Genome-Wide Evaluation of Populations

Hans D. Daetwyler

Genome-Wide Evaluation of Populations

Hans Dieter Daetwyler

Thesis Committee

Thesis supervisors
Prof. Dr. Ir. Johan A.M. van Arendonk
Professor of Animal Breeding and Genomics
Wageningen University

Prof. Dr. John A. Woolliams
Senior Principle Investigator
The Roslin Institute and Royal (Dick) School of Veterinary Studies
University of Edinburgh, Roslin, UK

Thesis co-supervisor
Dr. Beatriz Villanueva
Researcher
Departamento de Mejora Genética Animal
INIA, Madrid, Spain

Other members

Prof. Dr. Fred A. van Eeuwijk	Wageningen University
Prof. Dr. Michel Georges	University of Liège, Belgium
Prof. Dr. Theo H.E. Meuwissen	Norwegian University of Life Sciences, Ås, Norway
Prof. Dr. Miguel Pérez-Enciso	Universitat Autonoma de Barcelona, Spain

This thesis was conducted under the auspices of the Wageningen Institute of Animal Sciences.

Genome-Wide Evaluation of Populations

Hans Dieter Daetwyler

Thesis
Submitted in partial fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr. M.J. Kropff
in the presence of the
Thesis Committee appointed by the Doctorial Board
to be defended in public
on Wednesday 9 December 2009
at 11 AM in the Aula.

Hans. D. Daetwyler
Genome-Wide Evaluation of Populations
192 pages.

Ph.D. Thesis: Animal Breeding and Genomics Centre, Wageningen University,
Wageningen, NL (2009)

With references and including summaries in English and Dutch.

ISBN: 978-90-8585-528-6

Contents

Abstract

Genome-Wide Evaluation of Populations

A large amount of genetic marker data is now available in many species. This thesis investigated the use of this type of data to estimate genetic values in both animal and human populations. Two different general approaches are followed. The first approach aimed at detecting quantitative trait loci (QTL) associated with phenotypic traits in dairy cattle. A large number of QTL affecting different traits were detected using both a linkage analysis variance component method, which found 102 potential QTL, and a linkage disequilibrium regression method, which resulted in 144 SNP associations. The remainder of this thesis investigated a second general approach, called genome-wide evaluation (GWE), which allows us to estimate the effects of all QTL affecting a particular trait simultaneously and to predict breeding values by summing the effects of all loci. Deterministic predictions of the accuracy of a GWE least squares approach were derived and tested using simulated data. The factors that affect this accuracy and their relationships were clearly defined. Although many GWE methods exist, they can be broadly grouped into variable and non-variable selection methods. Variable selection methods, such as BayesB, attempt to identify a subset of SNP from which to estimate breeding value, whereas non-variable selection methods, such as genomic best linear unbiased prediction (GBLUP), assume all SNP have an effect. Two GWE methods representative of both groups, BayesB and GBLUP, were compared for varying effective population sizes and numbers of QTL affecting the trait. Population and trait genetic architecture were found to have a large influence on the relative performance of methods. The variable selection method was only found to be advantageous when the number of QTL was less than the number of independent chromosome segments. In addition, deterministic formulae derived for the least squares approach were extended to be predictive of the accuracies of both BayesB and GBLUP. Predictions for GBLUP accuracy were tested using real dairy cattle data and were found to be generally accurate. In addition, the reasons why selection on genomic breeding values is expected to result in lower inbreeding rates per generation than traditional genetic evaluation methods, when compared at the same rate of genetic gain, are summarised. Furthermore, a chromosomal phasing algorithm was developed to phase and impute missing genotypes in complex pedigrees. The algorithm was tested in varying depths of pedigree in simulated data and was able to impute a high percentage of genotypes correctly. Imputation of missing genotypes could be used to increase sample sizes for GWE. In addition, a method was developed to estimate the proportion of genetic variation tagged by a particular SNP chip and this method was used to estimate the proportion of variation tagged by the chip currently in use in dairy cattle. Finally, the expected impact of sequence data on GWE and issues related to implementation of GWE were discussed.

Chapter 1

General Introduction

Hans D. Daetwyler[1,2]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin, Midlothian, EH25 9PS, UK; [2]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6700 AH, The Netherlands

Traditional genetic evaluation methods, such as best linear unbiased prediction, use phenotypic data and pedigrees to estimate breeding values. Additional data sources, such as molecular information, can be incorporated into these approaches with the aim of increasing genetic gain. Development of methods to include blood group data (Neimann-Sorensen & Robertson 1961) or DNA marker data (Smith 1967) in animal breeding actually preceded the discovery of genetic markers. Once genetic markers, such as restriction fragment length polymorphisms, microsatellites and single nucleotide polymorphisms (SNP) were found, further development of statistical methods to detect quantitative trait loci (QTL) followed (e.g. Haley & Knott 1992; Georges *et al.* 1995; George *et al.* 2000; Meuwissen & Goddard 2000; Knott & Haley 2000; Meuwissen *et al.* 2002; Grapes *et al.* 2004).

The initial expectation that QTL will be widely used in animal breeding programs has not fully materialised because of a number of implementation challenges. Early marker maps were very sparse and therefore QTL could not be fine-mapped. Therefore, using them in marker-assisted breeding schemes required the determination of the phase between markers and QTL. Secondly, effects of significant QTL tended to be overestimated (e.g. Beavis 1998; Goring *et al.* 2001). This meant that effects needed to be re-estimated in independent population samples before using QTL information in breeding schemes. Studies also turned up false positives and validation studies were initiated to attempt to identify true QTL. Unfortunately many QTL could not be consistently validated in independent samples, likely as a result of studies being underpowered. The incorporation of QTL into breeding programs has been further complicated by the fact that the genetic variance accounted for by a QTL must be properly weighted with the polygenic variance. Furthermore, genotyping was expensive and collectively these issues have mainly prohibited widespread implementation of marker-assisted selection. Nevertheless, some causative mutations have been identified (e.g. Grisart *et al.* 2004; Cohen-Zinder *et al.* 2005; Clop *et al.* 2006) and there are breeding schemes which have successfully applied marker-assisted selection in quantitative traits (e.g. Guillaume *et al.* 2008).

Sequencing of various livestock genomes has accelerated the development of high throughput genotyping and SNP discovery. These developments have given rise to chips with thousands of SNP at reducing cost. The density of these SNP chips is such that markers can now be expected to be in population-wide linkage disequilibrium (LD) with QTL, which allows for easier use of marker data.

The availability of dense marker maps has opened new opportunities for genetic evaluation of individuals with high accuracy. Meuwissen et al. (2001) proposed a genetic evaluation method called genome-wide evaluation (GWE) and its wide-spread use depended on dense marker maps being available. Similarly, wide-spread application of genetic markers in animal breeding also seems to depend on effective and, preferably, simple methodology being available. Genome-wide evaluation fulfils that role and its beauty is its simplicity. Rather than concentrating on finding particular QTL of moderate to large effect, GWE estimates effects for all markers or haplotypes in the genome. These effects are then summed to a breeding value for an individual. It is important that each QTL is in significant LD with at least one marker, so that all of the genetic variance is accounted for. Most GWE methods estimate marker effects simultaneously and significance thresholds are generally not applied. In addition, marker effects are regressed towards the mean depending on a variance parameter which can differ across methods. These features reduce the overestimation of effects. The application of marker information pre-GWE required at least three steps, such as detecting the QTL, confirming them and re-estimating effects, and finally incorporating them into animal breeding programs to estimate breeding values. In contrast, GWE uses one step in which marker effects and breeding values are estimated.

There are a number of advantages of GWE over classic genetic evaluation methods. Assuming that enough phenotypes and genotypes are available for estimation of marker effects, breeding value accuracies from GWE can be substantially higher than those obtained with traditional approaches (Chapter 3 this thesis; Goddard 2008). The main reason for this increase in accuracy is that with molecular information the Mendelian sampling term can be better quantified, though the accuracy of genomic breeding values is also, in part, due to relationships among individuals in the sample (Habier *et al.* 2007). The reliance of GWE on the Mendelian sampling term is also the reason why lower inbreeding rates per generation are expected when compared to best linear unbiased prediction (Chapter 6 this thesis; Woolliams *et al.* 2002). In GWE, marker effects are estimated in one population sample of individuals with phenotypes and genotypes. Once effects have been estimated, breeding values can be predicted for animals only genotyped from the same population with only genotypes. Being able to calculate breeding values for juveniles without phenotypes or progeny can potentially shorten generation intervals substantially, leading to even greater annual genetic gains. Furthermore, GWE can potentially be applied in un-pedigreed populations as, in principle, a pedigree is not required. On the other hand, genotyping and, in some cases, phenotyping can be costly especially if a large number of genotypes and phenotypes is required for estimating effects, as it is the case for traits of low heritability. Thus, it may be difficult to justify GWE in populations with narrow profit margins. However, genotyping costs are expected to continue to decline in the future and wide-spread application of GWE may be possible. The large potential impact on genetic gain of GWE may also lead to substantial changes in the design of breeding programs.

