher than for the OCD dataset but were still relatively low and were not significant except in CV set five (using Bayesian analysis). As expected, GEBV accuracies, calculated as the correlation between true and predicted breeding values ($r_{g\hat{g}}$), were greater than predictive ability ($r_{y\hat{y}}$) and correlations were significant in three out of five test sets. When the five CV sets were combined, both predictive ability ($r_{y\hat{y}}$) and accuracy ($r_{g\hat{g}}$) were significantly different from zero. The reason for the lack of significance in some of the individual CV sets may be a lack of power, caused by having only 66 samples per test set. The accuracy based on the Bayesian analyses was significantly higher than that from the GBLUP analysis.

The variation in accuracy across the CV sets was relatively high in both the OCD and simulated data sets, but showed similar trends across the methods, indicating systematic variation in groups. This shows the importance of CV in the assessment of prediction performance. Here, horses (aged nine to 12 months) were sampled over two years but essentially represent a single generation, so I did not have to consider the presence of ancestors within the different sets. However, the chance sampling of half-sibs and other relatives within and between CV sets may have caused the relatively large variations in accuracy seen across them.

Given the small sample sizes and apparently small amount of information coming from the data, the proportion of SNPs with an effect in the model (π) was fixed to reduce the number of parameters that needed to be estimated. This method more closely resembles the original BayesB model of Meuwissen *et al.* (2001) [52]. Fixing π=0.10 in the analysis of the OCD data, resulted in the correlation between corrected true phenotype and GEBV ($r_{p\hat{p}}$) in the reference sets increasing considerably, but the corresponding statistic in the test sets showing no improvement.

The estimated genetic variance also increased. These findings suggest that the increased correlation observed in the reference sets was the result of over-fitting, that is, the result of marker effects explaining error, rather than any improvement in the estimation of the true effects. Similar results were seen in the simulated data where the value of π=0.0693 was chosen under the assumption that two adjacent SNPs per QTL would be fitted in the model. Clearly, this is a very simplistic view because the actual number of SNPs that it may be desirable to fit for every QTL so as to capture all its variance through LD, would be highly dependent on the degree of LD between each QTL and its surrounding SNPs.

Deterministic formulae which predict the accuracy of predicted genetic risk using a genome-wide approach allow the comparison of these results to expectation [61]. Two versions of the formula derived by Daetwyler *et al.* (2008) [61] are relevant here. The first is for a dichotomous disease phenotype assuming a population sample. This assumes that the proportion of samples with disease is the same as the prevalence of the disease in the population and is relevant to the simulation scenario. The formula takes the form $r^2_{g\hat{g}} = \frac{\lambda h^2_o}{\lambda h^2_o + 1}$, where, $r^2_{g\hat{g}}$ is the squared correlation between the true and the estimated additive genetic value, and h^2_o is the heritability of the disease on the observable (0,1) scale. $\lambda = {n_P}/{M_e}$, where n_P is the number of samples with phenotypes in the reference set and M_e is the number of independent segments in the genome and can be calculated across *i* chromosomes as $\sum_i \frac{2N_e L_i}{\log(4N_e L_i)}$, where L_i is chromosome length in Morgans and N_e is the effective population size [155, 248]. The second version of the formula incorporates an adjustment for case control study design such as that used for the OCD data and takes the form $r^2_{g\hat{g}} = \frac{\lambda h^2_o}{\lambda h^2_o + c}$ and $c = q(1-q)\left(1 - h^2_l \bar{i}(\bar{i} - x)\right) w^{-1}(1-w)^{-1}$, where, q is the disease prevalence, w is the proportion of cases in the sample, h^2_l is the heritability on the liability scale, $\bar{i} = wi_q - (1-w)i_{(1-q)}$, and x and i are as defined in

Falconer & MacKay (1996) [124], assuming *q* here is equivalent to *p*. Under both versions of the equation, it is assumed that all of the genetic variance is captured by the markers.

Using these formulae, together with predictions of N_e for the US Thoroughbred population from Chapter Three, further conclusions may be drawn from the results. Assuming an N_e of 100 (and a corresponding M_e of 939), predicted accuracies for the OCD dataset given heritabilities on the liability scale of 0.10 and 0.30, would be $r_{g\hat{g}} = 0.14$ and $r_{g\hat{g}} = 0.24$, respectively; clearly, this level of accuracy was not achieved. In contrast, the predicted accuracy for the simulated scenario of $r_{g\hat{g}} = 0.19$ was realised when a Bayesian approach was used. In the simulation, QTL SNP genotypes were removed prior to the genomic analysis to more closely resemble the likely situation in the OCD analysis. The fact that the expected accuracy was still achieved suggests much of the genetic variance was captured by remaining markers through linkage, making inadequate genome coverage an unlikely explanation for the poor results of the OCD analysis. Similarly, whilst there was some evidence in the simulation of poor power due to small sample size, with a significant correlation ($r_{g\hat{g}}$) only being observed in three out of five test sets, by combining test sets the power was sufficient to achieve strong significance ($p<0.01$) and so the small sample size may not fully account for the poor predictive ability in the OCD dataset. Whilst the use of a proxy measure ($r_{p\hat{p}}$) for $r_{g\hat{g}}$ in the OCD model would have lowered the expected correlation, the comparison of $r_{g\hat{g}}$ and $r_{y\hat{y}}$ in the simulated data showed only a small difference (0.03). Therefore, it seems likely that specific properties of the OCD sample affected the accuracy with which GEBV could be predicted and some such properties are explored below.

In order to negate any possible farm effect on the data, an approximately equal number of cases and controls were sampled from each farm (see Chapter 2). However, it is also likely that half sibs were clustered by farm as individual farms will tend to use the same stallion on several of its mares. The result of this may have been the incidental matching of cases and controls for pedigree, adding structure to

the data that was subsequently unaccounted for. The power of the analysis will also have been reduced by the fitting of the fixed effect, contemporary group. Whilst this was necessary due to the uneven distribution of cases and controls amongst classes, the true power of the resulting analysis might be expected to more closely resemble that of a study with 144 samples, that is, twice the sum of the minimum out of the number of cases or controls for each class, rather than the actual sample size of 330. Furthermore, although horses were rigorously screened for OCD in this study, misclassification may also have contributed to a reduction in overall power. The combination of these factors may go some way to explaining the difference in success between the OCD and the simulated trait analyses.

Despite the disappointing results from the OCD dataset, the simulation results were promising. By imposing QTL on existing genotypes, many real features of the genome are retained, along with their impact on the efficacy of the method. A criticism of some early simulation studies might be that the low number of QTL simulated led to unrealistically high estimated accuracies of prediction that could not be achieved for highly complex traits. Therefore, in this study, a relatively large number of QTL, relative to the genome structure, were simulated. To see some predictive ability in the simulated data with such small sample sizes indicates that, whilst the OCD study may have been under-powered, the 50k SNP chip is sufficient to make such genome-wide methods efficacious in the horse. With an increase in sample size and therefore a corresponding increase in accuracy, such an approach has the potential to be used to inform breeding decisions and this potential has been recognised by several authors to date. Haberland *et al.* (2012) [249] showed by simulation that, given a young horse without phenotypic data (or progeny), the additional information from GEBVs with accuracies of 0.3, could increase the correlation between true and predicted breeding values from 0.27, achieved using sire and dam data only, to 0.39. These results were based on a trait of low heritability (h^2=0.15) intended to represent OC, but results were also shown for a more highly heritable trait. Furthermore, Sitzenstock *et al.* (2010) [250] concluded from their simulation work, that the use of GEBV in horse breeding programmes

could increase profitability above that achieved with traditional selection strategies across a range of GEBV accuracies from 0.3 to 0.8.

Whilst only Thoroughbred samples were used in this study predictions may need to be made across a wide range of breeds and crossbreds. Work in other livestock species and using simulation has shown that across breed predictions can be much lower than within breed predictions [134, 251, 252]. Whilst this is also expected to be the case in horses, preliminary haplotype analyses have shown a large degree of haplotype sharing across breeds [23] which will aid across breed evaluations. Furthermore, many modern sport horse breeds comprise a large proportion of Thoroughbred genes, for example, the Hanoverian has a predicted 35% Thoroughbred genes [225] and this will aid transferability across breeds. Further studies are warranted to evaluate the true potential of GS in horses but are currently not possible due to an insufficient number of well-phenotyped samples. One possible approach to address this problem is low-density genotyping and this is explored in the next chapter.

Chapter 9: The utility of low-density genotyping in the Thoroughbred

9.1 Introduction

The introduction of high-throughput single-nucleotide polymorphism (SNP) chips which permit the analysis of large numbers of SNPs in parallel, has enabled large-scale studies of human and livestock populations to be carried out. The increased marker density has led to genome-wide association studies (GWAS) such as that carried out in Chapter 7, replacing linkage analyses as the dominant method in the hunt for genetic variants, due to their superior potential for informing us about complex as well as monogenic conditions. In livestock, genomic selection (GS) has become a realistic alternative, or at least an adjunct, to the pedigree based selection methods of the last 40 years, helping to overcome the uncertainty due to the Mendelian sampling term [253]. A common feature of GWAS and other population scale analyses, is the requirement for large sample sizes which are needed to ensure sufficient power to detect what are hypothesised to be relatively small quantitative trait loci (QTL) effects. As well as needing a substantial number of samples for the initial analysis, a second independent sample is required for the validation of any QTL identified. Furthermore, any underlying data structure such as that caused by differing ancestries, e.g. different breeds, and the presence of environmental factors, has the potential to reduce power for a given sample size. Whilst GS does not aim to identify QTL as such, the method is still reliant on large sample sizes in order to minimise the error in the estimation of the SNP effects which are then used to predict genomic estimated breeding values (GEBV) in the selection candidates.

In the equine setting, the accumulation of large numbers of samples represents a significant challenge. Since the introduction of the first equine SNP chip by Illumina in 2007, several GWAS of monogenic traits have been successful in identifying QTL and in several cases, causal mutations [66, 196, 254]. However, success in complex traits has been less convincing; some studies have reported QTL, but many of these QTL have been defined under ad hoc significance thresholds, as authors attempt to balance the risk of Type I and Type II errors. The apparently low signal to noise ratio in such studies is an indication of low power caused, in part, by small sample

sizes. Moreover, insufficient validation studies have been done to confirm whether or not these initial findings are true associations or false positives. The discontinuation of the 50K SNP chip after only a few years production is evidence in itself that the numbers of horses being genotyped has been below that which was expected. The reason for the typically small sample sizes is multifactorial but can be partly attributable to the cost of genotyping. Whilst the cost of purchasing and running SNP chip analyses has undoubtedly fallen during the last few years, it is relative cost which is important. In the equine industry, within the UK sport horse sector in particular, the potential to make significant returns from breeding superior animals is limited. Therefore, the relative cost of genotyping remains high and the potential to genotype large numbers of samples is correspondingly low. The integration of approaches such as GS into the equine industry will require genotyping to become more cost effective.

One opportunity to lower genotyping costs comes in the form of low-density genotyping. If a reference panel of individuals genotyped at high density is available, study samples (or selection candidates) may be genotyped for a subset of these loci on a low-density panel (LDP), and imputation carried out to fill in the 'missing' SNP genotypes. Provided the reference panel and study sample originate from a genetically similar population, population genetic models can be used to extrapolate allelic correlations measured in the former to predict unobserved genotypes in the latter [255]. The dependence of imputation accuracy on SNP density means that but there will always be a trade-off to be made between the cost of genotyping and the accuracy of imputation. Other factors that affect the accuracy of imputation include levels of linkage disequilibrium (LD) in the population, the degree of similarity between the reference panel and the study sample and to some extent, the size of the reference population. Efforts to develop improved imputation algorithms have resulted in a wide range of programs, most of which have evolved from programs written to infer haplotype phase from large-scale diplotype data sets, being available. Commonly used programs include fastPHASE [256], MACH [257], IMPUTE [258] and BEAGLE [259] and their relative efficacies have been explored under various scenarios [259-263]. Whilst some of these imputation methodologies

may use linkage analysis to exploit known relationships between individuals, in many cases, knowledge of relationships is not required.

Due to the reliance of the method on LD between LDP SNPs and remaining loci, SNP selection for LDPs also plays a role in determining the accuracy of imputation. As such, significant effort has been devoted to optimising LDP SNP selection and several algorithms have been developed along this vein. Many programs utilise LD between pairs or groups of markers to select LDP SNPs in a so-called block-free approach, e.g. Tagger [264], LDSelect [265]. Another common approach is to use haplotype information in a block-based approach, e.g. Hapblock [266], whilst other more novel algorithms have also been developed such as the neighbourhood graph approach of Halldórsson *et al.* (2004) [267] or the multiple linear regression approach of He & Zelikovsky (2006) [268]. In situations where LDP SNPs are selected to predict haplotypes, they are commonly referred to as 'tag SNPs'. A useful review of many of the currently available software programs can be found in Halldórsson *et al.* (2004) [269].

In this study, I use genotypes from the Illumina Equine SNP50 BeadChip to investigate the accuracy of imputation which can be achieved in Thoroughbred horses, without pedigree information, and using a typical imputation program (BEAGLE). Three methods of LDP SNP selection were tested across six LDP sizes in order to evaluate the impact of various SNP selection criteria including SNP informativeness and LD. The effect of geographical substructure on the accuracy of imputation was also investigated.

9.2 Materials and methods

9.2.1 Genotypes

The data for this study comprised all 1,201 available samples (n=853 Thoroughbreds from the UK and n=348 Thoroughbreds from the US), genotyped at the 50,707 SNPs that passed preliminary quality control (QC). The UK dataset was used to evaluate three methods of LDP SNP selection across six LDP densities in a within-population analysis. The US dataset was subsequently used to evaluate the impact on

imputation accuracy of geographical substructure between reference panels and study samples in a between-population analysis (at a single LDP density).

Samples in the UK dataset were randomly assigned to one of the following three subsets: Set A containing 200 samples that were used to select LDP SNPs; Set B containing 490 samples that were used as the reference panel; Set C containing the remaining 163 samples that were used as the test panel (representing the study sample). Two analyses were performed using the US dataset. In the first analysis, Set B was used as the reference panel for imputation and the entire US dataset used as the test panel. In the second analysis, samples in the US dataset were randomly assigned to one of two subsets: Set D containing 261 samples that were used as the reference panel; Set E containing 87 samples that were used as the test panel. A graphical representation of this data flow can be seen in Figure 9-1.

Quality control was applied in Set A in order to generate a list of SNPs to be used in the subsequent stages. SNPs that were genotyped in less than 95% of the set and those with a minor allele frequency (MAF) of less than 0.01 were excluded. In the within-population analysis, four *Equus caballus* (ECA) chromosomes were analysed: ECA1, ECA10, ECA20 and ECA26; these chromosomes were chosen to represent the shortest (ECA26), longest (ECA1), and median length (ECA20) chromosomes (as measured in cM according to Swinburne *et al.* (2006) [22]) and to include two chromosomes with centromeres (ECA1 and ECA10) and two acrocentric chromosomes (ECA20 and ECA26). In the between-population analysis ECA1 and ECA26 were analysed.

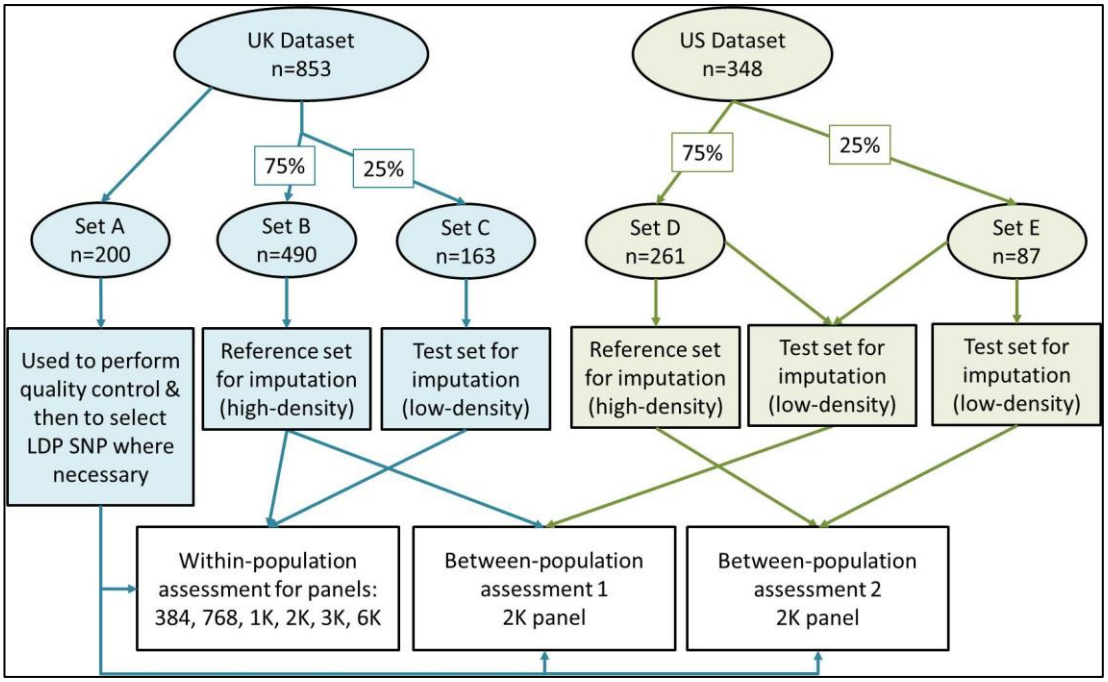

Figure 9-1 Data flow for analysis

9.2.2 Low-density panel SNP selection

The three methods used to select LDP SNPs are detailed below. The methods were tested at six different densities representing genome-wide panels with 384, 768, 1K, 2K, 3K and 6K SNPs. The equivalent densities, expressed in terms of N_e SNP per Morgan [270], were 0.09, 0.18, 0.24, 0.48, 0.72 and 1.44N_e/Morgan (assuming N_e=150 and total genome length is 27.72 Morgans). The number of LDP SNPs selected per chromosome was proportional to the length, in base pairs, of the chromosome for all methods.

Method 1: Equidistant (Mb) selection (bpEQ)

Low-density panel SNPs were selected so that they approached equidistant spacing, by base pairs, along the chromosome. This was achieved by dividing the total length (Mb) of the chromosome into equally sized segments, the number of segments being equal to one less than the number of desired LDP SNPs for the given density. The closest SNP to each segment boundary was then chosen to be a SNP in the LDP, irrespective of MAF.

Method 2: Equidistant (Mb) optimised for MAF (bpMAF)

The program described in this section was written and run by Andreas Kranis (of Aviagen Ltd.). Low-density panel SNPs were selected so that they approached equidistant spacing, by base pairs, along the chromosome and had a high MAF. The following paragraph was written in part by Andreas Kranis to describe his program.

In order to meet both objectives, SNP selection was performed using a custom python program employing a genetic algorithm. The cost function to be minimized included two components. The first one aimed to drive the MAF of selected SNPs towards 0.5 by employing a penalty equal to $\left(0.5 - MAF_{SNP_i}\right)^2$ (1). The second component ensured the equal spacing. An ideal distance was calculated as: $d = \frac{chr_{length}}{n_{chr}-1}$ and then, the spacing between consecutive SNPs *i* and *i*+1 in the LDP was forced to approach *d* using the function: $\left((S_i - S_{i+1}) - d\right)^2$ (2), where *S* is the base pair position of the SNP. By combining (1) and (2), the selected set of n_{chr} SNPs was derived by iteratively minimizing this function over all SNPs:

$$\sum_n \left[\left(0.5 - MAF_{SNP_i}\right)^2 + \left((S_i - S_{i+1}) - d\right)^2\right]$$

In order to ensure good coverage on the telomeres, where recombination events are more frequent and hence accuracy of imputation is expected to be lower, the first and last SNPs in the chromosome qualified in the LDP.

Method 3: Equidistant (LDU) optimised for MAF (lduMAF)

Low-density panel SNPs were selected according to the same algorithm as used in Method 2, but with SNP location given, not as a base pair position, but in linkage disequilibrium units (LDU). The LDMAP program (http://cedar.genetics.soton.ac.uk/pub/PROGRAMS/LDMAP) described and developed by Maniatis *et al.* (2002) [271] was used to construct an LD map for each chromosome in turn using the diplotypes of samples in Set A. First, LDMAP computes the absolute *D′*-statistic [272] between all pairs of markers. For map construction, each of the chromosomes was divided into several overlapping segments with no more than 250 SNPs per segment and an overlap of 50 SNPs. The

number of segments per chromosome ranged from four (ECA26) to 15 (ECA1). For each segment, LDMAP fits the Malecot model to *D′ vs.* inter-marker distance (*d*) data in kilobases. This quantifies the average rate of decline of LD for the segment which is a useful starting value for computing the interval-specific estimates used in LD map construction.