Genome-wide evaluation research is, in many ways, still in its infancy. Several studies have confirmed the potential of GWE and each study reveals another piece of the puzzle. This thesis adds more pieces to this puzzle.

Outline of Thesis

The overall aim of this thesis was to investigate the use of genomic marker data in genetic evaluation of populations. The thesis deals with both QTL detection and GWE methods. However, the focus is on GWE after Chapter 2 and breeding value accuracy is the unifying theme among these remaining chapters.

Chapter 2 is a QTL detection study in production and functional traits of Holstein cattle using both a linkage analysis variance component method and an association approach. A 10K Bovine SNP chip is used, and numerous potential QTL and significant SNP associations are detected.

Chapter 3 derives deterministic equations for the prediction of accuracy for continuous and for dichotomous traits both in population and case control studies using a least squares approach. The equations are extensively tested using stochastic simulation and the factors affecting GWE accuracy are identified.

Chapter 4 investigates the impact of population genomic structure and trait genetic architecture on GWE methods. A variable selection (BayesB) and a non-variable selection method (GBLUP) are compared at three different effective population sizes and a wide range of QTL affecting the trait. Furthermore, deterministic equations of Chapter 3 are extended to predict the accuracy of these two methods.

Chapter 5 compares predictions from Chapter 3 to accuracies achieved in real Holstein and Jersey dairy cattle populations both in the USA and Australia and also to predictions from a formula proposed by Goddard (2008). Both deterministic predictions match real accuracies generally well, though equations may need to be extended to account for the proportion of the genetic variance captured by a SNP chip, when predicting the accuracy in a denser SNP chip.

Chapter 6 uses theoretical concepts on inbreeding established from traditional pedigree-based methods to extrapolate what inbreeding rates are expected from selection using GWE. It concludes that selecting on genomic breeding values will result in lower rates of inbreeding per generation when compared to traditional best linear unbiased prediction at the same rate of genetic gain.

Chapter 7 describes a chromosomal phasing algorithm computationally efficient for imputing missing genotypes. The approach is tested in simulated data with varying number of generations and with three different proportions of loci missing. The performance of the algorithm is very good when more than two generations of individuals are available.

Chapter 8 is the General Discussion which raises four main topics. First, a method is presented to estimate the proportion of the total genetic variation tagged by current SNP chip, and this proportion is estimated for the 50K Illumina Bovine chip using US Holstein data. Secondly, GWE methods are discussed in terms of their performance according to different trait and population genetic architectures. The third topic is the impact that sequence data is likely to have on GWE. Finally issues related to implementation are discussed.

REFERENCES

Beavis, W. D., 1998 QTL analyses: power, precision, and accuracy, pp. 145-162 in *Molecular Dissection of Complex Traits*, edited by A. C. Paterson. CRC Press, New York.

Clop, A., F. Marcq, H. Takeda, D. Pirottin, X. Tordoir *et al.* 2006 A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genetics* **38**: 813-818.

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, W. A. Everts-van der *et al.* 2005 Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res.* **15**: 936-944.

George, A. W., P. M. Visscher, and C. S. Haley, 2000 Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**: 2081-2092.

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto *et al.* 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907-920.

Goddard, M. E., 2008 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**: 245-252.

Goring, H. H. H., J. D. Terwilliger, and J. Blangero, 2001 Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Gen.* **69**: 1357-1369.

Grapes, L., J. C. Dekkers, M. F. Rothschild, and R. L. Fernando, 2004 Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics* **166**: 1561-1570.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim *et al.* 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc.Natl.Acad.Sci.U.S.A* **101**: 2398-2403.

Guillaume, F., S. Fritz, D. Boichard, and T. Druet, 2008 Short communication: Correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. *J. Dairy Sci.* **91**: 2520-2522.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**: 2389-2397.

Haley, C. S., and S. A. Knott, 1992 A Simple Regression Method for Mapping Quantitative Trait Loci in Line Crosses Using Flanking Markers. *Heredity* **69**: 315-324.

Knott, S. A., and C. S. Haley, 2000 Multitrait least squares for quantitative trait loci detection. *Genetics* **156**: 899-911.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**: 421-430.

Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**: 373-379.

Neimann-Sorensen, A., and A. Robertson, 1961 The association between blood groups and several production charactersitics in three Danish cattle breeds. *Acta Agric.Scand.* **11**: 163-196.

Smith, C., 1967 Improvement of metric traits through specific genetic loci. *Anim.Prod.* **9**: 349-358.

Woolliams J.A., Pong-Wong R. & Villanueva B. Strategic optimisation of short- and long-term gain and inbreeding in MAS and non-MAS schemes. 7th World congress of genetics applied to livestock production. Proc.7th WCGALP. 19-8-2002.

Chapter 2

A Genome Scan to Detect Quantitative Trait Loci for Economically Important Traits in Holstein Cattle Using Two Methods and a Dense Single Nucleotide Polymorphism Map

Hans D. Daetwyler, Flavio S. Schenkel, Mehdi Sargolzaei, and J. Andrew B. Robinson

Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science,
University of Guelph, Guelph, ON, Canada

Published in Journal of Dairy Science, 2008. 91:3225-3236.

ABSTRACT

Quantitative trait loci (QTL) detection bovine genome scans were performed via variance component linkage analysis (VCLA) and linkage disequilibrium single locus regression (LDRM). Four hundred and eighty four Holstein sires, of which 427 were from 10 grandsire families, were genotyped for 9,919 single nucleotide polymorphisms (SNP) using the Affymetrix MegAllele GeneChip Bovine Mapping 10K SNP array. A hybrid of the granddaughter and selective genotyping designs was applied. Four thousand eight hundred fifty six of the 9,919 SNP were located to chromosomes in base-pairs and formed the basis for the analyses. The mean polymorphism information content of the SNP was 0.25. The SNP cM position was interpolated from their base-pair position using a microsatellite framework map. Estimated breeding values were used as observations and the following traits were analyzed: 305-day lactation milk, fat and protein yield, somatic cell score, herd life, interval of calving to first service, and age at first service. Both approaches were effective in detecting potential QTL with a dense SNP map. The VCLA analysis detected 102 potential QTL, while LDRM analysis found 144 significant SNP associations after accounting for a 5% false discovery rate. Twenty potential QTL and 49 significant SNP associations were in close proximity to QTL cited in the literature. Both methods found significant regions on *Bos taurus* autosome (BTA) 3, 5, and 16 for milk yield, BTA 14 and 19 for fat yield, BTA 1, 3, 16 and 28 for protein yield, BTA 2 and 13 for calving to first service, and BTA 14 for age at first service. LDRM was well suited for a first genome scan due to its approximately eight times lower computational demands. Further fine mapping should be applied on the chromosomal regions of interest found in this study.

INTRODUCTION

Traditional methods of genetic improvement in livestock species have relied solely on phenotype and pedigree information. The discovery of genetic markers has made it possible to detect regions of the genome that are significantly associated with differences in the expression of a phenotype such as milk production, so called quantitative trait loci (QTL). Genetic response can be improved by including the QTL in marker assisted selection (MAS), which is a method of selection that makes use of phenotypic, genotypic and pedigree data (Smith 1967). In MAS, selection does not occur on the QTL directly, unless the genetic marker is the causal mutation, but on the marker that is linked to the QTL through linkage disequilibrium (LD).