The Malecot model predicts the decline of association with distance as $\rho(d) = (1-L)Me^{-\varepsilon d} + L$, with $\rho(d) = E(D')$, and where *L* is the residual association at large distances, *M* is the proportion of the youngest haplotype (assumed to be the rarest haplotype) that is monophyletic (descended from one source), and ε is the rate of exponential decline of $\rho(d)$ with distance *d*. For LD map construction, the LDMAP program estimates the *ε* and *M* parameters in the Malecot model for each interval (between adjacent SNPs) in the map, using data from marker pairs that include that interval in sliding windows. Here, three rounds of iteration were performed. In the first round, the *ε* parameter was estimated with *M* set equal to 0.5. In the second round, both *ε* and *M* were estimated, with the starting values taken from the first round. The third round ensured convergence. The third Malecot model parameter *L* was kept constant at a value predicted according to Morton *et al.* (2001) [273] and generated by the program internally. Any larger interval that includes the one being estimated contains some information about *ε* unless the markers in the pair are at such a large distance that they contain no more useful information about LD. Here, the maximum distance between any pair was set to 10Mb. The length of the *i*th interval is $\varepsilon_i d_i$ LDUs, where ε_i is the Malecot decay rate parameter specific to that interval and d_i is the length of the interval on the physical map in kilobases. A chromosome has a total of $\sum_i \varepsilon_i d_i$ LDUs [271]. Once all SNPs in all segments have been mapped, the fit of the completed LD map is checked in a final round of iterations and parameter estimates. In cases where SNPs were allocated the same position in the LD map, a small addition was made to subsequent locus positions (10^{-6}) before entry of the SNP locations to the LDP SNP selection algorithm in order that SNP order (according to the physical map) was maintained. Linkage disequilibrium maps are shown in Figure 9-2 and Figure 9-3.

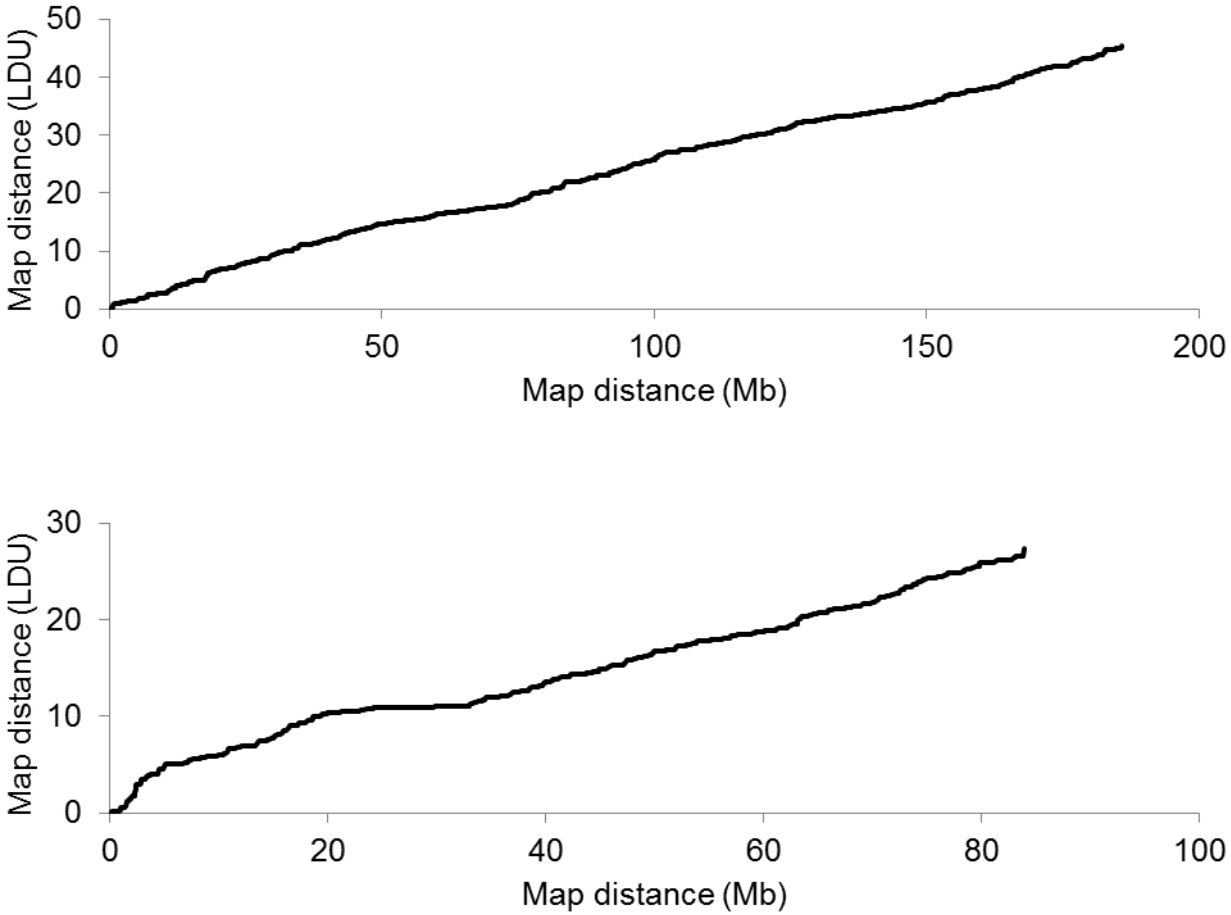

Figure 9-2 LD maps

a) ECA1; b) ECA10

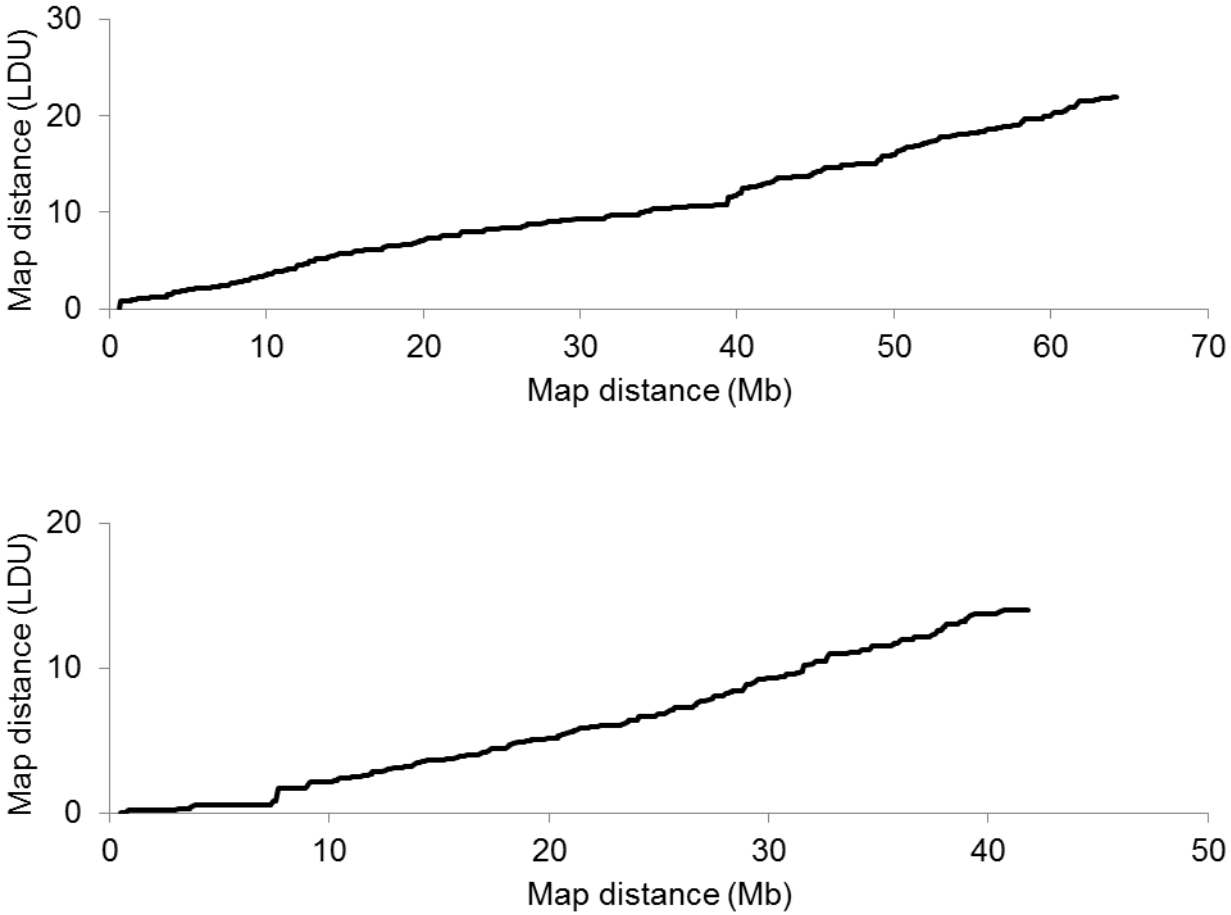

Figure 9-3 LD maps

a) ECA20; b) ECA26

9.2.3 Imputation

The program BEAGLE [274] was used to impute from low to high density markers. This program is based on a Hidden Markov model (HMM) in which the observed data are the observed unphased genotypes (which may include errors and missing data) and the hidden state represents the haplotype phase and the true genotypes [275]. In the model, the observed haplotypes are clustered at each marker position, based on the similarity of the haplotypes in the local vicinity [275]. The default parameters of the program were used throughout and the most likely genotype taken to be the imputed genotype. Input files consisted of the reference panel genotype data (Set B/Set D) with a corresponding list of all SNPs contained in the high density panel after QC and the test panel genotype data in which all genotypes except those selected to be LDP SNPs were set to missing (Set C/Set E/Set D + E). For reference purposes, missing genotypes were also imputed by random assignment of alleles to missing loci, assuming that the probability of a particular allele was equal to its observed frequency in the reference panel (Set B/Set D).

The accuracy of imputation in the test panel was assessed in several ways. Measures of accuracy were made both at the sample level and at the marker level. Firstly, the proportion of correctly imputed genotypes was calculated as the number of correctly imputed genotypes divided by the total number of imputed genotypes. Secondly, the correlation between the imputed and the true genotypes (coded as 0, 1, 2) was calculated. In the case of monomorphic SNPs, loci were excluded from the correlation calculation. These measures were made for the three different LDP SNP selection methods, six different LDP densities and four chromosomes.

9.3 Results

9.3.1 Within population assessment

The accuracy of imputation from low to high density panels increased as the number of SNPs in the low density panel increased, as shown in Table 9-1 for ECA1 and in Figure 9-4 and Figure 9-5 for all chromosomes. The increase was greatest at the lower densities, with the increase in accuracy from 1K to 6K being relatively less. A large range in the proportion of genotypes correctly imputed was observed between

animals, particularly when the number of LDP SNPs was at its lowest when the range was typically around 0.4. Whilst the difference in accuracy across the three LDP SNP selection methods was small or indistinguishable, selection methods 2 and 3 did appear to reduce the variation in imputation accuracy across SNPs, particularly at lower SNP densities.

The random imputation of genotypes based on allele frequencies in the reference panel demonstrates the minimum imputation accuracy that can be expected. Figure 9-6 shows the strong dependency of the accuracy of imputation on MAF when random imputation is used; the results follow closely the expectation, calculated as $p^4 + 4p^2q^2 + q^4$ [92]. The relationship between MAF and imputation accuracy is less strong when BEAGLE is used to impute but becomes increasingly strong as the SNP density of the LDP decreases.

In order to explore possible causes of heterogeneity in SNP imputation accuracy, imputation accuracy was plotted on SNP position (bp) (Figure 9-7). In addition, average pairwise LD (calculated as in Chapter 6) was calculated in 1Mb sliding windows (0.5Mb overlap) and plotted alongside imputation accuracy. Hypothesised centromere positions are also marked on the plots; these are regions that share some similarity with centromeric satellite sequences. ECA1 centromere position is hypothesised to be at 66Mb or 89Mb; ECA10 centromere position is hypothesised to be at 28.2Mb although there is a second region, ranging from 81Mb to 83Mb, that also contains some centromeric satellite-like sequences; ECA 20 and ECA26 are not centromeric but regions identified here may, if the similarity with centromeric satellite sequences is real, represent locations harbouring centromeres in the past (C. M. Wade 2012, Pers. Comm.). Plots show considerable variation in imputation accuracy across the chromosome which is often correlated with levels of LD. The variation is particularly marked in the case of ECA10, where the high LD around the proposed centromere position led to a peak in imputation accuracy. Using Method 3 (lduMAF) to select LDP SNPs reduces the variation in accuracy across the chromosome, leading to a more consistent level of accuracy and a reduction in its correlation with LD levels. This corresponds with the decreased variance of Method

3 shown in Figure 9-4 and Figure 9-5. In general, the decrease in accuracy seen with Method 3 in regions of high LD was greater than the corresponding increase in low LD areas. This explains the inability of this method improve mean accuracies above those achieved using Method 2 (bpMAF).

Table 9-1 Mean (min., max.) proportion of correctly imputed genotypes by sample (ECA1)

No. of SNPs[I]	bpEQ	bpMAF	lduMAF
384	0.66 (0.52,0.93)	0.67 (0.55,0.94)	0.69 (0.55,0.92)
768	0.76 (0.59,0.94)	0.77 (0.62,0.95)	0.78 (0.59,0.96)
1K	0.79 (0.61,0.94)	0.84 (0.66,0.97)	0.83 (0.64,0.98)
2K	0.90 (0.70,0.99)	0.91 (0.71,0.99)	0.89 (0.68,0.99)
3K	0.94 (0.70,0.99)	0.95 (0.73,1.00)	0.92 (0.67,0.99)
6K	0.97 (0.79,1.00)	0.98 (0.78,1.00)	0.95 (0.75,1.00)

[I]**Total number of SNPs that would be on a genome-wide LDP of equivalent density**

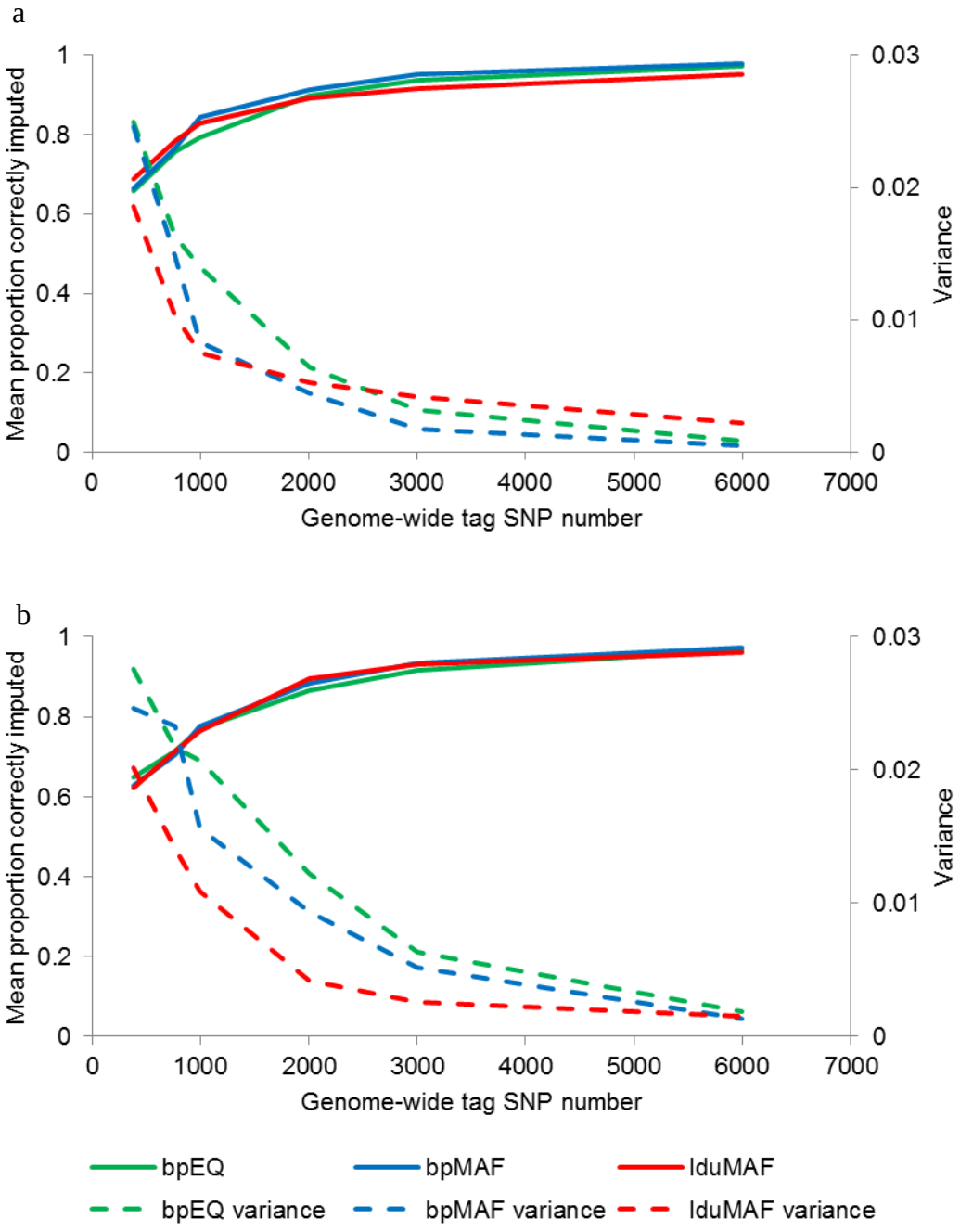

Figure 9-4 Mean proportion of correctly imputed genotypes by marker and its variance

a) ECA1

b) ECA10

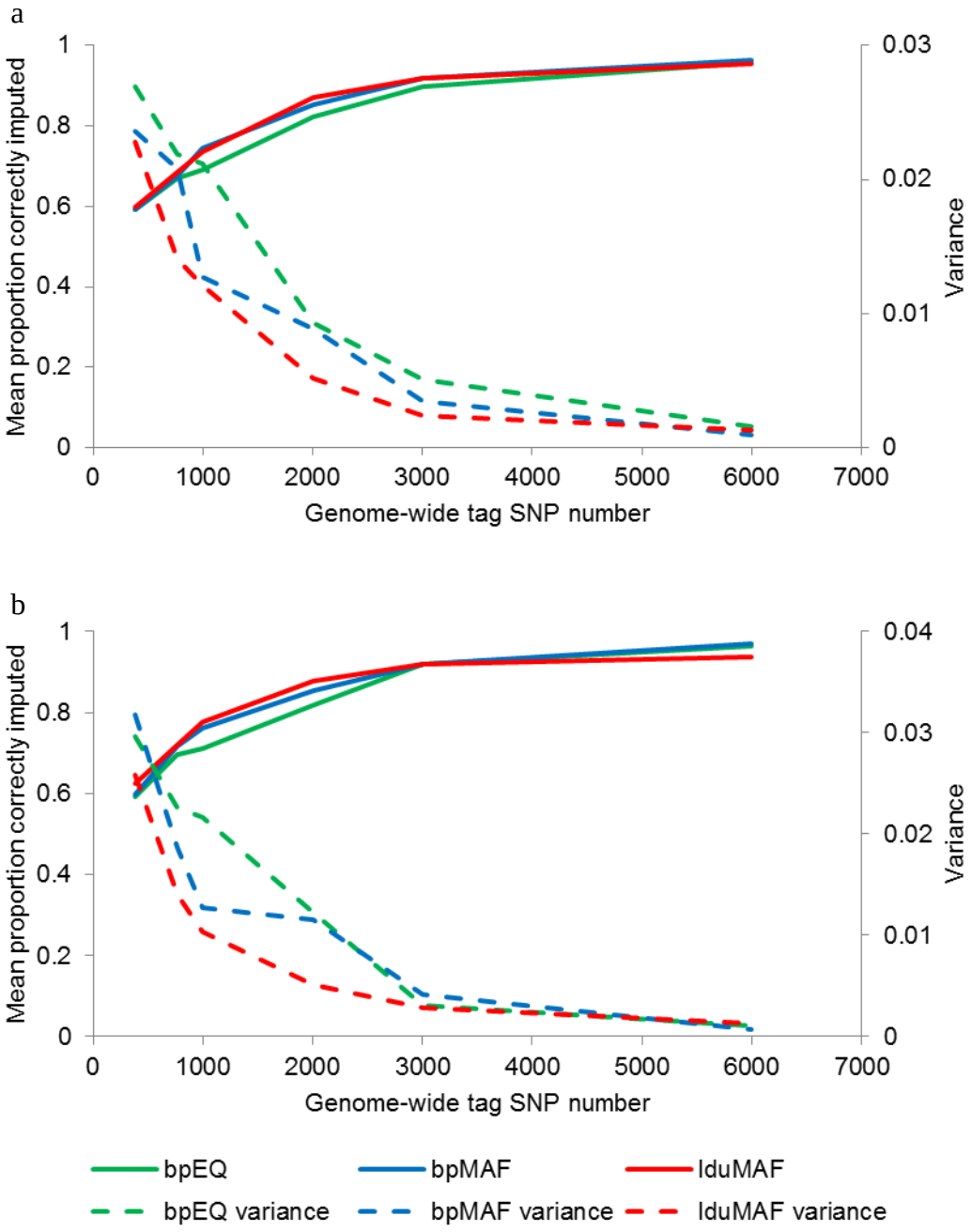

Figure 9-5 Mean proportion of correctly imputed genotypes by marker and its variance