In the past, genotyping many markers was expensive and therefore specific experimental designs were developed to reduce the impact of having fewer markers on statistical power. The granddaughter design in dairy cattle made use of the high sire estimated breeding value (EBV) accuracies due to progeny tests to maximize power while lowering the number of genotyped animals (Weller *et al.* 1990). However, more recently, high throughput methods have been developed to genotype markers such as single nucleotide polymorphisms (SNP) which have significantly reduced the cost. It is currently possible to genotype individuals for 10,000, 50,000 or more SNP with a GeneChip array and the bovine genome can be covered with a dense SNP map to potentially increase the power of association studies.

The QTL detection studies performed to date have found a large number of QTL in dairy cattle for traits of medium to high heritability, such as milk yield and composition traits (e.g. Khatkar *et al.* 2004; Polineni *et al.* 2006). Information on QTL that are associated with conformation and functional traits is becoming more readily available (e.g. Schrooten *et al.* 2000; Ashwell *et al.* 2005) and traits of lower heritability, such as fertility traits, have been successfully mapped for QTL (e.g. Boichard *et al.* 2003; Kuhn *et al.* 2003).

The advent of high throughput genotyping technology gives hope to finding more QTL for functional and fertility traits where heritability is usually low. The aim of this study was to use the higher power gained from a dense SNP map and perform scans of the *Bos taurus* genome to detect potential QTL in traits of medium to low heritability via variance component linkage analysis (VCLA) (George *et al.* 2000) and linkage disequilibrium single locus regression (LDRM) (Grapes *et al.* 2004).

MATERIALS AND METHODS

Experimental Design. The experimental design was a hybrid of the granddaughter (Weller *et al.* 1990) and the selective genotyping (Darvasi & Soller 1992) designs. Ten Holstein grandsires with sufficiently large groups of progeny tested sons in Canada were chosen. From these ten families, the lowest and highest four to five sons were chosen according to their EBV for each of the following four traits (305-day lactation protein yield, mammary system, somatic cell score and daughter fertility). Some bulls overlapped across traits, so that the number of bulls resulting from this process was 333, for a mean of 33 sons per grandsire. In addition, 88 grandsons from six of the 10 grandsires, 46 potential sires of sons, and 17 Holstein bulls imported from Europe were added. In total 484 bulls were genotyped and 421 of the bulls were part of the 10 core families. Up to 6 generations of genotyped sires were represented in the dataset and the mean inbreeding coefficient of all genotyped bulls was 5.9%. All the 484 sampled bulls contributed genetically to the current Canadian Holstein cow population.

The dataset was checked for stratification between the European and North American bulls by tracing back the pedigrees and by calculating allele frequency correlations between the 17 European and a random sample of 17 North American bulls in the dataset for all the SNP that showed significant associations with any of the traits analyzed (5000 replicates). The expected correlation of allele frequencies within only the North American bulls was also calculated from 5000 random samples of two groups of 17 North American bulls.

Observations. The observations used were EBVs obtained from the Canadian Dairy Network, Guelph, Ontario from the May 2006 genetic evaluation (Van Doormaal 2007). Multiple Across Country Evaluation was used, if needed and available, according to the minimum criteria for official bull proofs of the (Canadian Dairy Network 2007). The EBV statistics can be found in Table 1 and show that the mean EBV accuracy was high (range 88.3% to 94.9%). The following traits were analyzed: 305-day lactation milk yield (MY), 305-day lactation protein yield (PY), 305-day lactation fat yield (FY), somatic cell score (SCS), herd life (HL), interval from calving to first service (CTFS), and age at first service (AFS). Herd life is a measure of longevity expressed as the number of lactations a cow stays in the herd. Somatic cell score refers to the amount of somatic cells a cow has in her milk and is an important indicator trait for mastitis. Calving to first service is the period from parturition to first insemination in days and AFS is the age in days at which a heifer is artificially inseminated for the first time.

Table 1. Mean estimated breeding value (EBV $\bar{x}$), EBV standard deviation (EBV SD), Mean EBV accuracy (EBV acc. $\bar{x}$), and percentage of bulls with an EBV accuracy equal or greater than 90.0 (% acc. ≥ 90.0) for milk yield (MY), fat yield (FY), protein yield (PY), somatic cell score (SCS), herd life (HL), calving to first service (CTFS), and age at first service (AFS)

Trait	EBV $\bar{x}$	EBV SD	EBV acc. $\bar{x}$	% acc. ≥ 90.0
MY	631	773.9	94.9	91.8
FY	19	28.6	94.9	91.8
PY	22	22.4	94.9	91.8
SCS	3.04	0.28	92.0	90.8
HL	3.01	0.22	88.3	55.1
CTFS	0.20	5.22	91.0	63.6
AFS	-1.02	8.79	92.7	84.6

Genotype assays. Maxxam Analytics Inc., Guelph, Ontario, Canada extracted DNA from the semen samples and Affymetrix Inc., South San Francisco, California, USA performed the SNP genotyping via the Affymetrix MegAllele GeneChip Bovine Mapping 10K SNP array (Affymetrix Inc. 2006). Four hundred and eighty four bulls were tested for 9,919 SNP, but 56 of bulls failed to produce genotyping results due to possible phenolchloroform contamination. 9,628 SNP produced data and of these, 4,856 SNP were physically located to chromosomes (in base-pairs) using the bovine genome sequence (Btau-2.0) obtained from the International Bovine Genome Sequencing Consortium (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20050310-freeze/) at the time of this research.

Table 2. Single nucleotide polymorphism (SNP) frequency, *Bos taurus* autosome (BTA) length, and average number of SNP per centiMorgan (cM) per BTA

BTA	Frequency	BTA Length (cM)	SNP/cM
1	257	147	1.75
2	250	120	2.08
3	267	129	2.07
4	202	110	1.84
5	171	132	1.30
6	226	127	1.78
7	177	138	1.28
8	162	121	1.34
9	171	110	1.55
10	217	106	2.05
11	285	127	2.24
12	135	113	1.20
13	174	93	1.88
14	141	92	1.54
15	160	96	1.67
16	203	98	2.08
17	148	110	1.35
18	147	84	1.75
19	144	93	1.55
20	148	76	1.96
21	100	95	1.05
22	149	80	1.86
23	152	71	2.13
24	150	69	2.18
25	126	64	1.97
26	112	75	1.50
27	78	66	1.18
28	94	56	1.69
29	116	69	1.68
Mean	167.66	98	1.71

Calculation of SNP statistics. Deviation from Hardy-Weinberg equilibrium of the SNP was tested using a Chi-square test with one degree of freedom. The departure from random mating, heterozygosity (H) and polymorphism information content (PIC) was also determined. Departure from random mating was calculated as the mean difference between the observed and expected number of heterozygotes under Hardy-Weinberg equilibrium and SNP H was calculated based on observed allele frequencies. Polymorphism information content for the SNP was calculated as shown in Guo and Elston (1999) and was the probability that one can determine if it was the maternal or paternal allele that an offspring has inherited from its parent assuming no cross-over during meiosis.

Building the linkage map. The cM positions for the SNP were interpolated using a microsatellite framework map available from the National Centre for Biotechnology Information, Bethesda, MD, USA (National Centre for Biotechnology Information 2006). This framework map was edited to allow for interpolation of SNP cM positions. When the marker order between the bp and cM maps changed, the crossing microsatellite with a pattern contrary to the other microsatellites in the same section was deleted. All microsatellite markers that had the same bp positions or had a cM position of 0.0 were removed. When two or more microsatellite markers had the same cM position only the marker with the lowest bp position was retained in the framework map. Once both the microsatellite bp and linkage map had the same order, the SNP's bp locations were interpolated to cM based on its location within a microsatellite bracket.

Variance component linkage analysis (VCLA). Due to the structure of the genotyped population and the complexity of the pedigree information, a restricted maximum likelihood method with variance component estimation was chosen (George *et al.* 2000). Single trait analysis was performed on a chromosome by chromosome basis. Single nucleotide polymorphisms with a PIC of 0.0 were excluded from the analysis. Using Loki (Heath 1997), the IBD probabilities were estimated at 1cM intervals starting at 0.0cM and ending after the last SNP position on a chromosome. Mixing in Loki was improved by setting the LM ratio (proportion of locus versus meiosis updates) to 0.5 (Daw *et al.* 1999). Two hundred thousand iterations were performed with a burn-in period of 1,000 iterations to achieve satisfactory IBD convergence. If a marker was closer than 0.1cM to the point where the IBD probability was to be estimated the IBD matrix sometimes turned out to be singular. In these cases, to avoid singularity, the QTL location was moved to either the right or the left of the original position, so that the minimum distance to the closest SNP was 0.1cM. ASReml (Gilmour *et al.* 2000) computed the mixed linear models at each IBD location to obtain parameter estimates for the random factors. The full fitted model was:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Z_1a} + \mathbf{Z_2v} + \mathbf{e}$$

where **y** is a vector of EBVs, μ is the population mean, $\mathbf{Z_1}$ is the incidence matrix for animal effects, **a** is a vector of the additive polygenic animal effects, $\mathbf{Z_2}$ is the incidence matrix for the QTL effects, **v** is a vector of the additive QTL effects, and **e** is the vector of residuals. The random effects **a**, **v**, and **e** were assumed to be independent and normally distributed: $\mathbf{a} \sim N\ (0, \mathbf{A}\sigma^2_a)$, $\mathbf{v} \sim N\ (0, \mathbf{G}\sigma^2_{QTL})$, and $\mathbf{e} \sim N\ (0, \mathbf{I}\sigma^2_e)$. Where **A** is the numerator relationship matrix, σ^2_a is the variance of the additive polygenic effects, **G** is the IBD probability matrix, σ^2_{QTL} is the additive QTL variance, **I** is the identity matrix, and σ^2_e is the residual variance. This model was then refitted without the $\mathbf{Z_2v}$ term. The QTL test performed was a likelihood ratio test (LR), where the maximum restricted likelihood of the full model was compared to the maximum restricted likelihood of the model missing the QTL effect. The additive relationship matrix (**A**) was the same for both models and included all relevant animals in the pedigree (5,615 animals). The sires' EBVs were assumed to have equal residual variances, given that most of the bulls had highly accurate proofs (Table 1).

Linkage disequilibrium single locus regression (LDRM). A primary linkage disequilibrium screen using regression on individual SNP genotypes was carried out (Grapes *et al.* 2004). Markers were assumed to be in LD with QTL in close proximity and the effect evaluated was additive only (QTL allele substitution effect). SNP with a minor allele frequency of less than or equal to 0.1 were excluded from the analysis. The following model was calculated at each SNP genotype location using ASReml:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z_1a} + \mathbf{e}$$

with $\mathbf{a} \sim N\ (0, \mathbf{A}\sigma^2_a)$ and $\mathbf{e} \sim N\ (0, \mathbf{I}\sigma^2_e)$, where **X** is the design matrix in which SNP genotypes were coded 0, 1 and 2 for 1-1, 1-2, and 2-2 allele combinations, respectively, and

b is the vector of coefficients of the regression on recoded SNP genotypes. As for VCLA, the sires' EBVs were assumed to have equal residual variances. Thus regressions were not weighted by the EBV accuracies.

Statistical Inference. The false discovery rate (FDR) (Benjamini & Hochberg 1995) was used to account for multiple hypotheses testing. All significance values computed in this study were on a 5% chromosome-wise FDR level. The significance values for VCLA were obtained from a mix of two chi-square distributions (Self & Liang 1987). In VCLA, due to the nature of FDR, the largest LR in a peak could possibly not be significant at 5% FDR, while positions with a lower LR around it were significant. In those cases the position with the largest LR was still reported as the peak, because FDR is not monotonic with respect to the probability of the test statistic and does not recognize the dependency of the tests. In LA, the tests for QTL at IBD positions close to each other are likely dependent (Fernando *et al.* 2004). The p values for the LDRM were taken from a two tailed t-test distribution and only SNP below a chromosomal FDR of 5% were reported.

Confidence intervals were not calculated as bootstrapping would have been too computationally intensive to carry out. In its stead logarithmic odds scores (LOD) were computed and two QTL on a chromosome were considered distinct if the LOD score dropped more than one point from the higher peak at a position between the two LR peaks.

Potential QTL and significant SNP associations were considered in agreement with QTL cited in the literature if they were within a confidence interval of a QTL in a published study, or, if such an interval was not available, they were within 5cM of a QTL position in a published analysis. When comparing the results to previously reported QTL, it is important to recognize that cM locations are relative and depend on the linkage map used in each study. Differences occur because linkage maps (cM) are calculated based on the amount of recombination between the genetic makers relative to the first marker evaluated on the chromosome. Therefore, comparisons to literature positions give only a coarse measure of QTL location agreement.

RESULTS

SNP and dataset statistics. The located SNP were moderately evenly distributed across the genome. Table 2 shows the SNP frequency and density per chromosome. The number of SNP and SNP density per *Bos taurus* autosome (BTA) varied from 78 to 285 and 1.18 to 2.24 SNP/cM, respectively. Seventeen percent of SNP were not in Hardy-Weinberg equilibrium at the 5% significance level and the departure from random mating was 0.04 (SE 0.003) which confirms that positive assortative mating had occurred in the sires' pedigrees. The mean H for the SNP was 0.31 (range 0.0 to 0.5) and per chromosome mean SNP H ranged between 0.26 and 0.35. The mean PIC was 0.25 (range 0.0 to 0.375) and the chromosomes with lowest and highest means were the same as the H results because PIC is highly dependent on H. Four hundred and thirty one of the 4856 SNP in this dataset had a PIC of 0.0. These SNP were non-segregating, or fixed, and would not have added any information to the analysis and were therefore removed.

Tracing back of the 17 European bull pedigrees revealed that 86% of their founders (sires with unknown parents) were of North American origin. Thus, the European bulls were strongly related to the North American bulls. In addition, the mean correlation of allele frequency of significant SNP associations between the European bulls and North American bulls in the dataset was 0.85 (range 0.70 to 0.93). This was very similar to the allele frequency correlations within the North American bulls of 0.87 (range 0.71 to 0.94). These results indicate that population stratification was likely not responsible for the significant results in our study.

Potential chromosomal regions of interest detected. The potential QTL found with VCLA, as well as the previous studies that are in agreement, can be seen in Table 3. Variance component linkage analysis detected a total of 102 potential QTL, including 15 for MY, six for FY, 52 for PY, four for SCS, 20 for CTFS and five for AFS. Twenty of these QTL were in agreement with QTL previously reported in the literature including four for MY, four for FY, 11 for PY and one for SCS. New potential QTL were found for MY (11), FY (2), PY (41), SCS (3), CTFS (20) and AFS (5).

Linkage disequilibrium single locus regression found 144 significant SNP associations, which are reported in Table 4. When there were more than one SNP within a 1cM bracket they were grouped together and the maximum absolute t-value is reported. The number of significant SNP associations found per trait was: 31 for MY, seven for FY, 22 for PY, 32 for SCS, 17 for HL, 14 for CTFS and 21 for AFS. As was the case with VCLA, a proportion of the positions, 48 of the 140 SNP, were in agreement with previous findings. These SNP included 23 for MY, two for FY, 14 for PY, and nine for SCS. The individual literature studies can be found in the footnotes of Table 4. New significant SNP associations were found in the respective traits: eight for MY, five for FY, eight for PY, 23 for SCS, 17 for HL, 14 for CTFS and 21 for AFS. A number of SNP were associated with phenotypic variation in both milk and protein yield and this could likely be explained by large genetic correlations between the two traits. The direction (positive versus negative) of the regression coefficients was in all cases the same if one SNP had a significant association with differences in phenotype for both traits.