a) ECA20

b) ECA26

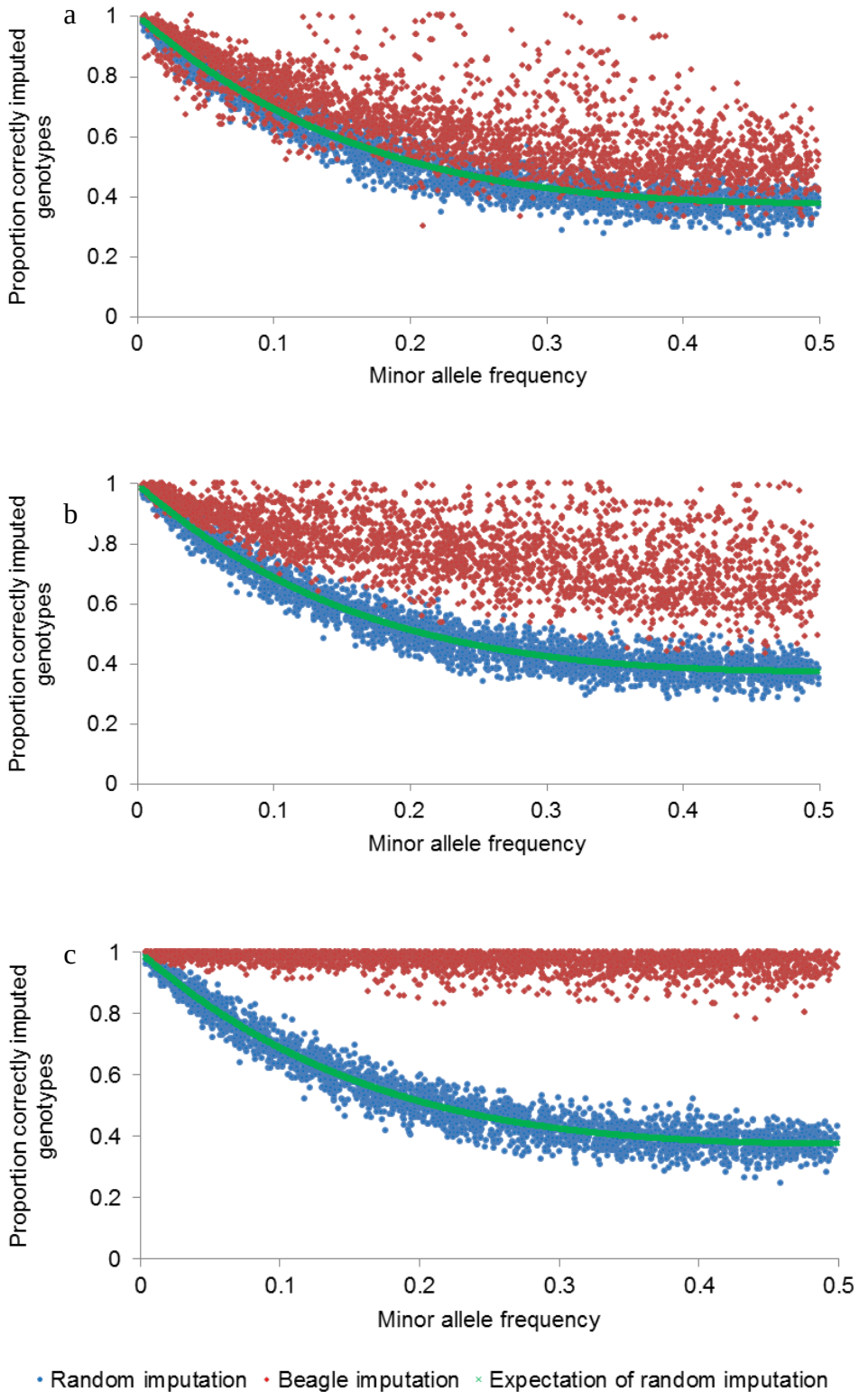

Figure 9-6 Proportion of correctly imputed genotypes plotted on the MAF of the SNPs (calculated in the reference panel) for ECA1 (bpEQ)

a) 384 panel; b) 1K panel; c) 6K panel

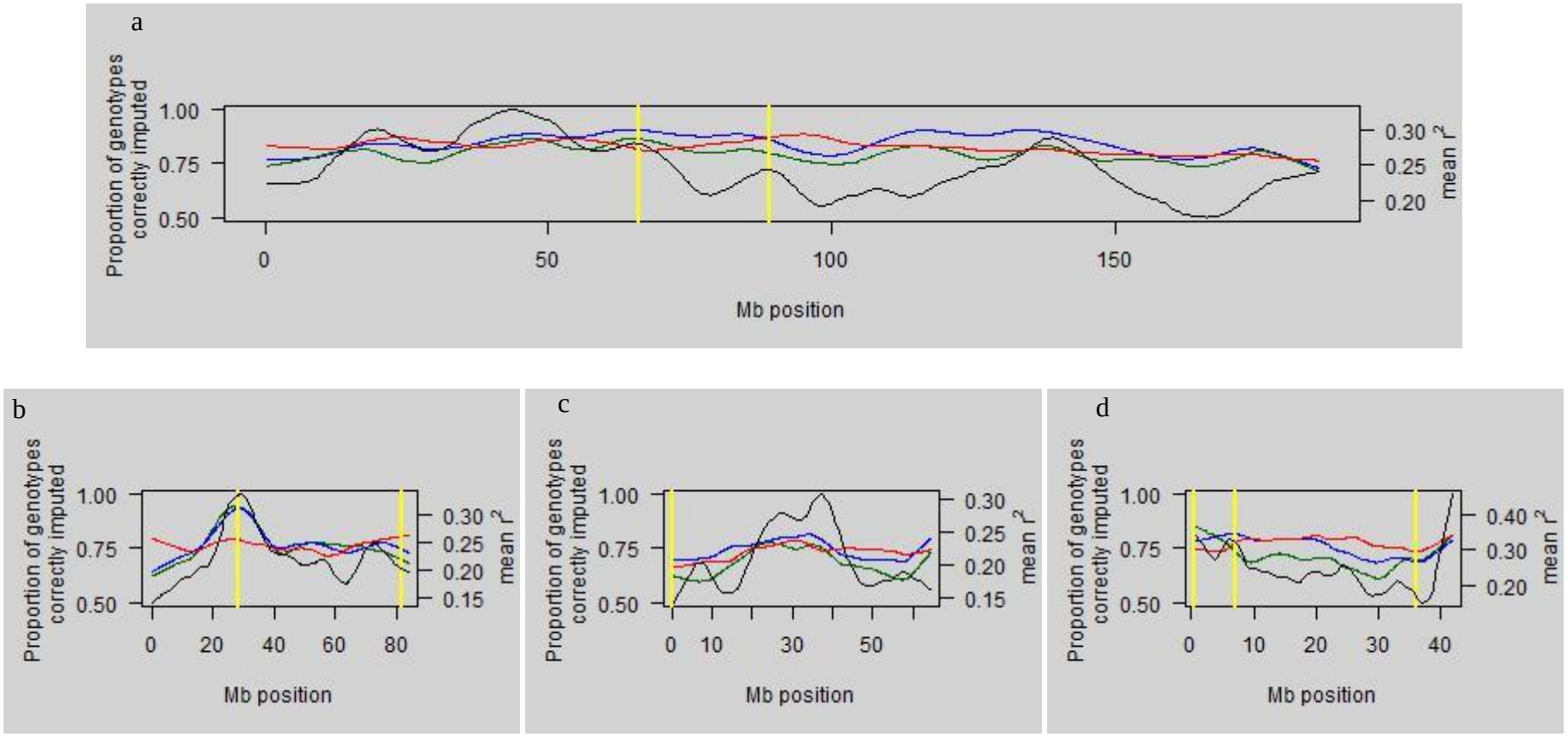

Figure 9-7 Proportion of correctly imputed genotypes and mean LD, presented as Lowess curves as calculated in R [140] [276-278], and plotted on SNP position (1K panel). Key: Green – bpEQ; blue – bpMAF; red – lduMAF; black – mean r^2 calculated in sliding windows of 1Mb (0.5Mb overlap); yellow – hypothesised centromere position. a) ECA1; b) ECA10; c) ECA20; d) ECA26

An alternative measure of imputation accuracy, the correlation between true and predicted genotypes, was also calculated. Accuracy rates were generally lower when expressed as correlations with considerable variation being observed across animals and SNPs (Table 9-2). A comparison of the two accuracy measures (proportion correctly imputed and correlation between true and predicted genotypes) showed some correspondence but also MAF dependent variation in the relationship (Figure 9-8a). The adjustment of the proportion of correctly imputed genotypes by the expected accuracy that would be achieved using random imputation, was calculated as: $\frac{\text{accuracy} - \text{random accuracy}}{1 - \text{random accuracy}}$ [263], where accuracy is the proportion of correctly imputed genotypes achieved and random accuracy is the expected accuracy that would be achieved using random imputation according to the aforementioned equation, $p^4 + 4p^2q^2 + q^4$. This statistic adjusts for the fact that SNPs with low MAF are likely to be imputed to high accuracy by chance alone. Plotting the adjusted statistic against the correlation between true and imputed genotypes shows a much stronger relationship between the two measures than when the raw accuracy was used. Furthermore, the relationship is almost independent of MAF, although SNPs with lower MAF tend to illustrate more variation in both imputation accuracy and correlation (Figure 9-8b).

Table 9-2 Mean correlation (min., max.) between true and predicted genotypes by sample (ECA1)

No. of SNPs[I]	bpEQ	bpMAF	lduMAF
384	0.46 (0.14,0.89)	0.49 (0.20,0.91)	0.53 (0.22,0.89)
768	0.64 (0.36,0.93)	0.66 (0.38,0.93)	0.69 (0.37,0.94)
1K	0.70 (0.41,0.93)	0.78 (0.51,0.96)	0.75 (0.47,0.98)
2K	0.86 (0.53,0.99)	0.88 (0.62,0.98)	0.85 (0.52,0.99)
3K	0.92 (0.59,0.99)	0.94 (0.60,1.00)	0.88 (0.48,0.99)
6K	0.97 (0.73,1.00)	0.97 (0.71,1.00)	0.93 (0.61,1.00)

[I]**Total number of SNPs that would be on a genome-wide LDP of equivalent density**

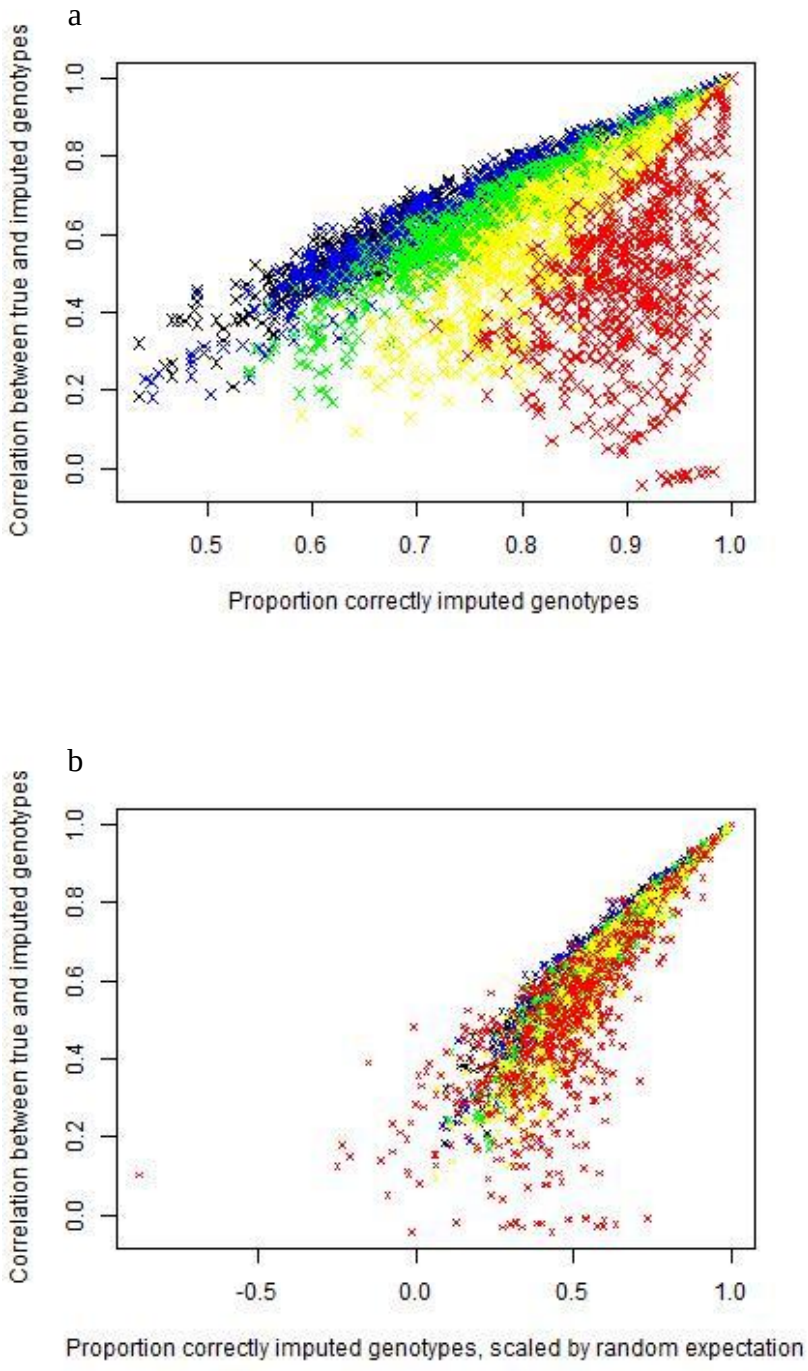

Figure 9-8 Correlation between true and imputed genotypes plotted on: a) proportion of correctly imputed genotypes; b) proportion of correctly imputed genotypes, scaled by random expectation. Black – SNPs with MAF ≥ 0.40; blue – SNPs with 0.30 ≤ MAF < 0.40; green – SNPs with 0.20 ≤ MAF < 0.30; yellow – SNPs with 0.10 ≤ MAF < 0.20; red – SNPs with MAF < 0.10. Data for ECA1 and 1K panel.

9.3.2 Between-population assessment

When Set B was used as the reference panel for imputation in the US dataset there was very little change in the mean proportion of correctly imputed genotypes relative to the within-population results (Table 9-3). Using random imputation, no change in accuracy was observed for ECA1, whilst a small but consistent decrease in the mean of 0.01 was seen for ECA26. This small difference is presumably due to the high correlation in MAF between the two populations which was calculated to be 0.91 for ECA1 and 0.90 for ECA26. Using BEAGLE to impute gave a similar pattern of results with no change for ECA1 and a slight decrease in accuracy for ECA26.

When Set D and E were used as reference and test sets, a slight increase in imputation accuracy was seen compared with the UK within-population results across both imputation methods and for all three LDP SNP sets (Table 9-3). This increase is likely due to higher average genomic relationship of horses in the US dataset compared to the UK dataset; when average genomic relationships were calculated for all samples using SNPs on ECA1 (as in Chapter 8), the mean relationship between horses in the US dataset was 0.022, compare to 0.003 in the UK dataset. There appeared to be no relative improvement of the bpEQ LDP SNP set over the more population specific LDP SNP sets (bpMAF and lduMAF), with the relative ranking of LDP SNP sets remaining unchanged when compared to the within-population assessment of the UK dataset.

9.3.3 Relative Cost

Of interest is the cost of genotyping horses for the LDPs relative to both the cost of genotyping at high-density and the loss of accuracy incurred. In the following, it is assumed that a reference population of horses, of comparable genetic background, have already been genotyped at high density (either 50K or 70K). Estimated genotyping costs were provided by Illumina for 3,072 to 6K SNPs based on their iSelect Infinium Assay and by Neogen for 384 to 2K SNPs, with 1K and 2K panel prices based on a custom chip construction. A minimum order size of 1,152 samples applies to all custom panels from both companies. Neogen are also the current providers of the 74K SNP chip which costs £115 per sample. The 50K chip is no longer available so the 74K chip is used to represent the high-density genotyping

option although it is possible that imputation accuracies up to 74K would be slightly different to those observed here. Figure 9-9 shows the cost of the different sized chips plotted against the average imputation accuracy (expressed as the correlation between the imputed and the true genotypes). In this plot, the most advantageous place to be would be the bottom right-hand corner, as here you benefit from both low cost and high accuracy. It is assumed that 100% accuracy is achieved by genotyping horses for all SNPs on the high-density panel, although the realised accuracy would be slightly less than this due to genotyping errors.

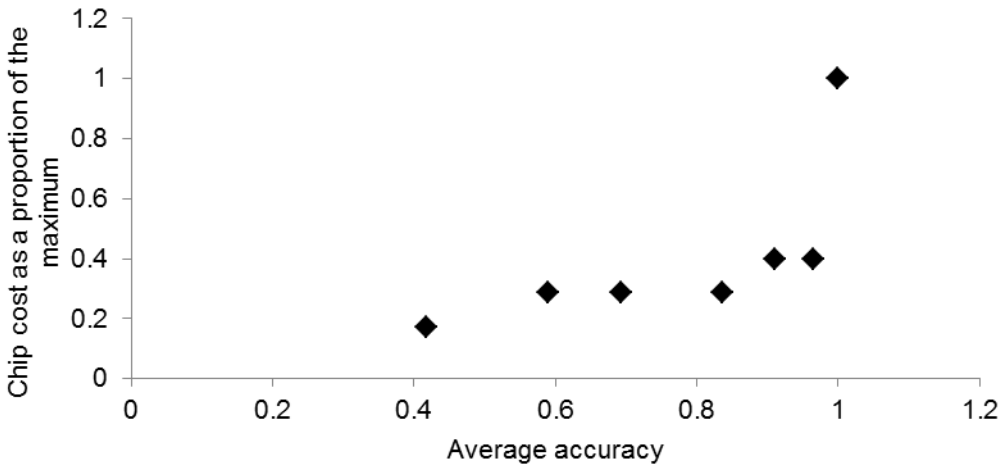

Figure 9-9 The cost of genotyping relative to the maximum (high-density) and plotted against average accuracy, expressed as the average correlation between imputed and true genotypes across the four chromosomes (bpMAF)

Table 9-3 Mean (standard error across SNPs/samples) proportion of correctly imputed genotypes on ECA1/ECA26 (2K panel)

Chr	Imputation method	Selection method	Within-population assessment	Between-population assessment 1	Between-population assessment 2
ECA1	Random	bpEQ	0.55 (0.003/0.001)	0.55 (0.003/0.0009)	0.56 (0.003/0.002)
		bpMAF	0.55 (0.003/0.001)	0.55 (0.003/0.0009)	0.56 (0.003/0.002)
		lduMAF	0.56 (0.003/0.001)	0.56 (0.003/0.0009)	0.57 (0.003/0.002)
	Beagle	bpEQ	0.90 (0.001/0.003)	0.90 (0.001/0.002)	0.92 (0.001/0.004)
		bpMAF	0.91 (0.001/0.003)	0.91 (0.001/0.002)	0.93 (0.001/0.004)
		lduMAF	0.89 (0.001/0.004)	0.89 (0.001/0.002)	0.92 (0.001/0.005)
ECA26	Random	bpEQ	0.51 (0.006/0.002)	0.50 (0.006/0.001)	0.51 (0.006/0.002)
		bpMAF	0.52 (0.006/0.002)	0.51 (0.006/0.001)	0.51 (0.006/0.002)
		lduMAF	0.52 (0.006/0.002)	0.51 (0.006/0.001)	0.51 (0.006/0.003)
	Beagle	bpEQ	0.82 (0.004/0.008)	0.81 (0.004/0.004)	0.86 (0.004/0.01)
		bpMAF	0.85 (0.004/0.006)	0.84 (0.004/0.004)	0.89 (0.003/0.01)
		lduMAF	0.88 (0.003/0.007)	0.85 (0.003/0.004)	0.90 (0.002/0.008)

9.4 Discussion

In this study, the efficacy of imputation from low to high density in a sample of Thoroughbred horses was investigated. Using equidistant LDP SNPs (bpEQ), the mean proportion of correctly imputed genotypes ranged from 0.59 at the minimum LDP SNP density of 0.09N_e/Morgan, to 0.97 at a density of 1.44N_e/Morgan. The relative increase in accuracy was greater between 0.09 and 0.48N_e/Morgan, than between 0.48 and 1.44 N_e/Morgan. By calculating equivalent N_e/Morgan marker densities in other studies, my results can be compared to those conducted in other species. Using a 2K low-density panel in Border Leicester sheep, equivalent to a SNP density of 0.23N_e/Morgan (assuming N_e=242 [279] and a total genome length of 36.3 Morgans [280]), Hayes *et al.* (2011) [263] achieved an imputation accuracy of approximately 0.73. Results here using the 1K bpMAF SNP panel (0.24N_e/Morgan) thus compare favourably, with accuracies ranging from 0.74 to 0.84 across chromosomes. Equivalent results from a study of Jersey cattle were in the range 0.7 to 0.8 [262].

By comparing results here to those of other studies which have gone on to examine the impact of imputation accuracy on subsequent use of the imputed genotypes, it can be determined whether the imputation accuracies here are likely to be high enough for implementation in the future. Daetwyler *et al.* (2011) [263] observed that the accuracy of GEBV achieved for a sparse subset of markers, as a percentage of that achieved for the dense accuracy, was in all cases greater than the proportion of correctly imputed genotypes. Specifically, when 87.8% of missing genotypes were correctly imputed, 95% of the dense genomic evaluation accuracy was achieved. Furthermore, the imputation accuracy achieved here for the 3K panel (≥0.90), was very similar to that reported by Weigel *et al.* (2010) [281] (0.912) with a 2,942 SNP panel, despite the actual SNP density likely being lower in the Thoroughbred (0.72N_e/Morgan compared to 0.93N_e/Morgan in the Jersey cow, assuming N_e=100 [92] and total genome length of 31.6 Morgans [282]). Weigel *et al.* (2010) [281] demonstrated GEBV accuracies for daughter pregnancy rate (DPR) of 0.642 when imputation (from 2,942 SNPs) was used compared to 0.674 when all SNPs were genotyped (42,552 genotypes). Given these results, it seems likely that a 2K to 3K

LDP would enable sufficiently high imputation accuracies to be useful in the Thoroughbred. The accuracy of imputation and therefore the predictive ability is likely to be slightly greater in the Thoroughbred than in other horse breeds because of the higher LD (and smaller effective population size) [23].

When correlation was used to assess imputation accuracy, the relative rank of the different methods remained unchanged. When compared across SNPs, the two measures of accuracy were correlated (Figure 9-8) but there was a strong bias in the proportion of correctly imputed genotypes due to MAF of the SNPs. Adjusting the proportion of correctly imputed genotypes by the expected accuracy using random imputation significantly reduced the bias, whilst emphasising the greater variation in imputation accuracies seen for SNPs with low MAF. Correlation therefore provides a MAF independent assessment of imputation accuracy, a property which makes it a preferable measure of accuracy to the proportion correctly imputed (a similar conclusion was also reached by Hickey *et al.* (2012) [283] in a study published after this work was done). However, accuracy expressed as the proportion of correctly imputed genotypes is more easily compared to other studies and so is also presented here.

There was considerable variation in imputation success both across SNPs and across samples. The range in imputation accuracies was greatest for the lowest density SNP panels. For example, at a LDP SNP density of $0.09N_e$/Morgan, some SNPs were imputed correctly in all samples, whilst others were correct in only 30% of samples (based on ECA1 bpMAF results). Increasing the density of the LDP led to an increase in the minimum accuracy so that by $1.44N_e$/Morgan, the minimum was around 80%. A similar pattern was seen in the variation across samples, albeit with a slightly smaller range of less than 50%, even at the lowest LDP SNP density. As expected, using random imputation, there was a direct and predictable relationship between the MAF of the SNPs and the accuracy with which they were imputed. This relationship was much less strong when using BEAGLE to impute, except at the lowest LDP SNP densities when presumably the amount of additional information available from LD was low.

A major source of the variation in SNP imputation accuracy was the extent of LD, with variation being seen both between and within chromosomes. There was a tendency for imputation accuracies to be higher for the longer chromosomes (ECA1 and ECA10) and this coincides both with the higher average LD of these chromosomes demonstrated in Chapter 3 and results of Weigel *et al.* (2010) [262]. Within chromosomes, SNPs located within regions of high LD were typically imputed more accurately than those in regions of low LD. The strength of this relationship varied across the four chromosomes being most obvious for ECA10 and least convincing for ECA1. In the case of ECA10, the region of highest LD (and imputation accuracy) also coincided with the hypothesised position of the centromere on this chromosome. However, this relationship was not seen on ECA1, which suggests that the processes underlying the LD may be important.

Given the relationship between MAF and SNP imputation accuracy, one hypothesis is that, by preferentially selecting LDP SNPs to be more informative, that is, to have a high MAF, whilst at the same time ensuring relatively consistent coverage across the chromosome, imputation accuracies could be improved. The increase in accuracy achieved by using a stochastic algorithm to do this (bpMAF) was small but consistent, ranging from 1.6% to 4.4% for the 2K panel. The biggest increase was seen in the range of SNP panel densities from 768 to 3K and for ECA20 and ECA26. This corresponds with results of Hayes *et al.* (2011) [263] who also saw little benefit in using MAF to select SNPs for LDPs of 5K or more. Using bpMAF LDP SNPs also resulted in a decrease in the across SNP variation in imputation accuracy compared to the bpEQ method.

Given the relationship between LD and SNP imputation accuracy, a further hypothesis is that by taking account of LD during the LDP SNP selection process, the across SNP variation in imputation accuracy could be reduced. The use of LD in the selection of LDP SNPs is not new, and several algorithms exist that use either pairwise or multimarker LD statistics to help select the most informative LDP SNP set [264, 265]. However, in this case, a more natural extension to the approach being used (bpMAF) was simply to adapt the input so that SNP locations were made relative to the amount of LD in the region. In fact, the stochastic algorithm used here

was initially designed to be used with cM locations, in which case patterns of LD related to recombination rate would naturally be incorporated. However, in the case of the Illumina Equine SNP50 BeadChip, all marker locations are given in base pairs and whilst some markers may have corresponding positions on the equine linkage map [22] to re-position all SNPs in this way would be unfeasible. As an alternative, LD map distance was used, which has been shown to have a close relationship with linkage maps [284] and recombination rates, at least to the extent that recombination hot spots can be identified [285].

Using lduMAF LDP SNPs did result in a decrease in the across SNP variation in imputation accuracy as hypothesised, particularly at the lower SNP densities. Figure 9-7 shows that by using this approach, the relationship between LD and imputation accuracy across chromosomes was indeed broken down. The change in mean imputation relative to results using bpMAF was small and inconsistent, with an increase in accuracy seen for ECA26 and a decrease for ECA1. The results shown in Figure 9-7 suggest that when using lduMAF LDP SNPs, whilst there is an increase in imputation accuracy where LD is low (due to a greater concentration of SNPs), the decrease in accuracy where LD is high is relatively greater, leading to a tendency for the mean imputation accuracy to be reduced. It may be possible to increase the overall mean imputation accuracy achieved with the lduMAF SNP set by inserting an additional SNP selection criterion into the algorithm such that the LD between the selected SNPs is reduced (B. Kinghorn 2012, Pers. Comm.). In contrast to the other chromosomes, using lduMAF LDP SNPs on ECA26 led to an increase in imputation accuracy across the entire latter half of the chromosome (for the 1K panel), which suggests some dependence on LD at the level of the chromosome. Further benefit may therefore be seen if this approach was applied at a genome-wide level such that the number of LDP SNPs per chromosome was proportional to the total LD map distance of the chromosome, rather than the base pair length as in the current implementation.

The fact that by using lduMAF the across SNP variation in accuracy can be reduced, but the mean accuracy of imputation not reliably increased, leads to the question of which properties are most important when assessing efficacy of imputation.

Specifically, are higher mean accuracies actually advantageous, even though some SNPs and some individuals may actually be poorly imputed? Or are lower mean accuracies preferable if a greater proportion of SNPs (or individuals) are imputed at a given threshold? Furthermore, should the focus be on the distribution of accuracies across SNPs or across samples? The answer to these questions almost certainly lies in the intended use of the imputed genotypes and there does not seem to be any methodology for assessing imputation performance in the context of its application at this time. Given the likely dependence on use of the imputed genotypes, a solution might be derived in the form of a loss function, whereby, a strategy is selected that is consistent with the actual loss experienced, in the context of a particular problem, for example, a GWAS. Alternatively, the success of imputation might be measured according to some utility function, such that when plotted against, for example, chromosome position as in Figure 9-7, the area under the curve becomes informative with respect to the optimum scenario. The development of such a method to assess imputation success in the context of its application would allow a more quantitative comparison of the different SNP sets used in this study, in particular the relative usefulness of the novel lduMAF approach and might also help to evaluate the relative cost of a missing genotype, as compared to a wrong genotype. At the current time, the uncertainty in imputed genotypes is generally dealt with either by setting a minimum threshold on the confidence with which a genotype is assigned or by using a genotype probability in place of the most likely genotype [92]. However, these approaches do not take into account the variance associated with the estimates. Meanwhile, several approaches have been proposed to take into account the uncertainty of imputations when performing GWAS and tend to be based on Bayesian frameworks [258, 286-290]. Utilising these with simulated data might be the next step to resolve the issue of whether uniform accuracy is more desirable than a high mean, at least for GWAS.

The transferability of LDPs both across breeds and across countries within breeds is an important consideration when designing a LDP. Here data from a cohort of US Thoroughbreds was used to evaluate the impact of geographical origin on the efficacy of imputation. A 2K panel density was used to assess the between-

population efficacy since at this density, the within-population accuracy was always greater than 0.8. Whilst comparisons involved both reference sets and test sets of differing sizes, my work (data not shown) and that of others [263, 281] has shown reference set size not to have a major impact on imputation accuracy across the range used here. Using a UK reference set to impute genotypes in a US test set had no impact on the proportion of correctly imputed genotypes for ECA1, regardless of the SNP selection strategy. A slight decrease in accuracy was observed for ECA26, but accuracies remained above 0.8.

When the UK reference set was replaced with a US reference panel for predicting the US data, imputation accuracies increased above those seen both in the first between-population assessment and in the within-population assessment of UK horses, despite the bpMAF and lduMAF LDP SNP sets being tailored to the UK dataset. This increase is presumably due to the higher average relationships in the US dataset which has been shown to improve imputation accuracy [283]. Whilst this difference in sample properties between the US and the UK samples does not allow the direct comparison of the corresponding within-population results, the fact that there was no difference in the relative increase between the bpEQ LDPs (not population dependent) and the bpMAF and lduMAF LDPs (selected using the UK dataset), suggests that the LDP SNP sets are equally appropriate for both populations. This implies that similar LD patterns exist in both populations. This in turn indicates that either, the genetic differentiation between the UK and US populations is small, or, that similar LD structure exists due to a common recombinatorial background. The high correlation between MAF across the two populations lends some credence to the former argument, whilst the relationship between LD and the centromere position in ECA10 suggests the latter is also relevant. The small difference in accuracy observed here correspond to results of Weigel *et al.* (2010) [262] who found that subdividing their population of Jersey cattle by country of registration, allele frequency similarity or by sire family, did not improve imputation accuracy. Whilst there appears to still be an advantage to using the bpMAF or lduMAF LDP SNP sets in a population of the same breed but different geographical origin, whether this will still be true when looking across breeds remains to be seen. The frequent sharing of

major haplotypes among diverse horse populations [23] suggests some accuracy should be maintained; further indications may be sought from across breed comparisons of allele frequencies. Furthermore, with allele frequencies and patterns of LD constantly changing over time, LDPs based on these properties may become less effective as generations progress.

The loss in accuracy that occurs as a result of using lower density SNP panels must also be considered alongside the cost savings that would be achieved and this is demonstrated in Figure 9-9. The fact that some SNPs will require more than one beadtype on the chip has not been accounted for, although preliminary analysis suggests there are no A/T or C/G SNPs on the chromosomes analysed in this study and so the impact of this is expected to be minimal. There is no difference in cost between genotyping 768 and 2K SNPs, or between genotyping 3,072 and 6K SNPs and therefore, the logical choice is between a 384, a 2K and a 6K SNP panel; these options offer 42%, 84% and 96% of the accuracy for 17%, 29% and 40% of the cost, respectively. Whilst the increase in cost is almost linear across this range, increasing by $20 (USD) from 384 to 2K and another $20 from 2K to 6K, the increase in accuracy is more than three times greater from 384 to 2K than from 2K to 6K, suggesting that a 2K SNP panel represents better value for money. However, certain uses may demand certain accuracies, in which case cost may be less important.

The results of this study show that it is possible to impute genotypes from low to high density in the Thoroughbred with reasonable to high accuracy. The haplotype phasing and imputation program used here (BEAGLE) has been shown to perform similarly to other available software [261, 263, 275] and therefore the results are thought to be representative. An investigation of the source of variation in imputation accuracy revealed dependence both on MAF of the SNPs being imputed and on the underlying LD structure. Confirmation of some of these results has since become available in a publication which explored similar dependencies when imputation was carried out in maize [283]. Whilst equidistant LDP SNPs work well, optimising LDP SNP selection to increase their MAF was advantageous, even when LDPs were subsequently used in a population of different geographical origin. By using LD map distance in place of base pair position, the variation in imputation

accuracy across SNPs was reduced. Whilst a 1K panel was generally sufficient to ensure more than 80% of genotypes were correctly imputed, inference from other studies suggests that a 2-3K panel would ensure the subsequent loss in accuracy in, for example, GS analyses was minimal. Furthermore, the relationship between accuracy and genotyping costs for the different LDPs, suggest a 2K SNP panel would represent good value for money. More work is needed to evaluate the impact of between breed differences on imputation accuracy. As well as enabling use of low-density SNP panels as a low cost alternative to high-density genotyping, imputation provides a means by which datasets from different genotyping platforms can be combined, something which will be necessary as researchers start to use the recently developed equine 70K SNP chip.

Chapter 10: General discussion

This chapter provides a general overview of the contents of this thesis and is presented in four parts. Part 1 is a brief summary of the main contributions of the thesis. Parts 2 and 3 comprise of a précis of the results contextualised within the limitations of the methods used. The final section considers the future for the application of genomic technologies to the horse.

10.1 Summary

This thesis represents one of the first explorations of the use of high-density SNP genotyping technology in the horse. The potential of the SNP chip to be used to investigate quantitative traits, in particular complex diseases, has been investigated both from a theoretical perspective and empirically, using osteochondrosis (OC) in Thoroughbreds as a case in point. An assessment of Thoroughbred genome coverage offered by the 50K SNP chip, based on a comprehensive analysis of linkage disequilibrium (LD), revealed good average coverage. However, this work also showed considerable variability both in LD and in the distribution of SNPs across the genome. Whilst this is to be expected, especially given that the 50K chip was the first high-density genotyping chip developed for the horse, researchers should be aware that some areas of the genome are not well captured. My efforts to use LD to explore the demographic history of the Thoroughbred population suggested that some information about a population's history can be gleaned in this way. However, a subsequent critique of the method identified a number of limitations relating to the application of the theoretically derived models to data from real populations. This work highlights the need for researchers to be aware of the assumptions underlying such models and to interpret the results of such analyses accordingly. Linkage disequilibrium was also used to examine the quality of the sequence assembly itself. As well as providing molecular geneticists with a potential approach for targeting regions of the genome for re-sequencing in the future, this work served as a reminder of the preliminary nature of the current assembly.

The conclusion from the series of LD analyses performed was that the SNP chip should, in theory, be a useful tool for studying the genetic background of complex

diseases such as OC. However, the empirical evidence from this study suggests that major challenges remain in the application of new genomic approaches to the horse. At the outset of this study, considerable effort had to be invested in deriving a dataset suitable for genomic analysis. The issues faced during this process demonstrate a clear need for the development and implementation of standard collection methodologies for use by equine clinicians and researchers. The results of both the genome-wide association study (GWAS) and the genome-wide evaluation (GWE) of OCD conducted as part of this study were disappointing given the theoretical potential of the SNP chip to capture genetic variation in the population of Thoroughbred horses used. A significant conclusion from this work is therefore that large, well-phenotyped cohorts of horses will be essential if the true potential of genomic technologies are to be realised in the horse. One recommendation to both equine researchers and the equine industry more generally, is to focus their efforts on sample collection.

10.2 Linkage disequilibrium and effective population size: Theory and practice

Linkage disequilibrium has a long history as a parameter of interest to population geneticists, as evidenced by the date range of the references given in Chapters 2 to 5. At the outset of this thesis, the extent and decline of LD in the Thoroughbred had not been comprehensively examined, but was of considerable interest because of its relevance in determining how well genetic variation would be captured by the 50K SNP chip. Whilst the use of r^2 to measure LD has been criticised in the past for reasons discussed in Chapter 5, it remains the most widely used statistic to quantify LD using high-density genotype data. This is due to both its relative ease of calculation and its relevance to GWAS such that, to achieve approximately the same power at the marker locus as would be achieved by genotyping the susceptibility locus, the sample size must be increased by a factor of $1/r^2$ [291]. Therefore, in Chapter 3, r^2 was calculated in order to determine the likely efficacy of such methodologies as GWAS. The analysis revealed high average LD at short distances which declined fairly slowly, reaching baseline levels at around 50Mb. In terms of genome coverage, this high LD points towards a good level of coverage on average.

However, there is expected to be considerable variability in coverage due to both an uneven distribution of markers across the genome and variable recombination rates. Whilst this high average LD is an advantage with respect to power to detect QTL associated with traits, it may make the identification of causal variants by fine-mapping more difficult. This issue has already been raised by Orr *et al.* (2010) [292] who, despite genotyping additional markers, were unable to further refine a 2Mb region identified in a GWAS for dwarfism in Friesian horses. Whilst this problem may be overcome to some extent by exploiting across breed recombination, the high rate of haplotype sharing across breeds [23] may hinder progress by this route.

Whilst this characterisation of LD in the Thoroughbred has given some indication of how well GWAS and the plethora of other methodologies that are reliant on LD between markers and causal variants can be expected to perform, there are many more factors that play a role. The genomic architecture of the trait is another important factor in determining which, if any, causal variants will be identified. Several studies have looked at the impact of the allele frequency of variants and the strength of their phenotypic effects on power of detection, that is, on the ability to reject the null hypothesis of no association whilst controlling for the type I error rate [37, 42, 48]. In general, if causal alleles are rare (minor allele frequency <0.1) and effect sizes small (odds ratio <1.3), then they will not be detected, with realistic sample sizes (below 10,000) [48]. As well as having a direct impact on GWAS power, the allele frequency of variants is also important in the context of LD. In order for the r^2 between a causal variant and a marker variant to be high, the frequencies of the coupled loci must be similar. If, as has been suggested by some, disease causing alleles occur at low frequencies in the population, our power to detect them using relatively common SNPs will necessarily be low. So, whilst calculating summary statistics of r^2 as in Chapter 3 has provided some useful information, it is only a piece of the puzzle.

As well as being used directly as an indication of the genome coverage achieved using the 50K SNP chip, the extent and decline of LD was used here to infer the effective population size (N_e) of the two Thoroughbred populations from which samples were collected. Initially, this analysis simply represented a means to an end.

In order to estimate the predictive accuracy of GWE in the Thoroughbred, using deterministic equations derived by Daetwyler *et al.* (2008) [61], an estimate of N_e was required. Because no pedigree was available for the sampled horses, alternative marker-based methods were explored. Whilst there were several options available for estimating N_e using marker data, some assumed multi-allelic markers whilst others needed two or more samples to be collected from the population over a period of time. The most appropriate method, given the data available for this study, seemed to be that used by Tenesa *et al.* (2007) [139] in which the theoretical relationship between expected r^2, the distance between markers and the N_e of the population was used. This approach was thus implemented in Chapter 3 and two estimates of N_e calculated for each sample. The first, was a single estimate assuming constant N_e, and the second, a series of estimates going back in time, under the assumption of linear change in N_e [138]. The results from this analysis largely coincided with expectations based on pedigree analyses by other authors, the demographic history of the Thoroughbred and the origins of the samples. The small estimated N_e of the population, led to relatively high predicted accuracies for GWE of between 0.15 and 0.30 for trait heritabilities typical of complex diseases and for reference sample sizes similar to those available for this study.

Despite obtaining N_e estimates which were largely in agreement with expectations, a review of literature around the method used revealed considerable variation in the way in which the method had been applied in different studies in the past. Evidently, there was some confusion over exactly how the model should be implemented, for example, whether adjustments were necessary to account for sample size and mutation. This apparent disagreement surrounding the optimal model, combined with the failure of empirical datasets to conform to the theoretical assumptions under which the models had been derived, led to some doubt over the accuracy of the estimates calculated using this approach. Therefore, the application of this method progressed from being a means to an end to being the focus of a new investigation. This investigation (Chapter 4) served as a reminder of the limitations that must apply when theoretically derived formulae are applied to real world data. An alternative approach to estimating N_e from marker data that exploited linkage information in a

different way (the CLE method used in Chapter 5), though useful in certain circumstances, was found to have its own limitations. Such methods will likely continue to be refined with time and new alternative approaches developed to replace them as more is understood about what influences genome structure.

10.3 Genetic control of osteochondrosis

At the outset of this study, no GWAS for OC in the horse had been published. However, prior to the publication of the work in Chapter 7, three such studies had made it to press. This demonstrates both the current relevance of OC to the equine industry and the perceived benefits in tackling the disease using genomics over traditional pedigree-based selection methods. However, given the results to date and the wider context discussed below, it seems likely that there is a long way to go before QTL identified in GWAS will be making any contribution to a reduction in OC. At the present time, despite many QTL having been identified, including one in this thesis (Chapter 7), no validated markers for OC have been published. Furthermore, the attempted validation of QTL in Chapter 7 resulted in at best the tentative confirmation of two regions. In some sense, it is not a surprise that these studies have essentially failed to produce a marketable product. More than ten years ago, the limited potential for MAS to be useful in livestock breeding for complex traits was recognised and people started looking for alternative approaches, namely GWE [51, 52]. Those involved in the research of complex disease in humans are also increasingly turning towards GWE as it has become evident that few genetic variants exist that have a sufficiently large effect on the phenotype to be clinically relevant. Therefore, whilst GWAS of complex disease continue to be carried out, the main motivation must be to improve knowledge of the underlying pathogenesis of the condition which in turn can contribute to developing better treatments and management strategies, rather than to identify markers for the implementation of MAS. Meanwhile, the future for selection to reduce the incidence of complex disease appears to be focused on GWE.