Table 3. Potential QTL detected via variance component linkage analysis for milk yield (MY), protein yield (PY), fat yield (FY), somatic cell score (SCS), calving to first service (CTFS), and age at first service (AFS)

BTA	Location (cM)	Trait	LR[1]	FDR[2]
1	79	PY	13.8	-
1	109[3]	PY	9.0	-
1	135	PY	9.1	-
2	71	PY	11.0	0.06
2	105	CTFS	11.2	-
3	37[4]	MY	8.9	0.10
3	25[5]	PY	7.5	-
3	45[5]	PY	15.3	-
3	41	SCS	10.8	0.07
3	27	CTFS	10.2	0.09
3	34	CTFS	9.7	-
3	45	CTFS	9.1	-
3	68	CTFS	4.7	-
4	4	MY	4.8	-

BTA	Location (cM)	Trait	LR[1]	FDR[2]
4	14	MY	7.8	-
4	57	MY	4.9	-
4	86	MY	9.5	0.12
4	16	PY	10.7	0.06
4	85	PY	8.1	-
4	105	PY	6.9	-
5	1	MY	18.3	-
5	1	PY	23.9	-
5	2	CTFS	16.3	-
6	25[6]	PY	11.1	0.06
6	98	PY	10.5	-
6	50	CTFS	9.3	0.15
6	59	AFS	9.0	-
6	68	AFS	14.1	-
6	100	AFS	7.6	-
8	0	PY	11.1	-
8	38	PY	10.7	-
8	53	PY	12.3	-
8	85	PY	9.3	-
8	122	PY	8.1	-
9	8	PY	11.8	-
9	75[7]	PY	9.2	-
10	21	PY	5.8	-
10	99	PY	15.5	-
11	38	PY	6.5	-
11	66	PY	10.7	0.07
11	72	PY	8.9	-
11	96	PY	6.3	-
11	95	CTFS	11.4	-
11	108	CTFS	10.1	-
12	52	FY	6.4	-
12	67[3]	FY	8.4	0.21
12	90	FY	2.8	-
13	30	PY	13.2	-
13	74[4]	PY	11.8	-
13	20	CTFS	13.6	-

BTA	Location (cM)	Trait	LR[1]	FDR[2]
13	39	CTFS	6.8	-
13	90	CTFS	4.7	-
14	3[7]	FY	8.5	0.17
14	22[9]	SCS	11.1	-
14	53	SCS	9.2	-
14	74	SCS	6.2	-
14	3	AFS	9.0	0.13
14	62	AFS	5.0	-
15	17	PY	4.5	-
15	41	PY	10.0	-
15	62	PY	11.9	-
15	93	PY	7.5	-
15	6	CTFS	11.6	-
15	94	CTFS	11.0	-
16	86[3]	MY	11.5	-
16	59	PY	7.2	-
16	68	PY	9.6	-
16	85[3]	PY	19.6	-
16	98	PY	13.0	-
18	0	MY	10.1	-
18	13	MY	6.2	0.19
18	41[10]	MY	7.9	-
18	84	MY	5.7	-
18	0	PY	3.9	-
18	41	PY	9.8	-
18	65	PY	8.4	-
18	84[10]	PY	15.5	-
19	36	MY	16.2	-
19	60	MY	10.9	-
19	36[11]	FY	11.3	-
19	51[11]	FY	7.8	-
19	29	PY	20.3	-
19	59	PY	7.1	-
23	7[6]	MY	10.3	-
23	6	PY	10.4	-
23	26[11]	PY	5.7	-

BTA	Location (cM)	Trait	LR[1]	FDR[2]
23	42[11]	PY	7.5	0.08
24	7	PY	7.1	-
24	53	PY	11.6	-
24	54	CTFS	9.4	0.08
25	51	MY	11.2	-
25	47[6]	PY	9.0	-
25	52	PY	13.8	-
25	64	PY	6.5	-
25	57	CTFS	9.0	0.09
26	11	CTFS	7.6	-
26	72	CTFS	7.8	0.19
28	1	PY	8.1	-
28	11	PY	8.5	0.10
28	35	PY	6.0	-
29	8	CTFS	8.4	0.13
29	20	CTFS	4.0	-
29	60	CTFS	4.1	-

[1] LR = likelihood ratio test.

[2] FDR = false discovery rate, reported only if at the peak position FDR was higher than 0.05.

[3] In agreement with (Rodriguez-Zas *et al.* 2002)

[4] In agreement with (Ashwell *et al.* 2004)

[5] In agreement with (Heyen *et al.* 1999)

[6] In agreement with (Viitala *et al.* 2003)

[7] In agreement with (Georges *et al.* 1995)

[8] In agreement with (Khatkar *et al.* 2004)

[9] In agreement with (Zhang *et al.* 1998)

[10] In agreement with (Olsen *et al.* 2002)

[11] In agreement with (Bennewitz *et al.* 2003)

The total number of potential QTL and significant SNP associations found per trait and method of analysis are summarized in Table 5. This table also lists the number of BTA on which both methods found significant associations and agreement between the two methods was greatest for the milk production traits and less for SCS and HL.

DISCUSSION

Choice of dependent variable. Estimated breeding values were used as the dependent variable in both VCLA and LDRM. Previous studies have found that using EBVs instead of daughter yield deviations (DYD) or de-regressed EBVs either does not significantly reduce power (Israel & Weller 1998) or only slightly reduces power (Thomsen *et al.* 2001). The potential for EBVs to cause a downward bias was reduced by the high mean EBV accuracy in this study (Table 1).

The EBVs were not weighted in the analysis to account for accuracy because mean EBV accuracy was high and the potential increase in power would have been minimal. The extent to which EBVs are regressed towards zero decreases as the amount of information available to calculate the EBV increases. Thus, EBVs of bulls with lower accuracy have a smaller variance than the EBVs of bulls with higher accuracy (Israel & Weller 1998). Bulls with EBVs of lower accuracy would therefore have a less impact on the results and this would mitigate the potential for causing bias when not weighting EBVs. However, this might not be the case for DYD and de-regressed EBV, because bulls with EBVs of lower accuracy would de-regress more and potentially might influence the results to a larger extent.

Table 4. Significant single nucleotide polymorphism (SNP) associations from linkage disequilibrium single locus regression analysis for milk yield (MY), fat yield (FY), protein yield (PY), somatic cell score (SCS), herd life (HL), calving to first service (CTFS), and age at first service (AFS)

BTA	Location (cM)	Trait	No. of SNP	Max. |t-value|	Mean (r^2)[1]
1	47[3]	MY	1	3.5	
1	142[4]	MY	1	4.0	
1	140	FY	1	3.7	
1	47[5]	PY	1	3.5	
1	142	PY	1	3.8	
1	131[5]	SCS	4	4.3	0.86
1	140	AFS	1	3.7	
2	109[6]	SCS	1	3.7	
2	24	CTFS	1	4.4	
3	50[3]	MY	1	3.8	

BTA	Location (cM)	Trait	No. of SNP	Max. |t-value|	Mean (r^2)[1]
3	50[5]	PY	1	3.4	
3	103	PY	1	3.5	
4	32	SCS	1	3.9	
5	3	MY	1	3.4	
5	77	SCS	2	3.6	0.83
5	86	SCS	1	3.8	
5	51	AFS	1	3.4	
5	78	AFS	2	3.0	NA[2]
5	82	AFS	2	3.5	0.84
6	52	SCS	1	3.0	
6	72	SCS	1	4.2	
6	73	SCS	1	2.9	
6	74	SCS	1	3.4	
6	75	SCS	1	2.9	
6	78	SCS	1	2.8	
6	81	SCS	1	2.8	
6	82	SCS	1	3.2	
6	83	SCS	3	3.9	0.65
6	46	HL	1	3.4	
6	51	HL	1	3.5	
6	53	HL	2	3.2	0.08
6	56	HL	1	3.1	
6	59	HL	1	3.6	
6	61	HL	1	3.2	
6	73	HL	1	3.1	
6	83	HL	2	2.9	0.19
6	84	HL	2	2.9	0.49
7	60	MY	2	3.6	0.99
7	73[7]	MY	2	3.9	0.38
7	75[7]	MY	1	2.9	
7	82[7]	MY	1	3.1	
7	60	PY	2	3.1	0.99
7	68	PY	1	2.8	
7	73[7]	PY	2	3.5	0.38
7	75[7]	PY	1	3.1	
7	95	PY	1	2.9	

BTA	Location (cM)	Trait	No. of SNP	Max. |t-value|	Mean (r^2)[1]
8	101	FY	1	3.5	
9	57[3]	MY	2	3.4	0.58
9	36	HL	1	3.7	
9	39	HL	1	3.7	
10	10	CTFS	1	3.5	
10	20	CTFS	6	3.4	0.75
10	23	CTFS	1	2.8	
10	31	CTFS	1	3.4	
10	42	CTFS	1	2.8	
11	119	SCS	4	3.5	0.21
11	60	AFS	1	4.2	
12	59	MY	1	3.8	
13	63	HL	1	4.8	
13	35	CTFS	1	3.4	
14	4[8]	MY	8	4.0	0.97
14	5[8]	MY	1	3.3	
14	6[8]	MY	2	3.5	0.83
14	12[6]	MY	3	3.3	0.75
14	42	MY	1	3.0	
14	28[6]	FY	1	3.4	
14	4[6]	PY	4	2.9	0.95
14	12[6]	PY	3	3.3	0.73
14	4	AFS	8	3.9	0.99
14	5	AFS	1	2.8	
14	6	AFS	1	2.6	
14	50	AFS	1	3.0	
14	54	AFS	1	2.6	
15	75	MY	1	3.3	
16	46	MY	1	3.5	
16	46	PY	1	3.5	
16	48[5]	SCS	1	3.5	
16	8	CTFS	2	4.1	0.87
18	50	AFS	1	3.2	
19	9[3]	FY	1	3.6	
19	47[6]	SCS	1	3.2	
19	58	HL	1	3.6	