The potential for GWE to be used in the breeding of horses has already been recognised [249, 250]. Interestingly, in the only published application of GS to the horse to date, Ricard *et al.* (2012) [293] suggest GS has little advantage over

traditional BLUP-based selection when applied to showjumping. However, it could be argued that, if similar GEBV accuracies (0.37-0.51) could be achieved with respect to OC, significant progress could be made in reducing disease incidence. Unfortunately, given predicted accuracies (from Chapter 3) in the range of 0.11 to 0.21, the results of the GWE performed in this thesis (Chapter 8) were disappointing, with neither of the models used producing GEBV with any predictive value for the phenotype. However, GEBV generated for a simulated phenotype achieved predicted accuracies, suggesting the failure to achieve expected accuracies with the true data may well be due to characteristics of the disease and/or the sample or incorrect assumptions with respect to these properties, rather than problems with the expectations. For example, power may have been reduced due to factors such as population substructure and misclassification. It is likely that, because of the issues surrounding OC definition discussed in 1.3.4, OC studies will have less power than might be predicted when assuming a perfect sample. Greater sample sizes will therefore be needed to provide a proof of principle for GWE applied to complex disease in horses.

The results in Chapters 7 and 8 will have been influenced to some extent by the type of analysis performed, with both approaches making certain assumptions about the underlying genetic model of disease. The major driver of association analyses is the common disease common variant (CDCV) model, with analyses being designed to distinguish genetic variants which are shared by a large number of affected individuals. If, however, this assumption is false and complex disease occurs as a result of an accumulation of much rarer variants, GWAS will have little power to detect them [40]. Not surprisingly, this has been put forward as one of the potential explanations for the disappointing results of GWAS to date and the problem of the so-called 'missing heritability' (see 1.2). In contrast, GWE might be better placed to exploit these variants but would still require large sample sizes to accurately estimate their effects. In reality, it seems likely that neither of these theories truly represent the highly complex nature of the true genetic architecture of complex traits.

The methods underlying both the GWAS and the GWE performed in this study are biased towards capturing additive genetic variation and such models tend to be

adopted because of their superior power. For example, in GWAS a test of allele counts can be done with a single degree of freedom whilst a test of three genotypes means an additional degree of freedom. However, it should be recognised that, if the genotype risks are not additive, for example in cases of dominance or epistasis, assuming an additive model represents a loss of power [294, 295]. In GWAS, a possible alternative to fitting an exclusively additive model is to take the maximum test statistic from tests designed to detect additive, dominant and recessive models of inheritance [296, 297]. Whilst this strategy is a manageable prospect, the study of epistatic effects in association studies is much more daunting due to the large number of tests that have to be performed. However, the computational burden of such analyses is being rapidly eroded through the use of novel data processing techniques [298] and researchers should soon be able to quantify the contribution of epistasis to genetic variation more easily. In the case of GWE, the use of alternative semi-parametric or non-parametric approaches to GEBV prediction has been advocated to overcome problems relating to the assumption of additive inheritance. De los Campos *et al.* (2009) [299] state that a Reproducing Kernel Hilbert Spaces (RKHS) methodology can be used under any genetic model, whilst Gonzalez-Recio (2008) [53] advocate RKHS and kernel regression methods to account for non-additive genetic effects, including dominance and epistasis. More recently, Gianola *et al.* (2011) [300] have explored the possible application of Bayesian neural networks in capturing the non-linear relationships between predictors. Also of interest when dealing with binary outcome traits such as the OC phenotype used in Chapters 7 and 8, is the machine learning classification procedure of Long *et al*. (2007) [54]. This methodology involves a two-step feature selection procedure such that the most informative SNPs are first selected and then fed into a so-called ‘wrapper’ step, which uses a naïve Bayesian classifier to build the predictive model [54]. These approaches represent a potential alternative to parametric methods for use in cases where the mode of inheritance is not thought to be additive but, initial results suggest that the improvement in the accuracy of GEBV that can be achieved is relatively modest [92].

10.4 The future for genomics in the horse

The story told in this thesis parallels developments in the field of equine research as we have entered the so-called genomics era. As outlined in Chapter 1, fantastic progress has been made with regard to equine genetics and genomics over the last twenty years, with the efforts of a relatively small number of individuals pushing equine genetics up the research agenda worldwide. However, it seems likely that harder times are already upon us, with the funding available for equine research in the current economic climate being severely limited. The emphasis is now as much on exploiting current resources so that they can be maintained as it is on striving to develop new research tools. During the course of this study, Illumina ceased production of the 50K SNP chip due to lack of demand. Whilst Geneseek, now a Neogen company, took on the project of developing a new 70K chip, the future of this product may now also be under threat. Furthermore, the identification of potential sequence errors in Chapter 6 and the lack of information available on the equine specific function of many genes, demonstrates the importance of work being carried out to refine and annotate the draft genome sequence. So, the question becomes, how do we ensure genomics in the horse remains a viable proposition in the future?

For tangible benefits to be seen by the horse industry from using genomic methodologies such as those demonstrated in Chapters 7 and 8, greater sample sizes are needed across a wide range of relevant breeds both for preliminary analysis and for subsequent validation work. Factors that are currently inhibiting the large scale collection of well-phenotyped samples include: cost, logistics, suspicion in the industry and a lack of agreement about disease phenotypes amongst practitioners. Whilst these factors are also relevant to other species, they are often exaggerated in the case of the horse. The cost of producing and maintaining horses is much greater than in most other livestock and it would be extremely costly to establish experimental populations of horses to study complex diseases. Therefore, researchers generally have to work with existing populations of horses and this is a huge logistical task and typically involves visiting a large number of yards and veterinary clinics to collect samples. Researchers are then dependent on the good

will of horse owners whose permission is required to collect samples. Whilst many owners are willing to contribute to genetic studies, others are wary of the impact of such research on the value of their animals. How then can larger sample sizes be secured for future studies?

The genotyping of some samples at lower density and the use of imputation to fill in the 'missing' genotypes, could be a strategy that benefits the equine industry at both the research level and the implementation stage. Provided a reference population of several hundred horses has been genotyped at high density (either for the 50K or the 70K SNP chip), new samples need only be genotyped for several thousand SNPs which can be done at lower cost, thereby increasing the number of horses that can be genotyped for a given budget. The validity of this approach in the Thoroughbred has been demonstrated in Chapter 9 but a significant question that remains is how accurately imputation will be when applied across breeds. In particular, the potential for this approach to be applied to the UK sport horse is not clear given the likely genetic heterogeneity of this population. This represents an important piece of outstanding research that should be prioritised in future.

There is a great need for a more cost-effective approach to the study of disease and performance traits in the horse. A large proportion of the time and cost involved in setting up a project from scratch is invested in sample collection. This aspect of the project also represents the biggest risk in terms of project failure and can therefore make funding providers wary of supporting projects which are reliant on collecting large numbers of samples. Where projects are funded, the conclusion of the project typically sees the samples which were collected relegated to the back of a freezer. Whilst some researchers may intend to use samples again, the reality of the situation is that poor record keeping and a lack of standardised procedures frequently prohibits their further use. Furthermore, under the current system, a series of barriers make subsequent sharing of samples difficult. Taking OC as an example, whilst more than 1,000 horses with a confirmed presence or absence of OC have been genotyped for the 50K SNP chip (based on GWAS published to date), a meta-analysis of this data represents a challenging prospect at both a practical and a political level. Samples have been collected according to different criteria relating to the presence of

radiographic findings as specified by the project veterinarian(s). Therefore, unless full and complete records exist for each horse that was sampled, for example, radiographs and surgery reports, it would be impossible for the same phenotypic criteria to be applied retrospectively to all cases. It was this concern that motivated the collection of additional details about OC lesions in the cases for this study (2.4). Furthermore, all the factors that are relevant at a single project level, for example, issues of population substructure and environmental factors, still apply. Politically, there are issues relating to funding and intellectual property (IP) that make retrospective collaborations difficult to set up. Funding providers need to agree that data can be shared and complex legal documents need to be drafted to ensure that, in the event the results are of value in a patentable sense, the rights to the commercial exploitation of the IP have been clearly established.

An alternative to combining samples in a meta-analysis is to collect samples in a cooperative way in the first instance, for example, through the establishment of a biobank. A biobank is an organised collection of biological samples and associated data from large numbers of people or animals. Samples and data can be made available to researchers to use in genetic and epidemiological studies of complex traits and diseases. The potential of this approach was recognised in the field of biomedical research some time ago, with the first official collaborative biobanks set up at the beginning of the century and several biobanks now containing samples and data from tens of thousands of people. Whilst banks of animal samples exist, they tend to be on a small scale and set up on an *ad hoc* basis. A nationwide or international scale equine biobank could enable the collection of the large numbers of samples required for genomic analyses to be carried out more efficiently. Furthermore, if samples were banked prospectively, researchers would be able to present much less risky research proposals to funders in the future. However, the success of this approach is dependent on a wide range of factors relating both to the equine industry and to the research community. These factors will be investigated in a feasibility study, commissioned by the British Equestrian Federation, to determine whether a UK equine biobank would enable the horse industry to better exploit new

genomic technologies in the future. Some materials which have been produced for this project can be found in Appendix B.

References

1. **Shorter Oxford English Dictionary**. In., 6th edn: OUP Oxford; 2007.
2. Crick FHC: **On protein synthesis**. *Symposia of the Society for Experimental Biology* 1958, **12**:138-163.
3. Sandberg K: **Linkage between the K blood group locus and the 6-PGD locus in horses**. *Animal Blood Groups and Biochemical Genetics* 1974, **5**(3):137-141.
4. Sandberg K, Juneja RK: **Close linkage between the albumin and Gc loci in the horse**. *Animal Blood Groups and Biochemical Genetics* 1978, **9**(3):169-173.
5. Juneja RK, Gahne B, Sandberg K: **Genetic polymorphism and close linkage of two α1-protease inhibitors in horse serum**. *Animal Blood Groups and Biochemical Genetics* 1979, **10**(4):235-251.
6. Bailey E, Stormont C, Suzuki Y, Trommershausen Smith A: **Linkage of loci controlling alloantigens on red blood cells and lymphocytes in the horse**. *Science* 1979, **204**:1317-1319.
7. Chowdhary BP, Raudsepp T: **The Horse Genome Derby: racing from map to whole genome sequence**. *Chromosome Research* 2008, **16**(1):109-127.
8. Pollitt CC, Bell K: **Protease inhibitor system in horses: classification and detection of a new allele**. *Animal Blood Groups and Biochemical Genetics* 1980, **11**(3):235-244.
9. Bowling AT: **The use and efficacy of horse blood typing tests**. *Journal of Equine Veterinary Science* 1985, **5**(4):195-199.
10. Bowling AT, Clark RS: **Blood group and protein polymorphism gene frequencies for seven breeds of horses in the United States**. *Animal Blood Groups and Biochemical Genetics* 1985, **16**(2):93-108.
11. Marklund S, Ellegren H, Eriksson S, Sandberg K, Andersson L: **Parentage testing and linkage analysis in the horse using a set of highly polymorphic microsatellites**. *Animal Genetics* 1994, **25**:19-23.
12. Bowling AT, Eggleston-Stott ML, Byrns G, Clark RS, Dileanis S, Wictum E: **Validation of microsatellite markers for routine horse parentage testing**. *Animal Genetics* 1997, **28**(4):247-252.
13. Montgomery GW, Crawford AM, Penty JM, Dodds KG, Ede AJ, Henry HM, Pierson CA, Lord EA, Galloway SM, Schmack AE *et al*: **The ovine Booroola fecundity gene (FecB) is linked to markers from a region of human chromosome 4q**. *Nature Genetics* 1993, **4**(4):410-414.
14. Montgomery GW, Lord EA, Penty JM, Dodds KG, Broad TE, Cambridge L, Sunden SLF, Stone RT, Crawford AM: **The Booroola Fecundity (FecB) Gene maps to sheep chromosome 6**. *Genomics* 1994, **22**(1):148-153.
15. Charlier C, Coppieters W, Farnir F, Grobet L, Leroy PL, Michaux C, Mni M, Schwers A, Vanmanshoven P, Hanset R *et al*: **The *mh* gene causing double-muscling in cattle maps to bovine Chromosome 2**. *Mammalian Genome* 1995, **6**(11):788-792.
16. Ellegren H, Johansson M, Sandberg K, Andersson L: **Cloning of highly polymorphic microsatellites in the horse**. *Animal Genetics* 1992, **23**(2):133-142.

17. Lindgren G, Sandberg K, Persson H, Marklund S, Breen M, Sandgren B, Carlsten J, Ellegren H: **A primary male autosomal linkage map of the horse genome**. *Genome Research* 1998, **8**(9):951-966.
18. Guérin G, Bailey E, Bernoco D, Anderson I, Antczak DF, Bell K, Binns MM, Bowling AT, Brandon R, Cholewinski G *et al*: **Report of the International Equine Gene Mapping Workshop: male linkage map**. *Animal Genetics* 1999, **30**(5):341-354.
19. Swinburne J, Gerstenberg C, Breen M, Aldridge V, Lockhart L, Marti E, Antczak D, Eggleston-Stott M, Bailey E, Mickelson J *et al*: **First comprehensive low-density horse linkage map based on two 3-generation, full-sibling, cross-bed horse reference families**. *Genomics* 2000, **66**(2):123-134.
20. Guérin G, Bailey E, Bernoco D, Anderson I, Antczak DF, Bell K, Biros I, Bjornstad G, Bowling AT, Brandon R *et al*: **The second generation of the International Equine Gene Mapping Workshop half-sibling linkage map**. *Animal Genetics* 2003, **34**(3):161-168.
21. Penedo MCT, Millon LV, Bernoco D, Bailey E, Binns M, Cholewinski G, Ellis N, Flynn J, Gralak B, Guthrie A *et al*: **International equine gene mapping workshop report: a comprehensive linkage map constructed with data from new markers and by merging four mapping resources**. *Cytogenetic and Genome Research* 2005, **111**(1):5-15.
22. Swinburne JE, Boursnell M, Hill G, Pettitt L, Allen T, Chowdhary B, Hasegawa T, Kurosawa M, Leeb T, Mashima S *et al*: **Single linkage group per chromosome genetic linkage map for the horse, based on two three-generation, full-sibling, crossbred horse reference families**. *Genomics* 2006, **87**(1):1-29.
23. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR *et al*: **Genome sequence, comparative analysis, and population genetics of the domestic horse**. *Science* 2009, **326**(5954):865-867.
24. Rudolph JA, Spier SJ, Byrns G, Rojas CV, Bernoco D, Hoffman EP: **Periodic paralysis in Quarter Horses: a sodium channel mutation disseminated by selective breeding**. *Nature Genetics* 1992, **2**(2):144-147.
25. Rudolph JA, Spier SJ, Byrns G, Hoffman EP: **Linkage of hyperkalaemic periodic paralysis in Quarter horses to the horse adult skeletal muscle sodium channel gene**. *Animal Genetics* 1992, **23**(3):241-250.
26. Wiler R, Leber R, Moore BB, VanDyk LF, Perryman LE, Meek K: **Equine severe combined immunodeficiency: a defect in V(D)J recombination and DNA-dependent protein kinase activity**. *Proceedings of the National Academy of Sciences* 1995, **92**(25):11485-11489.
27. Shin E, Perryman L, Meek K: **A kinase-negative mutation of DNA-PK(CS) in equine SCID results in defective coding and signal joint formation**. *The Journal of Immunology* 1997, **158**(8):3565-3569.
28. Shin EK, Perryman LE, Meek K: **Evaluation of a test for identification of Arabian horses heterozygous for the**. *Journal of the American Veterinary Medicine Association* 1997, **211**(10):1268-1270.

29. Santschi E, Purdy A, Valberg S, Vrotsos P, Kaese H, Mickelson J: **Endothelin receptor B polymorphism associated with lethal white foal syndrome in horses**. *Mammalian Genome* 1998, **9**(4):306-309.
30. Breen M, Lindgren G, Binns M, Norman J, Irvin Z, Bell K, Sandberg K, Ellegren H: **Genetical and physical assignments of equine microsatellites—first integration of anchored markers in horse genome mapping**. *Mammalian Genome* 1997, **8**(4):267-273.
31. Scott IS, Long SE: **Examination of chromosomes in the stallion (Equus-Caballus) during meiosis**. *Cytogenetics and Cell Genetics* 1980, **26**(1):7-13.
32. Green ED, Chakravarti A: **The Human Genome Sequence Expedition: Views from the "Base Camp"**. *Genome Research* 2001, **11**(5):645-651.
33. Makalowski W: **Not Junk After All**. *Science* 2003, **300**(5623):1246-1247.
34. Collins A, Lonjou C, Morton NE: **Genetic epidemiology of single-nucleotide polymorphisms**. *Proceedings of the National Academy of Sciences* 1999, **96**(26):15173-15177.
35. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome**. *Nature Reviews Genetics* 2002, **3**(4):299-309.
36. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ: **Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies**. *PLoS Genetics* 2008, **4**(7):e1000130.
37. Zhao HH, Fernando RL, Dekkers JCM: **Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci**. *Genetics* 2007, **175**(4):1975-1986.
38. Meuwissen THE, Goddard ME: **Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size**. *Genetics* 2007, **176**:2551-2560.
39. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: **A comprehensive review of genetic association studies**. *Genetics in Medicine* 2002, **4**(2):45-61.
40. Iles MM: **What can genome-wide association studies tell us about the genetics of common disease?** *PLoS Genetics* 2008, **4**(2):e33.
41. Maher B: **Personal genomes: The case of the missing heritability**. *Nature* 2008, **456**(7218):18-21.
42. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges**. *Nature Reviews Genetics* 2008, **9**(5):356-369.
43. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing heritability and strategies for finding the underlying causes of complex disease**. *Nature Reviews Genetics* 2010, **11**(6):446-450.
44. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G: **Beyond missing heritability: Prediction of complex traits**. *PLoS Genetics* 2011, **7**(4):e1002051.
45. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**(7265):747-753.

46. Chan EKF, Hawken R, Reverter A: **The combined effect of SNP-marker and phenotype attributes in genome-wide association studies**. *Animal Genetics* 2008, **40**:149-156.
47. Neale BM, Purcell S: **The positives, protocols, and perils of genome-wide association**. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2008, **147B**(7):1288-1294.
48. Wang WYS, Barratt BJ, Clayton DG, Todd JA: **Genome-wide association studies: theoretical and practical concerns**. *Nature Reviews Genetics* 2005, **6**(2):109-118.
49. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, Carta E, Dardano S, Dive M, Fasquelle C *et al*: **Highly effective SNP-based association mapping and management of recessive defects in livestock**. *Nature Genetics* 2008, **40**(4):449-454.
50. Dekkers JCM: **Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons**. *Journal of Animal Science* 2004, **82**(s13):E313-E328.
51. Visscher PM, Haley CS: **On the efficiency of marker-assisted introgression**. *Animal Science* 1999, **68**:59-68.
52. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**:1819-1829.
53. González-Recio O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendano S: **Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers**. *Genetics* 2008, **178**:2305-2313.
54. Long N, Gianola D, Rosa GJM, Weigel KA, Avendaño S: **Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers**. *Journal of Animal Breeding and Genetics* 2007, **124**(6):377-389.
55. Bennewitz J, Solberg T, Meuwissen T: **Genomic breeding value estimation using nonparametric additive regression models**. *Genetics Selection Evolution* 2009, **41**(20).
56. Gianola D, Fernando RL, Stella A: **Genomic-assisted prediction of genetic value with semiparametric procedures**. *Genetics* 2006, **173**(3):1761-1776.
57. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: Genomic selection in dairy cattle: progress and challenges**. *Journal of Dairy Science* 2009, **92**(2):433-443.
58. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited review: reliability of genomic predictions for North American Holstein bulls**. *Journal of Dairy Science* 2009, **92**:16-24.
59. Yang JA, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al*: **Common SNPs explain a large proportion of the heritability for human height**. *Nature Genetics* 2010, **42**(7):565-U131.
60. Lee Sang H, Wray Naomi R, Goddard Michael E, Visscher Peter M: **Estimating missing heritability for disease from genome-wide association studies**. *The American Journal of Human Genetics* 2011, **88**(3):294-305.

61. Daetwyler HD, Villanueva B, Woolliams JA: **Accuracy of predicting the genetic risk of disease using a genome-wide approach**. *Plos One* 2008, **3**:e3395.
62. Gianola D: **Nonparametric and machine learning procedures for genome-enabled prediction of genetic value for quantitative traits**. *Canadian Journal of Animal Science* 2009, **89**(1):123-124.
63. Gabreski N, Brooks S, Miller D, Anczak D: **Mapping of Lavender Foal Syndrome using the EquineSNP50 Chip**. *Journal of Equine Veterinary Science* 2009, **29**(5):321-322.
64. Eberth J, Swerczak T, Bailey E: **Investigation of Dwarfism among miniature horses using the Illumina Horse SNP50 Bead Chip**. *Journal of Equine Veterinary Science* 2009, **29**(5):315-315.
65. Cook D, Gallagher P, Bailey E: **Illumina Equine SNP50 Bead Chip Investigation of Adolescent idiopathic lordosis among American Saddlebred Horses**. *Journal of Equine Veterinary Science* 2009, **29**(5):315-316.
66. Brooks SA, Gabreski N, Miller D, Brisbin A, Brown HE, Streeter C, Mezey J, Cook D, Antczak DF: **Whole-genome SNP association in the horse: Identification of a deletion in Myosin Va responsible for Lavender Foal Syndrome**. *PLoS Genetics* 2010, **6**(4).
67. Bierman A, Guthrie AJ, Harper CK: **Lavender foal syndrome in Arabian horses is caused by a single-base deletion in the MYO5A gene**. *Animal Genetics*, **41**:199-201.
68. Ytrehus B, Carlson CS, Ekman S: **Etiology and pathogenesis of osteochondrosis**. *Veterinary Pathology Online* 2007, **44**(4):429-448.
69. van Grevenhof EM, Ducro B, Weeren P, Tartwijk J, Belt A, Bijma P: **Prevalence of various radiographic manifestations of osteochondrosis and their correlations between and within joints in Dutch Warmblood horses**. *Equine Veterinary Journal* 2009, **41**(1):11-16.
70. Jeffcott LB, Henson FMD: **Studies on growth cartilage in the horse and their application to aetiopathogenesis of dyschondroplasia (osteochondrosis)**. *Veterinary Journal* 1998, **156**(3):177-192.
71. Jeffcott LB: **Osteochondrosis in horses**. *In Practice* 1997, **19**(2):64-71.
72. McIlwraith CW: **Inferences from referred clinical cases of osteochondritis dissecans**. *Equine Veterinary Journal* 1993, **25**(S16):27-30.
73. McIlwraith CW: **Lameness in the Young Horse: Osteochondrosis**. In: *Adams and Stashak's Lameness in Horses.* Edited by Baxter GM, 6 edn. Arnes: Iowa Sate University Press; 2011: 1155-1164.
74. van Weeren PR, Barneveld A: **The effect of exercise on the distribution and manifestation of osteochondrotic lesions in the Warmblood foal**. *Equine Veterinary Journal* 1999, **31**(S31):16-25.
75. van Weeren PR, Sloet vO-O, Barneveld A: **The influence of birth weight, rate of weight gain and final achieved height and sex on the development of osteochondrotic lesions in a population of genetically predisposed Warmblood foals**. *Equine Veterinary Journal* 1999, **31**(S31):26-30.
76. Lepeule J, Bareille N, Robert C, Ezanno P, Valette JP, Jacquet S, Blanchard G, Denoix JM, Seegers H: **Association of growth, feeding practices and exercise conditions with the prevalence of Developmental Orthopaedic**

Disease in limbs of French foals at weaning. *Preventive Veterinary Medicine* 2009, **89**(3-4):167-177.

77. Pieramati C, Pepe M, Silvestrelli M, Bolla A: **Heritability estimation of osteochondrosis dissecans in Maremmano horses**. *Livestock Production Science* 2003, **79**(2-3):249-255.
78. Dik KJ, Enzerink E, van Weeren PR: **Radiographic development of osteochondral abnormalities, in the hock and stifle of Dutch Warmblood foals, from age 1 to 11 months**. *Equine Veterinary Journal* 1999, **31**(S31):9-15.
79. Lykkjen S, Roed KH, Dolvik NI: **Osteochondrosis and osteochondral fragments in Standardbred trotters: Prevalence and relationships**. *Equine Veterinary Journal* 2012, **44**(3):332-338.
80. Schougaard H, Ronne JF, Phillipson J: **A radiographic survey of tibiotarsal osteochondrosis in a selected population of trotting horses in Denmark and its possible genetic significance**. *Equine Veterinary Journal* 1990, **22**(4):288-289.
81. Dempster ER, Lerner IM: **Heritability of threshold characters**. *Genetics* 1950, **35**(2):212-236.
82. Robertson A: **Experimental design in the evaluation of genetic parameters**. *Biometrics* 1959, **15**(2):219-226.
83. Tallis GM: **Sampling errors of genetic correlation coefficients calculated from analyses of variance and covariance**. *Australian Journal of Statistics* 1959, **1**(2):35-43.
84. Stock K, Distl O, Hoeschele I: **Bayesian estimation of genetic parameters for multivariate threshold and continuous phenotypes and molecular genetic data in simulated horse populations using Gibbs sampling**. *BMC Genetics* 2007, **8**(19).
85. Stock KF, Hoeschele I, Distl O: **Estimation of genetic parameters and prediction of breeding values for multivariate threshold and continuous data in a simulated horse population using Gibbs sampling and residual maximum likelihood**. *Journal of Animal Breeding and Genetics* 2007, **124**(5):308-319.
86. Wittwer C, Hamann H, Rosenberger E, Distl O: **Genetic parameters for the prevalence of osteochondrosis in the limb joints of South German Coldblood horses**. *Journal of Animal Breeding and Genetics* 2007, **124**:302-307.
87. van Grevenhof EM, Schurink A, Ducro BJ, van Weeren PR, van Tartwijk JMFM, Bijma P, van Arendonk JAM: **Genetic variables of various manifestations of osteochondrosis and their correlations between and within joints in Dutch warmblood horses**. *Journal of Animal Science* 2009, **87**(6):1906-1912.
88. Philipsson J, Andréasson E, Sandgren B, Dalin G, Carlsten J: **Osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints in Standardbred trotters. II. Heritability**. *Equine Veterinary Journal* 1993, **25**(S16):38-41.
89. Stock KF, Hamann H, Distl O: **Estimation of genetic parameters for the prevalence of osseous fragments in limb joints of Hanoverian**

Warmblood horses. *Journal of Animal Breeding and Genetics* 2005, **122**(4):271-280.

90. Stock KF, Distl O: **Genetic correlations between osseous fragments in fetlock and hock joints, deforming arthropathy in hock joints and pathologic changes in the navicular bones of Warmblood riding horses**. *Livestock Science* 2006, **105**(1-3):35-43.
91. Lepeule J, Seegers H, Rondeau V, Robert C, Denoix JM, Bareille N: **Risk factors for the presence and extent of Developmental Orthopaedic Disease in the limbs of young horses: Insights from a count model**. *Preventive Veterinary Medicine* 2011, **101**(1-2):96-106.
92. Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ: **High-resolution haplotype block structure in the cattle genome**. *BMC Genetics* 2009, **10**(19).
93. Castle K, Jeffcott LB, Raadsma HW, Tammen I, Nicholas FW: **Development of genomic tools to predict the occurrence of osteochondrosis in Australian Thoroughbreds**. Edited by Corporation RIRAD. Australia: RIRDR; 2010.
94. Robertson A: **The sampling variance of the genetic correlation coefficient**. *Biometrics* 1959, **15**(3):469-485.
95. Jørgensen B, Andersen S: **Genetic parameters for osteochondrosis in Danish Landrace and Yorkshire boars and correlations with leg weakness and production traits**. *Animal Science* 2000, **71**(3):427-434.
96. Löhring K: **Genome scan for quantitative trait loci (QTL) for osteochondrosis in Hanoverian Warmblood horses using as optimised microsatellite marker set**. University of Veterinary Medicine Hannover; 2003.
97. Dierks C, Löhring K, Lampe V, Wittwer C, Drögemüller C, Distl O: **Genome-wide search for markers associated with osteochondrosis in Hanoverian warmblood horses**. *Mammalian Genome* 2007, **18**:739-747.
98. Wittwer C, Löhring K, Drögemüller C, Hamann H, Rosenberger E, Distl O: **Mapping quantitative trait loci for osteochondrosis in fetlock and hock joints and palmar/plantar osseus fragments in fetlock joints of South German Coldblood horses**. *Animal Genetics* 2007, **38**(4):350-357.
99. Lampe V: **Fine mapping of quantitative trait loci (QTL) for osteochondrosis in Hanoverian warmblood horses**. University of Veterinary Medicine Hannover; 2009.
100. Komm K: **Fine mapping of quantitative trait loci (QTL) for osteochondrosis in Hanoverian warmblood horses**. University of Veterinary Medicine Hannover; 2010.
101. Teyssèdre S, Dupuis MC, Guérin G, Schibler L, Denoix JM, Elsen JM, Ricard A: **Genome-wide association studies for osteochondrosis in French Trotters**. *Journal of Animal Science* 2011.
102. Lykkjen S, Dolvik NI, McCue ME, Rendahl AK, Mickelson JR, Roed KH: **Genome-wide association analysis of osteochondrosis of the tibiotarsal joint in Norwegian Standardbred trotters**. *Animal Genetics* 2010, **41**:111-120.
103. Cardon LR, Bell JI: **Association study designs for complex diseases**. *Nature Reviews Genetics* 2001, **2**(2):91-99.

104. Lampe V, Dierks C, Distl O: **Refinement of a quantitative trait locus on equine chromosome 5 responsible for fetlock osteochondrosis in Hanoverian warmblood horses**. *Animal Genetics* 2009, **40**(4):553-555.
105. Purcell S: **PLINK**. v1.06; 2009
106. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ *et al*: **PLINK: A tool set for whole-genome association and population-based linkage analyses**. *The American Journal of Human Genetics* 2007, **81**(3):559-575.
107. McIlwraith CW: **Arthroscopic surgery for osteochondral chip fragments and other lesions not requiring internal fixation in the carpal and fetlock joints of the equine athlete: What have we learned in 20 years?** *Clinical Techniques in Equine Practice* 2002, **1**(4):200-210.
108. Preston SA, Zimmel DN, Chmielewski TL, Trumble TN, Brown MP, Boneau JC, Hernandez JA: **Prevalence of various presale radiographic findings and association of findings with sales price in Thoroughbred yearlings sold in Kentucky**. *Journal of the American Veterinary Medical Association* 2010, **236**(4):440-445.
109. McIlwraith CW: **Developmental orthopedic disease: problems of limbs in young horses**. *Journal of Equine Veterinary Science* 2004, **24**(11):475-479.
110. Baxter GM, Turner AS: **Diseases of the bone and related structures**. In: *Adams' Lameness in Horses.* Edited by Stashak TS, 5 edn. USA: Lippincott Williams & Wilkins; 2002: 401-457.
111. Sandgren B, Dalin G, Carlsten J: **Osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints in Standardbred trotters. I. Epidemiology**. *Equine Veterinary Journal* 1993, **25**(S16):31-37.
112. Dalin G, Sandgren B, Carlsten J: **Plantar osteochondral fragments in the metatarsophalangeal joints in Standardbred trotters; result of osteochondrosis or trauma?** *Equine Veterinary Journal* 1993, **25**(S16):62-65.
113. Kawkak CE, McIlwraith CW: **Proximodorsal first phalanx osteochondral chip fragmentation in 336 horses**. *Equine Veterinary Journal* 1994, **26**(5):392-396.
114. Bertone AL: **Lamness: The Fetlock**. In: *Adams' Lameness in Horses.* Edited by Stashak TS, 5 edn. USA: Lippincott Williams & Wilkins; 2002: 768-799.
115. Ellis DR: **Fractures of the proximal sesamoid bones in Thoroughbred foals**. *Equine Veterinary Journal* 1979, **11**(1):48-52.
116. Bramlage LR, Auer JA: **Diagnosis, assessment, and treatment strategies for angular limb deformities in the foal**. *Clinical Techniques in Equine Practice* 2006, **5**(4):259-269.
117. Vanderperren K, Saunders JH: **Diagnostic imaging of the equine fetlock region using radiography and ultrasonography. Part 2: The bony disorders**. *The Veterinary Journal* 2009, **181**(2):123-136.
118. Schnabel LV, Bramlage LR, Mohammed HO, Embertson RM, Ruggles AJ, Hopper SA: **Racing performance after arthroscopic removal of apical sesamoid fracture fragments in Thoroughbred horses age <2 years: 151 cases (1989–2002)**. *Equine Veterinary Journal* 2007, **39**(1):64-68.

119. Wang JL: **Estimation of effective population sizes from data on genetic markers**. *Philosophical Transactions of the Royal Society B-Biological Sciences* 2005, **360**:1395-1409.
120. Bannasch D, Lohi H, Wade CM, Mickelson JR, Hemman K, Haase B, Berger J, Raekallio M, Obexer-Ruff G, Kruft M *et al*: **Genome wide association analysis of a behavioural vice in horses**. In: *8th Dorothy Russell Havemeyer Foundation International Equine Genome Mapping Workshop: 22-25 July 2009; Suffolk, UK*. 15.
121. Blott S, Boursnell M, Bramlage L, Fox-Clipsham L, Helwegen M, Hillyer L, McIlwraith CW, Newton JR, Parkin TDH, Sibbons C *et al*: **Whole genome association mapping for catastrophic fracture, RER and OCD in Thoroughbred horses**. In: *8th Dorothy Russell Havemeyer Foundation International Equine Genome Mapping Workshop: 22-25 July 2009; Suffolk, UK*. 17.
122. Lykkjen S, Dolvik NI, Mickelson JR, Roed KH: **Genetic studies of osteochondrosis dissecans (OCD) in the hock and proximoplantar osteochondral fragments (POF) in the fetlock joints of Norwegian Standardbred trotters**. In: *8th Dorothy Russell Havemeyer Foundation International Equine Genome Mapping Workshop: 22-25 July 2009; Suffolk, UK*. 20.
123. Drögemüller M, Drögemüller C, Welle M, Straub R, Poncet PA, Rieder S, Leeb T: **Mapping of Caroli liver fibrosis (CLF) in Franches-Montagnes horses**. In: *8th Dorothy Russell Havemeyer Foundation International Equine Genome Mapping Workshop: 22-25 July 2009; Suffolk, UK*. 16.
124. Falconer DS, Mackay TFC: **Introduction to quantitative genetics**, 4 edn. Malaysia: Pearson Education Limited; 1996.
125. Hill EW, Bradley DG, Al-Barody M, Ertugrul O, Splan RK, Zakharov I, Cunningham EP: **History and integrity of thoroughbred dam lines revealed in equine mtDNA variation**. *Animal Genetics* 2002, **33**:287-294.
126. Tozaki T, Hirota K, Hasegawa T, Tomita M, Kurosawa M: **Prospects for whole genome linkage disequilibrium mapping in Thoroughbreds**. *Gene* 2005, **346**:127-132.
127. Heifetz EM, Fulton JE, O'Sullivan N, Zhao H, Dekkers JCM, Soller M: **Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations**. *Genetics* 2005, **171**(3):1173-1181.
128. Khatkar M, Nicholas F, Collins A, Zenger K, Cavanagh J, Barris W, Schnabel R, Taylor J, Raadsma H: **Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel**. *BMC Genomics* 2008, **9**(187).
129. McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J: **Linkage disequilibrium in domestic sheep**. *Genetics* 2002, **160**(3):1113-1122.
130. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME: **Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle**. *Genetics* 2008, **179**:1503-1512.

131. Qanbari S, E.C.G.Pimentel, J.Tetens, G.Thaller, P.Lichtner, A.R.Sharifi, H.Simianer: **The pattern of linkage disequilibrium in German Holstein cattle**. *Animal Genetics* 2009, **41**(4):346-356.
132. Thévenon S, Dayo GK, Sylla S, Sidibe I, Berthier D, Legros H, Boichard D, Eggen A, Gautier M: **The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies**. *Animal Genetics* 2007, **38**(3):277-286.
133. Goddard KAB, Hopkins PJ, Hall JM, Witte JS: **Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations**. *American Journal of Human Genetics* 2000, **66**(1):216-234.
134. Toosi A, Fernando RL, Dekkers JCM: **Genomic selection in admixed and crossbred populations**. *Journal of Animal Science* 2010, **88**(1):32-46.
135. Hill WG, Robertson A: **Linkage disequilibrium in finite populations**. *Theoretical and Applied Genetics* 1968, **38**(6):226-231.
136. Weir BS: **Genetic Data Analysis II: Methods for Discrete Population Genetic Data**. Canada: Sinauer Associates, Inc.; 1996.
137. Sved JA: **Linkage disequilibrium and homozygosity of chromosome segments in finite populations**. *Theoretical Population Biology* 1971, **2**:125-141.
138. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME: **Novel multilocus measure of linkage disequilibrium to estimate past effective population size**. *Genome Research* 2003, **13**(4):635-643.
139. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM: **Recent human effective population size estimated from linkage disequilibrium**. *Genome Research* 2007, **17**:520-526.
140. R Development Core Team VA: **R: A Language and Environment for Computing**. v2.10.0; 2009
141. Hill WG: **Estimation of effective population-size from data on linkage disequilibrium**. *Genetical Research* 1981, **38**(3):209-216.
142. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A: **Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels**. *American Journal of Human Genetics* 2001, **69**(4):831-843.
143. Hudson RR: **The sampling distribution of linkage disequilibrium under an infinite allele model without selection**. *Genetics* 1985, **109**(3):611-631.
144. DerSimonian R, Laird N: **Meta-analysis in clinical trials**. *Controlled Clinical Trials* 1986, **7**(3):177-188.
145. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A: **Linkage disequilibrium patterns of the human genome across populations**. *Human Molecular Genetics* 2003, **12**(7):771-776.
146. Abasht B, Sandford E, Arango J, Settar P, Fulton J, O'Sullivan N, Hassen A, Habier D, Fernando R, Dekkers J *et al*: **Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations**. *BMC Genomics* 2009, **10**(Suppl 2):S2.
147. Zhao H, Nettleton D, Soller M, Dekkers JCM: **Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of**

linkage disequilibrium between markers and QTL. *Genetical Research* 2005, **86**(1):77-87.
148. Muir WM, Wong GK, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, Megens HJ, Zhang HM, McKay JC, McLeod S *et al*: **Review of the initial validation and characterization of a 3K chicken SNP array**. *World's Poultry Science Journal* 2008, **64**(02):219-226.
149. Mahon GAT, Cunningham EP: **Inbreeding and the inheritance of fertility in the thoroughbred mare**. *Livestock Production Science* 1982, **9**(6):743-754.
150. Cunningham EP, Dooley JJ, Splan RK, Bradley DG: **Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses**. *Animal Genetics* 2001, **32**(6):360-364.
151. Woolliams JA, Bijma P: **Predicting rates of inbreeding in populations undergoing selection**. *Genetics* 2000, **154**(4):1851-1864.
152. Binns MM, Boehler DA, Bailey E, Lear TL, Cardwell JM, Lambert DH: **Inbreeding in the Thoroughbred horse**. *Animal Genetics* 2012, **43**(3):340-342.
153. Du FX, Clutter AC, Lohuis MM: **Characterizing linkage disequilibrium in pig populations**. *International Journal of Biological Sciences* 2007, **3**(3):166-178.
154. Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF: **Accuracy of genomic selection using different methods to define haplotypes**. *Genetics* 2008, **178**(1):553-561.
155. Goddard ME: **Genomic selection: prediction of accuracy and maximisation of long-term response**. *Genetica* 2008, **136**(2):245-257.
156. Hill WG: **Note on effective population-size with overlapping generations**. *Genetics* 1979, **92**(1):317-322.
157. Hill WG: **Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population**. *Theoretical Population Biology* 1975, **8**(2):117-126.
158. Ohta T, Kimura M: **Linkage disequilibrium due to random genetic drift**. *Genetical Research* 1969, **13**:47-55.
159. Ohta T, Kimura M: **Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation**. *Genetics* 1969, **63**(1):229-238.
160. Ohta T, Kimura M: **Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population**. *Genetics* 1971, **68**:571-580.
161. Sved JA, Feldman MW: **Correlation and probability methods for one and two loci**. *Theoretical Population Biology* 1973, **4**(1):129-132.
162. Weir BS, Hill WG: **Effect of mating structure on variation in linkage disequilibrium**. *Genetics* 1980, **95**(2):477-488.
163. Flury C, Tapio M, Sonstegard T, Drögemüller C, Leeb T, Simianer H, Hanotte O, Rieder S: **Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium**. *Journal of Animal Breeding and Genetics* 2010, **127**(5):339-347.
164. Kim ES, Kirkpatrick BW: **Linkage disequilibrium in the North American Holstein population**. *Animal Genetics* 2009, **40**(3):279-288.