BTA	Location (cM)	Trait	No. of SNP	Max. |t-value|	Mean (r^2)[1]
23	52[9]	SCS	2	3.5	0.98
23	2	AFS	1	3.8	
25	55	SCS	1	3.4	
26	55[7]	MY	1	3.3	
27	28	SCS	1	3.3	
28	33	FY	3	3.1	0.68
28	45	FY	1	3.3	
28	48	FY	1	3.0	
28	33	PY	1	2.9	
28	44[5]	PY	1	3.1	
28	46[5]	PY	1	3.3	
29	5	SCS	1	3.1	
29	67	SCS	1	3.3	

[1] Mean (r^2) = mean r^2 (linkage disequilibrium) between SNP

[2] NA= r^2 was not available for these two SNP

[3] In agreement with (Khatkar *et al.* 2004)

[4] In agreement with (De Koning *et al.* 2001)

[5] In agreement with (Rodriguez-Zas *et al.* 2002)

[6] In agreement with (Bennewitz *et al.* 2003)

[7] In agreement with (Boichard *et al.* 2003)

[8] In agreement with (Grisart *et al.* 2004)

[9] In agreement with (Ashwell *et al.* 1998)

Variance component linkage analysis. The VCLA method located 102 potential QTL. The relatively large number of new QTL found was promising. The great efficiency of QTL detection for protein yield (52 potential QTL) could be partly explained by the fact that this was one of the traits used for selective genotyping. While the selective genotyping approach seemed to have shown benefits for protein yield QTL, it has not shown an equally strong performance in SCS, where only four QTL were called significant. No potential QTL were found for HL with VCLA, which was possibly due to a lower mean EBV accuracy of HL (Table 1). The improved QTL detection in SCS over HL might have been due to the higher SCS EBV accuracy, the minor selective genotyping carried out and because SCS EBVs were available for all bulls genotyped, which was not the case for HL. It was also encouraging that for CTFS and AFS, 20 and five potential QTL were found, respectively. The improved result over previous studies in detecting QTL for traits of lower heritability showed that using a dense SNP map can increase power in linkage analysis.

Linkage disequilibrium regression method. The LDRM was successful and a total of 144 significant SNP associations were detected. In traits of low heritability, such as SCS, HL, CTFS and AFS, the LDRM found a large number of significant SNP associations. This makes it a viable choice for future genome scans in dairy cattle.

Single nucleotide polymorphisms in very close proximity to each other were often all significantly associated with a particular phenotype. The r^2 (Hill & Robertson 1968) was calculated according to the same guidelines as Sargolzaei et al. (2008) to determine the amount of LD between significant SNP associations within a 1cM bracket. As can be seen in Table 4, in most cases the SNP were in very strong LD and this suggests that they are all associated with the same QTL. However, in some cases the SNP exhibit low LD (e.g. BTA 6, 53cM, HL, $r^2 = 0.08$) and may be associated with the same or another QTL.

Comparison of results. The VCLA method did produce the expected results in locating DGAT (Grisart *et al.* 2004) at position 3cM, which is a QTL with a large effect in fat yield located on chromosome 14 at position 0.1cM (Khatkar *et al.* 2004). The LDRM showed a significant SNP association with fat yield at position 28cM which is within the confidence interval calculated by Bennewitz et al. (2003) but falls outside of the Khatkar et al. (2004) meta-analysis confidence range. The reason for this suboptimal result of the LDRM may be in the poor SNP distribution at the beginning of chromosome 14. No SNP were present until 4cM on that chromosome and the LDRM may be more sensitive to gaps in SNP distribution.

Neither VCLA or the LDRM showed conclusive results for the ABCG2 mutation located in the middle of BTA 6 (Cohen-Zinder *et al.* 2005) which has an effect on milk yield and composition in Holstein cattle. While VCLA detected two protein QTL on BTA 6, one at 25cM and another at 98cM, they were outside of the confidence interval of the meta-analysis for this trait (Khatkar *et al.* 2004). At the location of the causative mutation identified by Cohen-Zinder et al. (2005) the SNP distribution showed a gap and the LDRM possibly did not overcome that.

Table 5. Total number of potential QTL or significant SNP found per trait for variance component linkage analysis (VCLA) and linkage disequilibrium single locus regression (LDRM), the number of potential significant regions found with LDRM when accounting for a confidence interval (CI) of 10cM, and the *Bos taurus* autosomes (BTA) on which both methods found significant associations.

	Significant BTA regions			Agreement
Traits	VCLA	LDRM	LDRM (CI = 10)	VCLA + LDRM
MY	15	31	15	BTA 3, 5, 16
FY	6	7	6	BTA 14, 19
PY	52	22	13	BTA 1, 3, 16, 29
SCS	4	32	15	-
HL	-	17	7	-
CTFS	21	14	6	BTA 2, 13
AFS	5	21	9	BTA 14
Total	102	144	71	12

In some traits, such as SCS, HL and AFS, VCLA identified fewer potential regions than LDRM. While VCLA maps potential QTL, the LDRM detects significant SNP associations. Therefore multiple SNP in close proximity could be in LD with the same QTL. The extent to which not accounting for LD between several SNP and one QTL could have inflated the number of potential significant regions found with the LDRM was investigated. Significant SNP were counted as being in LD with the same QTL when they were within 5cM of each other (confidence interval = 10cM). When counted by this method, 71 significant chromosome regions were found with the LDRM. As can be seen in Table 5, even when grouping SNP into 10cM groups the LDRM detected more significant regions in traits of low heritability while finding equivalent or fewer numbers of significant associations in traits of moderate heritability.

Overall, the LDRM discovered a more uniform number of chromosomal regions across traits. There were few BTA on which both methods found significant associations as can be seen in Table 5. Significant regions for PY that resulted from both methods were found on four BTA, which was the most of any trait in this study. The differences in QTL discoveries among the two methods are likely related to the sample size that was likely not large enough and was limiting the power of both approaches, associated with the inherent differences in the methods with respect to the use of the LD information.

The VCLA utilizes LD within families to calculate IBD probabilities and it is more efficient when the average r^2 is low (< 0.2), as in the case of low density of markers, compared to LDRM. However LDRM becomes equally efficient to VCLM when the density of markers increases and the average r^2 is high (Grapes *et al.* 2004; Zhao *et al.* 2006; Goddard & Hayes 2007). This is so, because LDRM requires a dense and preferably uniformly distributed marker map, and, if there are gaps with no SNP, the power to detect a QTL diminishes (Grapes *et al.* 2004). In the current study, while the average marker density was 1.7 per cM, there were gaps in the distribution of SNP and differences in SNP density across chromosomes. This could have lead to different rates of success in detecting QTL between LDRM and VCLA, depending on the chromosomes and/or chromosomal regions analyzed. The r^2 between SNP differs across genomic regions and chromosomes (Sargolzaei *et al.* 2008) and could lead to differences between VCLA and LDRM depending on where the QTL were located.

Another option of QTL analysis is to use a combined linkage disequilibrium and linkage analysis in a LDLA approach (Olsen *et al.* 2005), which can narrow confidence intervals considerably when compared to LA, but this was beyond the scope of this study.

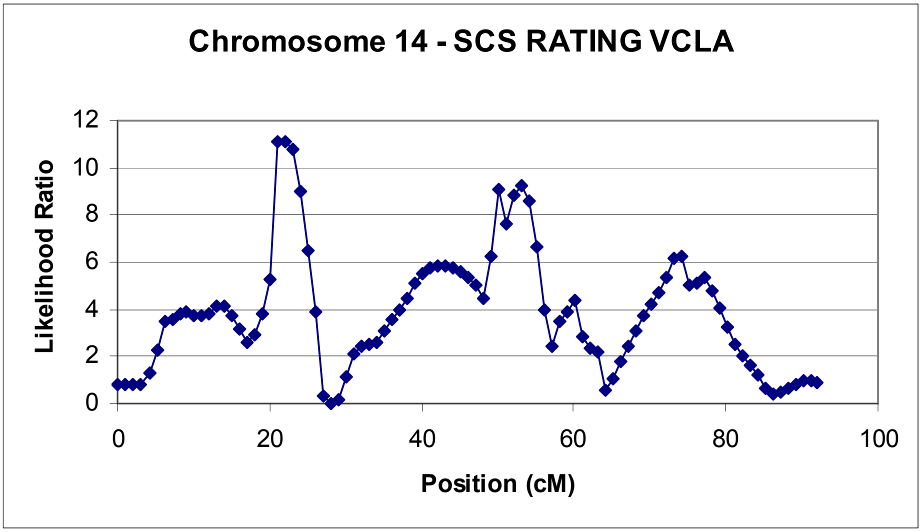
A LR profile of VCLA and a t-test profile representative of the LDRM can be seen in Figure 1. The t-test pattern of the LDRM was more erratic than the LR profile of VCLA, which was likely due to the LDRM treating each SNP as a separate regression whereas in VCLA all SNP on a chromosome were considered together to calculate IBD at each position. In VCLA, this lead to a moderation of the variability in the LR profile. The BTA 14 SCS analyses in Figure 1 also show an example of suggestive unison between the two methods even when only one of them showed significant results. Three significant peaks were found with VCLA, but none of the SNP associations were significant in LDRM. However, considering Figure 1b, it is possible to see that the LDRM yielded three regions where the SNP have higher t-test values, albeit not significant.

The primary focus of this study was on detecting significant chromosome regions for further analysis and not estimating the effects of these regions. In whole genome scans, QTL effects are known to be overestimated, because when the test statistic is maximized over the many point-wise tests in the genome, the estimates of the parameter(s) characterizing the locus-specific effects (e.g. QTL variance) are effectively maximized as well (Göring et al., 2001). There was strong evidence for both methods that this was the case in this study (results not shown). Another reason for an upward bias might be the presence of more than one QTL in a chromosomal region (Allison *et al.* 2002) which was not accounted for in the single-QTL analyses performed. In addition, the partial selective genotyping based on four traits carried out in this study may have added to the overestimation in some of the analyzed traits.

In LDRM, an effect estimate which was based on few genotypes could have been a source of bias. Single nucleotide polymorphisms with a minor allele frequency of less than or equal to 0.1 (1441 SNP) were excluded from the analysis for that reason. The most significant t values and effects were checked for all traits to determine if a possible bias existed and the minor genotype frequency was never below 0.05, which corresponds to 23 bulls with this genotype, in the SNP investigated. Therefore, this was likely not a large source of bias in this study.

The LDRM had the lower computational requirement of the two methods. The LDRM analyzed one chromosome in approximately 0.5 hour, whereas VCLA needed on average four hours, not including IBD calculation time, on a server with 16 Gigabytes (GB) of 400 MHz CL3 memory, eight 500 GB SATA disk drives (Iomega, San Diego, California) and four jobs running simultaneously. When computing power or time is limiting, LDRM is more useful than VCLA for a first QTL scan with a dense SNP map.

a) VCLA



b) LDRM

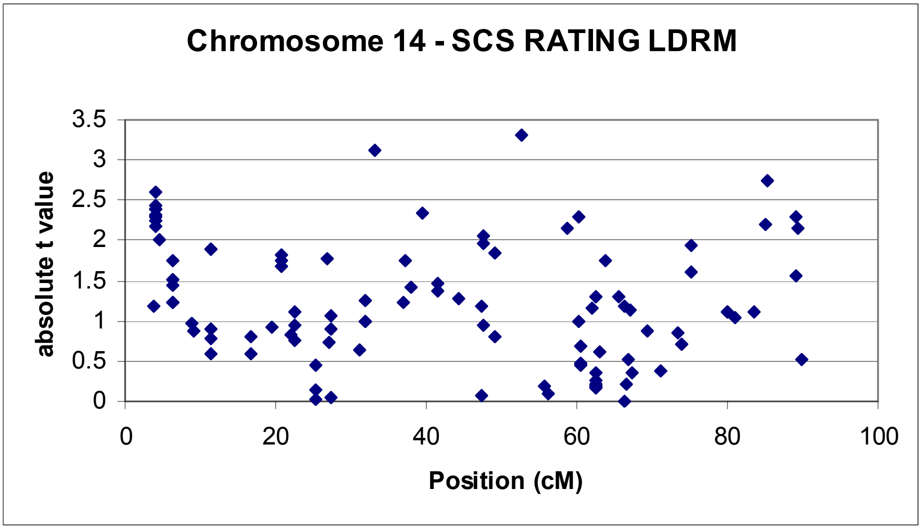


Figure 1. Test statistic profile for somatic cell score (SCS) on chromosome 14: a) Variance component linkage analysis (VCLA) likelihood ratio profile and b) Linkage disequilibrium single locus regression (LDRM) absolute t-test profile.

CONCLUSIONS

The two genome scans resulted in 102 potential QTL and 144 significant SNP associations for production, functional and reproduction traits in the Holstein dairy sires genotyped. A large number of potential chromosomal regions of interest for traits of low heritability were detected. This study was one of the first applications of the LDRM to real dense SNP data and it showed that the LDRM was capable of detecting significant SNP associations at an average SNP density of 1.7 SNP per cM. The LDRM located more potential chromosomal regions of interest than VCLA in traits of low heritability. Future work with the LDRM and the full set of SNP marker locations should increase its statistical power.

ACKNOWLEDGEMENTS

The authors thank L'Alliance Boviteq Inc. (SEMEX Alliance, Canada) and the Ontario Centre for Agriculture Genomics (Challenge Fund) for their financial support. We are also grateful for the input of John Woolliams and the two anonymous reviewers on this manuscript.

REFERENCES

Affymetrix Inc. Affymetrix MegAllele GeneChip Bovine 10K SNP Array. Affymetrix Inc., South San Francisco, CA, USA . 2006. 8-10-2006.

Allison, D. B., J. R. Fernandez, M. Heo, S. Zhu, C. Etzel *et al.* 2002 Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *Am.J.Hum.Genet.* **70**: 575-585.

Ashwell, M. S., Y. Da, C. P. Van Tassell, P. M. VanRaden, R. H. Miller *et al.* 1998 Detection of putative loci affecting milk production and composition, health, and type traits in a United States Holstein population. *J. Dairy Sci.* **81**: 3309-3314.

Ashwell, M. S., D. W. Heyen, T. S. Sonstegard, C. P. Van Tassell, Y. Da *et al.* 2004 Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. *J. Dairy Sci.* **87**: 468-475.

Ashwell, M. S., D. W. Heyen, J. I. Weller, M. Ron, T. S. Sonstegard *et al.* 2005 Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle. *J. Dairy Sci.* **88**: 4111-4119.

Benjamini, Y., and T. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc.* **85**: 289-300.

Bennewitz, J., N. Reinsch, C. Grohs, H. Leveziel, A. Malafosse *et al.* 2003 Combined analysis of data from two granddaughter designs: A simple strategy for QTL confirmation and increasing experimental power in dairy cattle. *Genet. Sel. Evol.* **35**: 319-338.

Boichard, D., C. Grohs, F. Bourgeois, F. Cerqueira, R. Faugeras *et al.* 2003 Detection of genes influencing economic traits in three French dairy cattle breeds. *Genet. Sel. Evol.* **35**: 77-101.

Canadian Dairy Network. Minimum Criteria for an Official Bull Proof. Canadian Dairy Network . 2007. Accessed 5-3-2007.

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, W. A. Everts-van der *et al.* 2005 Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res.* **15**: 936-944.

Darvasi, A., and M. Soller, 1992 Selective genotyping for determination of linkage between a marker locus and a quantitative trait loci. *Theor. Appl. Genet.* **85**: 353-359.

Daw, E. W., S. C. Heath, and E. M. Wijsman, 1999 Multipoint oligogenic analysis of age-at-onset data with applications to Alzheimer disease pedigrees. *Am.J.Hum.Genet.* **64**: 839-851.

De Koning, D. J., N. F. Schulman, K. Elo, S. Moisio, R. Kinos *et al.* 2001 Mapping of multiple quantitative trait loci by simple regression in half-sib designs. *J.Anim Sci.* **79**: 616-622.