165. England PR, Cornuet JM, Berthier P, Tallmon DA, Luikart G: **Estimating effective population size from linkage disequilibrium: severe bias in small samples**. *Conservation Genetics* 2006, **7**(2):303-308.
166. Waples RS: **A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci**. *Conservation Genetics* 2006, **7**(2):167-184.
167. Goyache F, Álvarez I, Fernández I, Pérez-Pardal L, Royo LJ, Lorenzo L: **Usefulness of molecular-based methods for estimating effective population size in livestock assessed using data from the endangered black-coated Asturcón pony**. *Journal of Animal Science* 2011, **89**(5):1251-1259.
168. McVean GAT: **A genealogical interpretation of linkage disequilibrium**. *Genetics* 2002, **162**(2):987-991.
169. Cockerham CC, Weir BS: **Digenic descent measures for finite populations**. *Genetical Research* 1977, **30**:121-147.
170. Charlesworth B, Charlesworth D: **Elements of Evolutionary Genetics**. Greenwood Village, CO: Roverts & Co.; 2010.
171. Hedrick PW: **Gametic disequilibrium measures: Proceed with caution**. *Genetics* 1987, **117**(2):331-341.
172. Nordborg M, Tavaré S: **Linkage disequilibrium: what history has to tell us**. *Trends in Genetics* 2002, **18**(2):83-90.
173. McEvoy BP, Powell JE, Goddard ME, Visscher PM: **Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs**. *Genome Research* 2011, **21**:821-829.
174. Sved JA, McRae AF, Visscher PM: **Divergence between human populations estimated from linkage disequilibrium**. *The American Journal of Human Genetics* 2008, **83**(6):737-743.
175. Uimari P, Tapio M: **Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds**. *Journal of Animal Science* 2011, **89**(3):609-614.
176. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J *et al*: **The diploid genome sequence of an Asian individual**. *Nature* 2008, **456**(7218):60-65.
177. Wray NR: **Allele frequencies and the r^2 measure of linkage disequilibrium: Impact on design and interpretation of association studies**. *Twin Research and Human Genetics* 2005, **8**(2):87-94.
178. Griffiths RC, Marjoram P: **Ancestral inference from samples of DNA sequences with recombination**. *Journal of Computational Biology* 1996, **3**(4):479-502.
179. Kuhner MK, Yamato J, Felsenstein J: **Maximum likelihood estimation of recombination rates from population data**. *Genetics* 2000, **156**(3):1393-1401.
180. Nielsen R: **Estimation of population parameters and recombination rates from single nucleotide polymorphisms**. *Genetics* 2000, **154**(2):931-942.
181. Hudson RR: **Two-locus sampling distributions and their application**. *Genetics* 2001, **159**(4):1805-1817.

182. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome**. *Science* 2004, **304**(5670):581-584.
183. McVean G, Awadalla P, Fearnhead P: **A coalescent-based method for detecting and estimating recombination from gene sequences**. *Genetics* 2002, **160**(3):1231-1241.
184. Andolfatto P, Wall JD: **Linkage disequilibrium patterns across a recombination gradient in African Drosophila melanogaster**. *Genetics* 2003, **165**(3):1289-1305.
185. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation**. *Bioinformatics* 2002, **18**(2):337-338.
186. Hudson RR: **Properties of a neutral allele model with intragenic recombination**. *Theoretical Population Biology* 1983, **23**(2):183-201.
187. Lanczos C: **A precision approximation of the gamma function**. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 1964, **1**:86-96.
188. Press WH, Flannery BP, Teukolsky SA, Vetterling WT: **Numerical Recipes, The Art of Scientific Computing**. Cambridge: Cambridge University Press; 1986.
189. Waples RS: **Genetic estimates of contemporary effective population size: to what time periods do the estimates apply?** *Molecular Ecology* 2005, **14**(11):3335-3352.
190. Russell JC, Fewster RM: **Evaluation of the linkage disequilibrium method for estimating effective population size**. In: *Modeling Demographic Processes In Marked Populations.* Edited by Thomson DL, Cooch EG, Conroy MJ, vol. 3. New York, U.S.: Springer US; 2009: 291-320.
191. Uleberg E, Meuwissen T: **The complete linkage disequilibrium test: a test that points to causative mutations underlying quantitative traits**. *Genetics Selection Evolution* 2011, **43**(1):1-8.
192. Woolliams J, Corbin L: **Coalescence theory in livestock breeding**. *Journal of Animal Breeding and Genetics* 2012, **129**(4):255-256.
193. Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES: **Whole-genome sequence assembly for mammalian genomes: Arachne 2**. *Genome Research* 2003, **13**(1):91-96.
194. Consortium MGS: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.
195. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
196. Andersson L, Juras R, Ramsey D, Eason-Butler J, Ewart S, Cothran G, Lindgren G: **Equine Multiple Congenital Ocular Anomalies maps to a 4.9 megabase interval on horse chromosome 6**. *BMC Genetics* 2008, **9**(1):88.
197. Go YY, Bailey E, Cook DG, Coleman SJ, MacLeod JN, Chen K-C, Timoney PJ, Balasuriya UBR: **Genome-Wide Association Study Among Four Horse Breeds Identifies a Common Haplotype Associated with the In Vitro CD3+ T Cell Susceptibility/Resistance to Equine Arteritis Virus Infection**. *Journal of Virology* 2011:JVI.06068-06011.

198. Hill EW, McGivney BA, Gu JJ, Whiston R, Machugh DE: **A genome-wide SNP-association study confirms a sequence variant (g.66493737C > T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses**. *BMC Genomics* 2010, **11**(552).
199. Zhang K, Sun F, Waterman MS, Chen T: **Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data**. *The American Journal of Human Genetics* 2003, **73**(1):63-73.
200. Khatkar MS, Zenger KR, Hobbs M, Hawken RJ, Cavanagh JAL, Barris W, McClintock AE, McClintock S, Thomson PC, Tier B *et al*: **A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein-Friesian cattle**. *Genetics* 2007, **176**(2):763-772.
201. Khatkar M, Hobbs M, Neuditschko M, Solkner J, Nicholas F, Raadsma H: **Assignment of chromosomal locations for unassigned SNPs/scaffolds based on pair-wise linkage disequilibrium estimates**. *BMC Bioinformatics* 2010, **11**(1):171.
202. Bohmanova J, Sargolzaei M, Schenkel F: **Characteristics of linkage disequilibrium in North American Holsteins**. *BMC Genomics* 2010, **11**(1):421.
203. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**(2):263-265.
204. Wigginton JE, Cutler DJ, Abecasis GR: **A note on exact tests of Hardy-Weinberg Equilibrium**. *The American Journal of Human Genetics* 2005, **76**(5):887-893.
205. Cannon GB: **The effects of natural selection on linkage disequilibrium and relative fitness in experimental populations of Drosophila melanogaster**. *Genetics* 1963, **48**(9):1201-1216.
206. Raudsepp T, Frönicke L, Scherthan H, Gustavsson I, Chowdhary B: **Zoo-FISH delineates conserved chromosomal segments in horse and man**. *Chromosome Research* 1996, **4**(3):218-225.
207. Scherthan H, Cremer T, Arnason U, Weier H-U, Lima-de-Faria A, Fonicke L: **Comparative chromosome painting discloses homologous segments in distantly related mammals**. *Nature Genetics* 1994, **6**:342-347.
208. Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A *et al*: **Extent and distribution of linkage disequilibrium in three genomic regions**. *The American Journal of Human Genetics* 2001, **68**(1):191-197.
209. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *The American Journal of Human Genetics* 2001, **68**(4):978-989.
210. Oliver LJ, Baird DK, Baird AN, Moore GE: **Prevalence and distribution of radiographically evident lesions on repository films in the hock and stifle joints of yearling Thoroughbred horses in New Zealand**. *New Zealand Veterinary Journal* 2008, **56**(5):202-209.
211. Wittwer C, Hamann H, Rosenberger E, Distl O: **Prevalence of osteochondrosis in the limb joints of south German coldblood horses**.

Journal of Veterinary Medicine Series A-Physiology Pathology Clinical Medicine 2006, **53**(10):531-539.

212. Olivier A, Nurton JP, Guthrie AJ: **An epizoological study of wastage in Thoroughbred racehorses in Gauteng, South Africa**. *Journal of South African Veterinary Association* 1997, **68**(4):125-129.
213. Rossdale PD, Hopes R, Digby NJ, Offord K: **Epidemiological study of wastage among racehorses 1982 and 1983**. *Veterinary Record* 1985, **116**(3):66-69.
214. Grøndahl AM, Dolvik NI: **Heritability estimation of osteochondrosis in the tibiotarsal joint and of bony fragments in the palmar/plantar portion of the metacarpo- and metatarsophalangeal joints of horses**. *Journal of the American Veterinary Medicine Association* 1993, **203**(1):101-104.
215. Andersson-Eklund L, Uhlhorn H, Lundeheim N, Dalin G, Andersson L: **Mapping quantitative trait loci for principal components of bone measurements and osteochondrosis scores in a wild boar x Large White intercross**. *Genetical Research* 2000, **75**(2):223-230.
216. Lee GJ, Archibald AL, Garth GB, Law AS, Nicholson D, Barr A, Haley CS: **Detection of quantitative trait loci for locomotion and osteochondrosis-related traits in Large White X Meishan pigs**. *Animal Science* 2003, **76**(2):155-165.
217. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: an R library for genome-wide association analysis**. *Bioinformatics* 2007, **23**(10):1294-1296.
218. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: A tool for genome-wide complex trait analysis**. *The American Journal of Human Genetics* 2011, **88**(1):76-82.
219. Gilmour AR, Gogel BJ, Cullis BR, Thompson R: **ASReml User Guide Release 3.0**. Hemel Hempstead, UK: VSN International Ltd.; 2009.
220. Henderson CR: **Best linear unbiased estimation and prediction under a selection model**. *Biometrics* 1975, **31**(2):423-447.
221. Cailliez F: **The analytical solution of the additive constant problem**. *Psychometrika* 1983, **48**(2):305-308.
222. Cox TF, Cox MAA: **Multidimensional Scaling**. London: Chapman and Hall; 1994.
223. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: **Score tests for association between traits and haplotypes when linkage phase is ambiguous**. *The American Journal of Human Genetics* 2002, **70**(2):425-434.
224. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al*: **The structure of haplotype blocks in the human genome**. *Science* 2002, **296**(5576):2225-2229.
225. Hamann H, Distl O: **Genetic variability in Hanoverian warmblood horses using pedigree analysis**. *Journal of Animal Science* 2008, **86**(7):1503-1513.
226. Göring HHH, Terwilliger JD, Blangero J: **Large upward bias in estimation of locus-specific effects from genomewide scans**. *The American Journal of Human Genetics* 2001, **69**(6):1357-1369.
227. Clark AG: **The role of haplotypes in candidate gene studies**. *Genetic Epidemiology* 2004, **27**(4):321-333.

228. Sue N, Jack BHA, Eaton SA, Pearson RCM, Funnell APW, Turner J, Czolij R, Denyer G, Bao S, Molero-Navajas JC *et al*: **Targeted disruption of the Basic Kruppel-Like Factor Gene (Klf3) reveals a role in adipogenesis**. *Molecular and Cellular Biology* 2008, **28**(12):3967-3978.
229. Zhang J, Yang C, Brey C, Rodriguez M, Oksov Y, Gaugler R, Dickstein E, Huang CH, Hashmi S: **Mutation in Caenorhabditis elegans Krüppel-like factor, KLF-3 results in fat accumulation and alters fatty acid composition**. *Experimental Cell Research* 2009, **315**(15):2568-2580.
230. Bieker JJ: **Krüppel-like Factors: Three fingers in many pies**. *Journal of Biological Chemistry* 2001, **276**(37):34355-34358.
231. Clarkin CE, Allen S, Kuiper NJ, Wheeler BT, Wheeler-Jones CP, Pitsillides AA: **Regulation of UDP-glucose dehydrogenase is sufficient to modulate hyaluronan production and release, control sulfated GAG synthesis, and promote chondrogenesis**. *Journal of Cellular Physiology* 2011, **226**(3):749-761.
232. Kuroki K, Cook JL, Tomlinson JL, Kreeger JM: **In vitro characterization of chondrocytes isolated from naturally occurring osteochondrosis lesions of the humeral head of dogs**. *American Journal of Veterinary Research* 2002, **63**(2):186-193.
233. Lillich JD, Bertone AL, Malemud CJ, Weisbrode SE, Ruggles AJ, Stevenson S: **Biochemical, histochemical, and immunohistochemical characterization of distal tibial osteochondrosis in horses**. *American Journal of Veterinary Research* 1997, **58**(1):89-98.
234. Bertone AL, Bramlage LR, McIlwraith CW, Malemud CL: **Comparison of proteoglycan and collagen in articular cartlage of horses with naturally developing ostechondrosis and healing osteochondral fragments of experimentally induced fractures**. *American Journal of Veterinary Research* 2005, **66**(11):1881-1890.
235. de Grauw JC, Brama PA, Wiemer P, Brommer H, van de Lest CA, van Weeren PR: **Cartilage-derived biomarkers and lipid mediators of inflammation in horses with osteochondritis dissecans of the distal intermediate ridge of the tibia**. *American Journal of Veterinary Research* 2006, **67**(7):1156-1162.
236. Teyssèdre S, Dupuis MC, Elsen JM, Guerin G, Schiblert L, Denoix JM, Ricard A: **Genome-wide SNP association study identifies region of interest associated with osteochondrosis in French Trotters**. In: *9th World Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig, Germany*.
237. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS *et al*: **Genome-wide association analysis identifies 20 loci that influence adult height**. *Nature Genetics* 2008, **40**:575-583.
238. Jørgensen B, Arnbjerg J, Aaslyng M: **Pathological and radiological investigations on osteochondrosis in pigs, associated with leg weakness**. *Journal of Veterinary Medicine Series A* 1995, **42**(1-10):489-504.
239. McIlwraith CW: **Recent advances in diagnosis of equine joint disease**. In: *Proceedings of the 17th Kentucky Equine research Nutrition Conference: 26-27 April 2010; Lexington, KY*. 23-33.

240. Simm G: **Genetic improvement of cattle and sheep**. Tonbridge, UK: Farming Press; 2000.
241. Muir WM: **Comparison of genomic and traditional BLUP estimated breeding value accuracy and selection response under alternative trait and genomic parameters**. *Journal of Animal Breeding and Genetics* 2007, **124**:342-355.
242. Michel G, Massey JM: **Velogenetics or the synergistic use of Marker Assisted Selection and Germ-Line Manipulation**. *Theriogenology* 1991, **35**(1):151-156.
243. Habier D, Fernando R, Kizilkaya K, Garrick D: **Extension of the bayesian alphabet for genomic selection**. *BMC Bioinformatics* 2011, **12**(1):186.
244. Hayes BJ, Visscher PM, Goddard ME: **Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009)**. *Genetics Research* 2009, **91**(2):143-143.
245. Legarra A, Robert-Granie C, Manfredi E, Elsen JM: **Performance of genomic selection in mice**. *Genetics* 2008, **180**(1):611-618.
246. Powell JE, Visscher PM, Goddard ME: **Reconciling the analysis of IBD and IBS in complex trait studies**. *Nature Reviews Genetics* 2010, **11**(11):800-805.
247. Visscher PM, Yang J, Goddard ME: **A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010)**. *Twin Research and Human Genetics* 2010, **13**(6):517-524.
248. Meuwissen T: **Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping**. *Genetics Selection Evolution* 2009, **41**(1):35.
249. Haberland AM, König von Borstel U, Simianer H, König S: **Integration of genomic information into sport horse breeding programs for optimization of accuracy of selection**. *Animal* 2012, **6**(9):1369-1376.
250. Sitzenstock F, Konig S, Ytournel F, Simianer H: **Evaluation of genomic selection for functional traits in horse breeding programs**. In: *9th World Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig, Germany*.
251. Kizilkaya K, Fernando RL, Garrick DJ: **Genomic prediction of simulated multi-breed and purebred performance using observed 50k SNP genotypes**. *Journal of Animal Science* 2010, **88**:544-551.
252. Ibánẽz-Escriche N, Fernando RL, Toosi A, Dekkers JCM: **Genomic selection of purebreds for crossbred performance**. *Genetics Selection Evolution* 2009, **41**(1):12.
253. Goddard ME, Hayes BJ: **Genomic selection**. *Journal of Animal Breeding and Genetics* 2007, **124**:323-330.
254. Fox-Clipsham LY, Carter SD, Goodhead I, Hall N, Knottenbelt DC, May PDF, Ollier WE, Swinburne JE: **Identification of a mutation associated with Fatal Foal Immunodeficiency Syndrome in the Fell and Dales Pony**. *PLoS Genetics* 2011, **7**(7):e1002133.
255. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies**. *PLoS Genetics* 2009, **5**(6):e1000529.

256. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase**. *The American Journal of Human Genetics* 2006, **78**(4):629-644.
257. Li Y, Abecasis GR: **Mach 1.0: rapid haplotype reconstruction and missing genotype inference**. *American Journal of Human Genetics* 2006, **S79**:2290.
258. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nature Genetics* 2007, **39**(7):906-913.
259. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering**. *The American Journal of Human Genetics* 2007, **81**(5):1084-1097.
260. Pei Y-F, Li J, Zhang L, Papasian CJ, Deng H-W: **Analyses and comparison of accuracy of different genotype imputation methods**. *Plos One* 2008, **3**(10):e3551.
261. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A: **A comprehensive evaluation of SNP genotype imputation**. *Human Genetics* 2009, **125**(2):163-171.
262. Weigel KA, Van Tassell CP, O'Connell JR, VanRaden PM, Wiggans GR: **Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms**. *Journal of Dairy Science* 2010, **93**(5):2229-2238.
263. Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, van der Werf JHJ: **Accuracy of genotype imputation in sheep breeds**. *Animal Genetics* 2012, **43**(1):72-80.
264. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: **Efficiency and power in genetic association studies**. *Nature Genetics* 2005, **37**(11):1217-1223.
265. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium**. *The American Journal of Human Genetics* 2004, **74**(1):106-120.
266. Zhang K, Qin Z, Chen T, Liu JS, Waterman MS, Sun F: **HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms**. *Bioinformatics* 2005, **21**(1):131-134.
267. Halldórsson BV, Bafna V, Lippert R, Schwartz R, De La Vega FM, Clark AG, Istrail S: **Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies**. *Genome Research* 2004, **14**(8):1633-1640.
268. He J, Zelikovsky A: **MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression**. *Bioinformatics* 2006, **22**(20):2558-2561.
269. Halldórsson BV, Istrail S, De La Vega FM: **Optimal selection of SNP markers for disease association studies**. *Human Heredity* 2004, **58**(3-4):190-202.

270. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE: **Genomic selection using different marker types and densities**. *Journal of Animal Science* 2008, **86**(10):209-216.
271. Maniatis N, Collins A, Xu C-F, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, Morton NE: **The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis**. *Proceedings of the National Academy of Sciences* 2002, **99**(4):2228-2233.
272. Lewontin RC: **Interaction of selection and linkage. I. General considerations - Heterotic models**. *Genetics* 1964, **49**(1):49-67.
273. Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok P-Y, Collins A: **The optimal measure of allelic association**. *Proceedings of the National Academy of Sciences* 2001, **98**(9):5217-5221.
274. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals**. *The American Journal of Human Genetics* 2009, **84**(2):210-223.
275. Browning S: **Missing data imputation and haplotype phase inference for genome-wide association studies**. *Human Genetics* 2008, **124**(5):439-450.
276. Becker RA, Chambers JM, Wilks AR: **The New S Language**: Wadsworth & Brooks/Cole; 1988.
277. Cleveland WS: **Robust locally weighted regression and smoothing scatterplots**. *Journal of the American Statistical Association* 1979, **74**(368):829-836.
278. Cleveland WS: **Lowess - A program for smoothing scatterplotd by robust locally weighted regression**. *American Statistician* 1981, **35**(1):54-54.
279. Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K *et al*: **Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection**. *PLoS Biology* 2012, **10**(2):e1001258.
280. Dalrymple B, Kirkness E, Nefedov M, McWilliam S, Ratnakumar A, Barris W, Zhao S, Shetty J, Maddox J, O'Grady M *et al*: **Using comparative genomics to reorder the human genome sequence into a virtual sheep genome**. *Genome Biology* 2007, **8**(7):R152.
281. Weigel KA, de los Campos G, Vazquez AI, Rosa GJM, Gianola D, Van Tassell CP: **Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle**. *Journal of Dairy Science* 2010, **93**(11):5423-5435.
282. Ihara N, Takasuga A, Mizoshita K, Takeda H, Sugimoto M, Mizoguchi Y, Hirano T, Itoh T, Watanabe T, Reed KM *et al*: **A comprehensive genetic map of the cattle genome based on 3802 microsatellites**. *Genome Research* 2004, **14**(10A):1987-1998.
283. Hickey JM, Crossa J, Babu R, de los Campos G: **Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs**. *Crop Science* 2012, **52**(2):654-663.
284. Khatkar MS, Collins A, Cavanagh JAL, Hawken RJ, Hobbs M, Zenger KR, Barris W, McClintock AE, Thomson PC, Nicholas FW *et al*: **A first-generation metric linkage disequilibrium map of bovine chromosome 6**. *Genetics* 2006, **174**(1):79-85.

285. Zhang W, Collins A, Maniatis N, Tapper W, Morton NE: **Properties of linkage disequilibrium (LD) maps**. *Proceedings of the National Academy of Sciences* 2002, **99**(26):17004-17007.
286. de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF: **Practical aspects of imputation-driven meta-analysis of genome-wide association studies**. *Human Molecular Genetics* 2008, **17**(R2):R122-R128.
287. Servin B, Stephens M: **Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits**. *PLoS Genetics* 2007, **3**(7):e114.
288. Guan Y, Stephens M: **Practical Issues in Imputation-Based Association Mapping**. *PLoS Genetics* 2008, **4**(12):e1000279.
289. Lin DY, Hu Y, Huang B: **Simple and efficient analysis of disease association with missing genotype data**. *American Journal of Human Genetics* 2008, **82**(2):444-452.
290. Zaitlen N, Eskin E: **Imputation aware meta-analysis of genome-wide association studies**. *Genetic Epidemiology* 2010, **34**(6):537-542.
291. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: Models and data**. *The American Journal of Human Genetics* 2001, **69**(1):1-14.
292. Orr N, Back W, Gu J, Leegwater P, Govindarajan P, Conroy J, Ducro B, Van Arendonk JAM, MacHugh DE, Ennis S *et al*: **Genome-wide SNP association–based localization of a dwarfism gene in Friesian dwarf horses**. *Animal Genetics* 2010, **41**:2-7.
293. Ricard A, Danvy S, Legarra A: **First results on genomic selection in French show-jumping horses**. In: *33rd Conference of the International Society of Animal Genetics: 15-20 July, 2012; Cairns, Australia*. 2012.
294. Balding DJ: **A tutorial on statistical methods for population association studies**. *Nature Reviews Genetics* 2006, **7**(10):781-791.
295. So H-C, Sham P: **Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates**. *Behavior Genetics* 2011, **41**(5):768-775.
296. Cooley P, Clark R, Folsom R, Page G: **Genetic inheritance and genome wide association statistical test performance**. *Journal of Proteomics & Bioinformatics* 2010, **3**:330-334.
297. Li Q, Zheng G, Li Z, Yu K: **Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies**. *Annals of Human Genetics* 2008, **72**(3):397-406.
298. Hemani G, Theocharidis A, Wei W, Haley C: **EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards**. *Bioinformatics* 2011, **27**(11):1462-1465.
299. de los Campos G, Gianola D, Rosa GJM: **Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation**. *Journal of Animal Science* 2009, **87**(6):1883-1887.
300. Gianola D, Okut H, Weigel K, Rosa G: **Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat**. *BMC Genetics* 2011, **12**(1):87.

Web references

1. Ensembl (2012). *Equus caballus.* [Online] Available from: http://www.ensembl.org/Equus_caballus. [Accessed: 3 April 2012].
2. Merriam-Webster (2012). *Dictionary*. [Online] Available from: http://www.merriam-webster.com/dictionary. [Accessed: 25 June 2012].
3. National Center for Biotechnology Information (2012). *Equus caballus (Horse)* [Online]. Available from: http://www.ncbi.nlm.nih.gov/genome?term=equus%20caballus. [Accessed: 12 July 2012].
4. National Center for Biotechnology Information (2009). *Map View* [Online]. Available from: http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9796. [Accessed: 19 May 2009].
5. University of Kentucky (2012) *Horse Genome Project* [Online]. Available from: http://www.uky.edu/Ag/Horsemap/abthgp.html. [Accessed: 12 July 2012].

Appendix A: Supplementary documents

A.i Details of Meta-Analysis

In the meta-analysis there are *k* chromosomes and for each chromosome a parameter μ is estimated. The value of μ may vary between chromosomes so that the i^{th} chromosome true value is μ_i, and this is estimated as $\hat{\mu}_i$, with sampling error $\text{var}(\hat{\mu}_i - \mu_i) = s_i^2$. Let $\Delta^2 = \text{var}(\mu_i - \mu)$ be the variance between chromosome parameter estimates, where μ is now the mean value over chromosomes. Each estimate $\hat{\mu}_i$ is then an independent estimate of μ with sampling error $\Delta^2 + s_i^2$. The best estimate of μ is then $\hat{\mu} = \sum_i w_i^* \hat{\mu}_i$, where $w_i^* = (\Delta^2 + s_i^2)^{-1}$, with $s.e.(\hat{\mu}) = \left(\sum_i w_i^* \right)^{-1/2}$. This requires an estimate of Δ^2 to use in the weighting. Following DerSimonian and Laird (1986) [144], the statistic $Q_w = \sum_i w_i (\hat{\mu}_i - \bar{\mu}_w)^2$ is used, where $\bar{\mu}_w = \sum_i w_i \hat{\mu}_i \Big/ \sum_i w_i$ and $w_i = s_i^{-2}$ {note that $\bar{\mu}_w$ is also an estimate of μ but using sub-optimum weighting w_i, not w_i^*. Since $E(Q_w) = (k-1) + m\Delta^2$, where $m = \sum_i w_i - \sum_i w_i^2 \Big/ \sum_i w_i$, equating the observed Q_w with its expectation provides an estimate of Δ^2,

$$\hat{\Delta}_w^2 = \max\left\{0, \{Q_w - (k-1)\} \Big/ \left[\sum_i w_i - \left(\sum_i w_i^2 \Big/ \sum_i w_i \right) \right] \right\}$$

The maximisation step provides an estimate of Δ^2 with lower mean square error should $\hat{\Delta}_w^2 < 0$. This value is then used to calculate w_i^*, $\hat{\mu}$ and $s.e.(\hat{\mu})$ as described above.

A.ii Details of simulation validation

The expected number of segregating sites was calculated according to the formula $S_k = \theta m a_k$, where S_k is the expected number of segregating sites, $\theta = 4N_e u$, u is the mutation rate, m is the number of nucleotides in the sequence (assuming 1Mb≈1cM) and a_k is Watterson's correction factor for k sampled alleles ($2n$) [170]. In this case, a_k was calculated as $\ln(k) + \gamma$, where γ is Euler-Mascheroni's constant which represents the limiting difference between the harmonic series ($a_k = 1 + \frac{1}{2} + \frac{1}{3} + ...$) and the natural logarithm ($\ln(k)$) which has been suggested as providing a suitable approximation to a_k.

Table A-1 Comparison of observed and expected number of segregating sites

Sample size (*n*)	Expected no. of segregating sites (S_k)	Mean observed no. of segregating sites (across 30 replicates)	Percentage difference
50	4144	4150.8	0.164
100	4696	4718.2	0.471
200	5256	5324.8	1.301
400	5808	5990.1	3.087

The expected heterozygosity $E(H)$ of the sample was calculated as follows[10].

$E(H) = \int_{p_0}^{p_1} 2p(1-p)f(p)\,dp = \int_{p_0}^{p_1} 2k\,dp = (p_1 - p_0)2k$, where $f(p) = k/[p(1-p)]$ and *k* is determined by: $\int_{p_0}^{p_1} f(p)\,dp = 1 = \int_{p_0}^{p_1} k\left[p^{-1} + (1-p)^{-1}\right]dp = 2k \ln\left(p_1/p_0\right)$, so $k = \left[2\ln\left(p_1/p_0\right)\right]^{-1}$ and $E(H) = (p_1 - p_0)/\ln\left(p_1/p_0\right)$. If N_e=200, then the minimum frequency is 1/400 and maximum frequency is 399/400, and the best continuous approximation has 399 intervals going from p_0 = 1/800 to p_1 = 799/800 so $E(H) = (798/800)/\ln(799) = 0.149$.

Table A-2 Observed heterozygosity

Sample size (*n*)	Mean heterozygosity of segregating sites[I]	Mean heterozygosity weighted by total no. of sites
50	0.1903	2,369,691.7
100	0.1679	4,753,081.1
200	0.1491	9,527,191.8
400	0.1327	19,077,270.5
	Overall Mean:	0.1438

[I] **Calculated assuming HWE (*H* = 2*pq*)**

[10] Derivation by John Woolliams (The Roslin Institute).

A.iii Description of contemporary groups

Table A-3 The distribution of samples across contemporary groups

		Presence (1) or absence (0)				No. of Samples		
No.	**Class**	**ALD[I]**	**Fetlock chip(s)**	**Sesamoid fracture(s)**	**Sex (f=female, m=male)**	**Controls**	**Cases**	**Total**
1	Af	0	0	0	F	3	40	43
2	Am	0	0	0	M	4	57	61
3	Bf	1	0	0	F	52	12	64
4	Bm	1	0	0	M	21	13	34
5	Cf	0	1	0	F	28	10	38
6	Cm	0	1	0	M	28	15	43
7	Df	0	0	1	F	8	4	12
8	Dm	0	0	1	M	13	1	14
9	Ef	1	1	0	F	7	6	13
10	Em	1	1	0	M	3	3	6
11	Ff	1	0	1	F	1	1	2
12	Fm	1	0	1	M	0	0	0
13	Gf	0	1	1	F	0	0	0
14	Gm	0	1	1	M	0	0	0
15	Hf	1	1	1	F	0	0	0
16	Hm	1	1	1	M	0	0	0
					TOTAL:	**168**	**162**	**330**

[I]ALD = angular limb deformity

A.iv Quantitative trait loci

Table A-4 Details of QTL regions from Lampe (2009) [99] and Komm (2010) [100] tested

QTL no.	Chr	From (bp)	To (bp)	QTL type[I]	Reference
1	1	43,130,411	44,675,827	C	Lampe (2009), ch. 8
2	2	17,546,278		A	Komm (2010), ch. 5
3	3	64,213,851	64,632,362	C	Lampe (2009), ch. 8
4	3	69,400,000	71,400,000	B	Komm (2010), ch. 5
5	4	7,614,842		A	Komm (2010), ch. 5
6	4	13,101,957[II]		A[†]	Komm (2010), ch. 4
7	4	15,993,651		A[†]	Komm (2010), ch. 4
8	4	38,260,000	40,260,000	B	Komm (2010), ch. 5
9	4	41,202,168	41,365,096	C	Lampe (2009), ch. 8
10	5	57,960,000	59,960,000	B	Komm (2010), ch. 5
11	7	16,020,000	18,020,000	B	Komm (2010), ch. 5
12	14	57,700,667	57,826,754	C	Lampe (2009), ch. 8
13	16	39,044,608		A	Komm (2010), ch. 5
14	16	79,400,000	81,400,000	B	Komm (2010), ch. 5
15	16	82,398,594	82,803,587	C	Lampe (2009), ch. 8
16	18	36,408,881	38,738,316	C	Lampe (2009), ch. 8
17	18	79,364,656	80,836,383	D	Lampe (2009), ch. 7
18	19	620,000	2,620,000	B	Komm (2010), ch. 5
19	20	13,220,000	15,220,000	B	Komm (2010), ch. 5
20	22	2,790,000	4,790,000	B	Komm (2010), ch. 5

QTL no.	Chr	From (bp)	To (bp)	QTL type[I]	Reference
21	26	26,610,000	28,610,000	B	Komm (2010), ch. 5
22	29	7,670,000	9,670,000	B	Komm (2010), ch. 5
23	30	12,182,691	12,239,797	C	Lampe (2009), ch. 8
24	X	55,573,885	55,574,350	C	Lampe (2009), ch.8

[I]A – a specific SNP locus which lies within a QTL previously identified in linkage analyses and was subsequently found to be associated with OCD and/or OC in a GWAS; A[†] - a specific SNP locus found to be associated with OC or OCD in a QTL refinement study; B – a region which represents a 2Mb window surrounding the approximate location of a significant QTL identified in a GWAS; C – a region which contained ≥3 significant SNP(s) in a GWAS; D – a region which contained a significant SNP haplotype found within a QTL previously identified in linkage analyses.

[II]BIEC2-893170 at position 13,101,957 was not typed in my sample. Therefore BIEC2-849331 at position 13,126,529 was tested in the mixed model.

Appendix B: Equine biobank documents

B.i Project brief

Background

During my PhD, there have been several discussions between myself, my supervisors (Prof. John Woolliams of RI and Sarah Blott of AHT) and representatives of the BEF (Jan Rogers, Graham Suggett and Jenny Hall) about the lack of sample availability for future genomic work in the horse. There is potential for genomic technologies to be used to increase the rate of genetic improvement in the UK sport horse population both with respect to disease reduction and performance improvement. A significant barrier to the implementation of such marker assisted breeding programs is a lack of DNA samples with associated data for use in research studies. Since the development of the Illumina Equine SNP50 BeadChip, groups from across the world have embarked on projects to elucidate the genetic component of various traits from coat colour to racing speed. Studies published to date nearly all share one thing in common – a lack of power to detect all but the largest genetic effects due to small sample size. Therefore, if the UK equine industry is to make use of the current and future genomic technologies, a strategy is needed to collect well-characterised DNA samples. This project is concerned with whether a national equine biobank could be part of such a strategy.

Scope, Objectives and Deliverables

This project has two main components:

1. An overview of biobank methodology. Objectives:
    - Produce a report that: describes the principal considerations in setting up a biobank and discusses the advantages and disadvantages of different biobank methodologies.
2. A survey of opinion within the equine industry. Objectives:
    - Undertake interviews with sample cross-section of industry representatives

- Document the opinions of a range of industry representatives on the establishment of a national biobank
- Document the opinions of the general horse owning public on the establishment of a national biobank

Further objectives are then:

- To consider different biobank methodologies in the context of the equine industry.
- To produce a list of stakeholders within the equine industry and their particular interests/needs with respect to a national equine biobank.
- To provide guidance to the BEF on the next steps with respect to the establishment of a national equine biobank.
- To produce a list of experts that could offer advice and/or assistance to the BEF on the topic of an equine biobank in the future.

This project **will not**:

- Provide a full stakeholder analysis
- Provide a costing for the establishment of a national equine biobank
- Provide a plan for the execution of a national equine biobank

Funding

Reasonable expenses will be provided by the BEF to cover costs including: travel to conduct interviews, printing costs.

Success Criteria

This project will be considered a success if:

- A report is compiled that improves the BEF's knowledge regarding the feasibility of setting up a national equine biobank, enabling them to make a decision regarding the future progression of this idea.
- A list of experts in the field is provided to the BEF.

Work/Task Breakdown Structure

1. An overview of biobank methodology:
    - Literature review of biobank methodology including different workflows, sample collection options, etc.
    - Investigation of biobank services offered by companies, e.g. processing and storage of samples, software packages to manage biobank collections
    - A superficial look at the cost of establishing a biobank
2. A survey of opinion within the equine industry:
    - Online survey open to all horse owners and riders
    - Interviews with industry representatives both to gather opinions and to compile a list of experts in the field that could be consulted in future
3. Presentation of results:
    - Written report in the style of a government white paper discussing biobank methodology in the context of the horse industry
    - Short report with key conclusions and recommendations

Milestones

	Date	Have completed by end	Time spent on project this sector
Start Date:	4th January 2012		
	27th January 2012	List of potential interviewees compiled	20%
	2nd March 2012	Contact made with majority of intended interviewees and plan of interviews developed	20%
	30th April 2012	Relevant information about biobanking methodology collected, interviews completed, online survey data collated	80%
	31st July 2012	All data analysed and abstract for ESBB written	35%
Planned Completion Date:	31st August 2012	Final report submitted	100%

Benefits

By the end of the project, it should be clearer whether a national equine biobank is a feasible solution to the current problem of a lack of well-characterised DNA samples for use in genomic studies of disease and performance in the horse.

Resources, Skills and Costs

The majority of the work will be undertaken by me (Laura Corbin), with the proportion of my time being spent on the project as indicated in the milestones section above. The main expenditure will be travelling costs.

Project Impacts and Dependencies

Whilst carrying out this project the remainder of my time (as indicated under milestones) will be spent writing up my thesis. Timings may need to be adjusted to ensure both this project and my thesis are completed on time.

B.ii Online survey questions

1. **Select one or more of the following to best describe your involvement within the horse industry.**

- I am a hobby/amateur rider
- I am a professional rider
- I am a breeder
- I own and/or manage a commercial yard
- I am a groom and/or instructor
- I am studying/have studied a further education or higher education course in horses
- I am an equine professional/paraprofessional (please use space below to specify, e.g. veterinarian,dentist)
- Other, please specify

2. **Do you compete in any of the following disciplines? Please select all that apply.**

- No, I don't compete
- Show Jumping
- Dressage
- Eventing
- Para Dressage
- Endurance
- Showing
- Vaulting
- Reining
- Driving
- Racing
- Polo
- Other, please specify

3. **Are you a member of any of the following equine organisations? Please select all that apply.**

- No, I am not a member of any equine societies/organisations
- British Horse Society
- British Show Jumping
- British Dressage
- British Eventing
- Endurance GB
- British Equestrian Vaulting
- British Reining
- British Horse Driving Trials Association
- Thoroughbred Breeders Association
- Other, please specify

4. **How many horses do you currently own?**

- None
- 1
- 2-4
- More than 5

5. **Is/are your horse/s registered with a breed society or studbook, either in the UK or overseas?**

- No
- Yes
- If yes, please specify

6. **Has/have your horse/s ever been tested for a genetic disease?**

- No
- Don't know
- Yes
- If yes, please provide further details below

7. **Biobanks have traditionally been associated with biomedical research. They are collections of biological materials (such as blood and/or tissues) and personal data (medical records, lifestyle data) from large numbers of people. Using such biobanks, researchers can identify the genetic and environmental factors associated with the risk of certain diseases. This information can help to improve prevention, diagnosis and treatment. Critics, however, raise questions about the privacy and confidentiality of biobanks and have concerns over commercial interests and regulation. As scientists unravel the genetic code of many animals, including the horse, veterinary researchers are creating collections of animal biological samples with associated data for use in similar genetic studies. Data in this setting may include diagnostic test results, treatment received, etc.**

Suppose you were asked to vote for or against the creation of an equine biobank in the UK, what issues would you like to know more about before voting? Rank the following from the issue you would be most interested in knowing more about (1) to the issue you would be least interested in knowing more about (5).

Rank the following items using numbers from 1 to 5.

- What the potential benefits would be What the possible risks would be
- Who would fund the biobank
- How the biobank would be managed and by
- whom
- How the biobank would be regulated and by whom

8. **Before finding out about this survey, had you ever heard anything about biobanks?**

- Yes
- No

9. **Other than for the purposes of this survey, have you ever....?**

- Talked about biobanks with anyone
- Searched for information about biobanks
- Talked specifically about an equine biobank with anyone
- Searched for information specifically about equine biobanks

10. **Below is a list of areas in which an equine biobank could be used. Indicate how important you think each of them are.**

- Conservation
- Genetic research of disease
- Genetic research of performance traits
- Breeding – to reduce disease by using marker assisted selection
- Breeding – to improve performance by using marker assisted selection

11. **In order to understand the causes of disease, researchers need as much information as possible about the horses in the biobank. Would you be concerned or reluctant about the collection of any of the following types of data about and materials from your horse/s?**

- Blood samples (left over from essential veterinary procedures)
- Blood samples (taken solely for the purpose of the biobank)
- Surplus tissue collected during essential veterinary procedures
- Hair (with roots attached)
- Buccal (cheek) swab
- Nasal swab
- Veterinary records
- Performance records, e.g. British Showjumping points
- Lifestyle information, e.g. time spent at grass, feeding regime
- Pedigree

12. **Assuming that you had agreed for a sample to be collected from your horse/s, where and by whom should the sample to be taken? Indicate whether you would participate in the following collection schemes.**

- Collection at a competition or other event that you were already taking your horse/s to.
- Collection at a nearby venue which you would take your horse/s to, specifically for a sample to be taken.
- Collection by yourself (only applies to hair, buccal (cheek) swabs and nasal swabs).
- Collection by your vet during a routine visit.
- Collection by your vet following the admission of your horse/s for treatment.

13. **The more information associated with a sample in the biobank, the more valuable that sample will be to researchers. Therefore, the ability to obtain follow-up data would be an important aspect of a biobank. For each of the following statements regarding methods of obtaining follow-up information relevant to your horse's sample, indicate whether you agree or disagree.**

- I would be willing to provide regular updates (e.g. every six months) about my horse/s via an online profile
- I would be willing to provide updates about my horse/s when asked to do so
- I would be happy for my horse's performance records to be accessed by researchers, as required, without my permission
- I would be happy for my horse's veterinary records to be accessed by researchers, as required, without my permission
- I would be happy for my horse's performance records to be accessed by researchers, as required, provided that my permission was sought first
- I would be happy for my horse's veterinary records to be accessed by researchers, as required, provided that my permission was sought first
- I would prefer for my horse's sample and associated data to be made anonymous following collection and therefore would not be willing for any further data to be collected

14. **Biobanks are expensive to maintain. Therefore, owners may be asked to pay a small fee to have their horse's sample and data collected. For each of the following statements regarding payment for biobank services, indicate whether you agree or disagree.**

- I would be happy to pay a small fee
- I would be happy to pay a small fee but I would expect something in return
- I do not think horse owners should have to pay, the people who use the samples and information should pay
- I do not think horse owners should have to pay, equine organisations should pay
- I do not think horse owners should have to pay, the government should pay
- I do not think horse owners should have to pay, but I do not know who else should pay

15. **If a UK equine biobank was created, how concerned would you be about the following?**
- Confidentiality of biological samples
- Confidentiality of data
- Potential effects of genetic test results on insurance premiums
- Potential effects of genetic test results on sale values
- Potential effects of genetic test results on breeding values

16. **Do you think that the UK should have an equine biobank? Please use the box below if you have any additional comments.**
- Yes, definitely
- Yes, probably
- No, probably not
- No, definitely not
- Don't know
- Additional Comments

17. **If a UK equine biobank were created, would you be willing to provide samples from and information about your horse/s? Please use the box below if you have any additional comments.**
- Yes, definitely
- Yes, probably
- No, probably not
- No, never
- Don't know
- Additional Comments

18. **Why would/might you not be willing to provide samples from and information about you horse/s to a biobank?**

19. **Please use the space below to add any further comments that you feel are relevant to the topic of an equine biobank.**

20. **Please indicate below how much of the online supplementary information you read before completing this questionnaire.**
- Survey Welcome Greeting (on the BEF website)
- Project Information (on The Roslin Institute website)
- What is a biobank? (on The Roslin Institute website)
- What could an equine biobank be used for? (on The Roslin Institute website)

B.iii Online survey results