Fernando, R. L., D. Nettleton, B. R. Southey, J. C. Dekkers, M. F. Rothschild *et al.* 2004 Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**: 611-619.

George, A. W., P. M. Visscher, and C. S. Haley, 2000 Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**: 2081-2092.

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto *et al.* 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907-920.

Gilmour A. R., R. Thompson, B. R. Cullis, and S. J. Wellham, 2000 *ASReml Reference Manual*. NSW Department for Primary Industries, New South Wales.

Goddard, M. E., and B. J. Hayes, 2007 Genomic selection. *J. Anim. Breed. Genet.* **124**: 323-330.

Grapes, L., J. C. Dekkers, M. F. Rothschild, and R. L. Fernando, 2004 Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics* **166**: 1561-1570.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim *et al.* 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc.Natl.Acad.Sci.U.S.A* **101**: 2398-2403.

Guo, X., and R. C. Elston, 1999 Linkage information content of polymorphic genetic markers. *Hum.Hered.* **49**: 112-118.

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am.J.Hum.Genet.* **61**: 748-760.

Heyen, D. W., J. I. Weller, M. Ron, M. Band, J. E. Beever *et al.* 1999 A genome scan for QTL influencing milk production and health traits in dairy cattle. *Physiol Genomics* **1**: 165-175.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. *Theor.Appl.Genet.* **38**: 226-231.

Israel, C., and J. I. Weller, 1998 Estimation of candidate gene effects in dairy cattle populations. *J. Dairy Sci.* **81**: 1653-1662.

Khatkar, M. S., P. C. Thomson, I. Tammen, and H. W. Raadsma, 2004 Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genet.Sel. Evol.* **36**: 163-190.

Kuhn, C., J. Bennewitz, N. Reinsch, N. Xu, H. Thomsen *et al.* 2003 Quantitative trait loci mapping of functional traits in the German Holstein cattle population. *J.Dairy Sci.* **86**: 360-368.

National Centre for Biotechnology Information. Bovine Microsatellite Framework cM Map. NCBI . 2006. Accessed 17-5-2006.

Olsen, H. G., L. Gomez-Raya, D. I. Vage, I. Olsaker, H. Klungland *et al.* 2002 A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. *J.Dairy Sci.* **85**: 3124-3130.

Olsen, H. G., S. Lien, M. Gautier, H. Nilsen, A. Roseth *et al.* 2005 Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6. *Genetics* **169**: 275-283.

Polineni, P., P. Aragonda, S. R. Xavier, R. Furuta, and D. L. Adelson, 2006 The bovine QTL viewer: a web accessible database of bovine Quantitative Trait Loci. *BMC.Bioinformatics.* **7**: 283.

Rodriguez-Zas, S. L., B. R. Southey, D. W. Heyen, and H. A. Lewin, 2002 Interval and composite interval mapping of somatic cell score, yield, and components of milk in dairy cattle. *J.Dairy Sci.* **85**: 3081-3091.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer, 2008 Extent of linkage disequilibrium in Holstein cattle in North America. *J.Dairy Sci.* **5**: 2106-2117.

Schrooten, C., H. Bovenhuis, W. Coppieters, and J. A. van Arendonk, 2000 Whole genome scan to detect quantitative trait loci for conformation and functional traits in dairy cattle. *J.Dairy Sci.* **83**: 795-806.

Self, S. G., and K. Y. Liang, 1987 Asymptoyic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J.Am.Stat.Assoc.* **82**: 605-610.

Smith, C., 1967 Improvement of metric traits through specific genetic loci. *Anim.Prod.* **9**: 349-358.

Thomsen, H., N. Reinsch, N. Xu, C. Looft, S. Grupe *et al.* 2001 Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for QTL. *J. Anim. Breed. Genet.* **118**: 357-370.

Van Doormaal B.J. Genetic Evaluation of Dairy Cattle in Canada. Canadian Dairy Network . 2007. Accessed 5-3-2007.

Viitala, S. M., N. F. Schulman, D. J. De Koning, K. Elo, R. Kinos *et al.* 2003 Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle. *J.Dairy Sci.* **86**: 1828-1836.

Weller, J. I., Y. Kashi, and M. Soller, 1990 Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J.Dairy Sci.* **73**: 2525-2537.

Zhang, Q., D. Boichard, I. Hoeschele, C. Ernst, A. Eggen *et al.* 1998 Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics* **149**: 1959-1973.

Zhao H.H., Dekkers J.C. & Fernando R.L. Power and precision of regression-based linkage disequilibrium mapping of QTL. 8th World Congress of Genetics Applied to Livestock Production. Proc.of the 8th WCGALP. 2006.

Chapter 3

Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach

Hans D. Daetwyler[1,2] , Beatriz Villanueva[3], and John A. Woolliams[1]

[1]Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin, Midlothian, EH25 9PS, UK; [2]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6700 AH, The Netherlands; [3]Sustainable Livestock Systems, Scottish Agriculture College, Edinburgh, EH9 3JG, UK

Published in PLoS ONE, 2008, 3(10): e3395.

ABSTRACT

The prediction of the genetic disease risk of an individual is a powerful public health tool. While predicting risk has been successful in diseases which follow simple Mendelian inheritance, it has proven challenging in complex diseases for which a large number of loci contribute to the genetic variance. The large numbers of single nucleotide polymorphisms now available provide new opportunities for predicting genetic risk of complex diseases with high accuracy.

We have derived simple deterministic formulae to predict the accuracy of predicted genetic risk from population or case control studies using a genome-wide approach and assuming a dichotomous disease phenotype with an underlying continuous liability. We show that the prediction equations are special cases of the more general problem of predicting the accuracy of estimates of genetic values of a continuous phenotype. Our predictive equations are responsive to all parameters that affect accuracy and they are independent of allele frequency and effect distributions. Deterministic prediction errors when tested by simulation were generally small. The common link among the expressions for accuracy is that they are best summarized as the product of the ratio of number of phenotypic records per number of risk loci and the observed heritability.

This study advances the understanding of the relative power of case control and population studies of disease. The predictions represent an upper bound of accuracy which may be achievable with improved effect estimation methods. The formulae derived will help researchers determine an appropriate sample size to attain a certain accuracy when predicting genetic risk.

INTRODUCTION

Genetic risk of disease is an important component of overall risk of disease in addition to environmental, socio-economic, and behavioral risk factors. Therefore, predicting the genetic risk of disease for an individual is a powerful tool in taking preventative measures against the onset of the disease. Such predictions from genetic testing are relatively straightforward when a disease is caused by one or few genes. However, when a disease is of complex inheritance, the genetic risk of the disease may be associated with many loci, each explaining only a small portion of the genetic variance (e.g. Hayes & Goddard 2001; Valdar *et al.* 2006). In this case, the prediction of genetic risk of disease of a particular individual becomes more challenging. Currently, prediction of risk for complex diseases is based mainly on pedigree analysis but this approach yields predictions of risk that are of low precision; for example predictions would be identical for full siblings without offspring, yet the genetic variation among them accounts for half or more of the genetic variance (Falconer & Mackay 1996; Bijma & Woolliams 1999).

The identification of very large numbers of single nucleotide polymorphisms (SNP) has enabled the use of genome-wide association studies (GWA) to detect alleles that are associated with risk for complex diseases (Hirschhorn & Daly 2005), such as Type II Diabetes and Crohn's disease (Wellcome Trust Case Control Consortium 2007). In tandem with this substantive increase of SNP data, several methods for quantifying and/or predicting genetic risk of disease from multiple genes have been put forward (Pharoah *et al.* 2002; Janssens *et al.* 2006). Wray et al. (2007) extended these methods by using an GWA approach to estimate the individual genetic risk of disease. Unlike the risk estimates obtained using only pedigree, the estimates resulting from such a GWA approach are more precise by allowing for differentiation among full-siblings. In addition, no pedigree or family history is needed either for estimating risk in one genotyped sample from the population or for predicting risk in a fresh sample. Similar genome-wide methodology has been proposed in animal and plant breeding to estimate additive genetic values for quantitative traits (Meuwissen *et al.* 2001; Xu 2003). One critical difference between the two genome-wide approaches is that Wray et al. (2007) set a significance threshold for the loci selected for disease prediction, whereas Meuwissen et al. (2001) use all loci regardless of whether they affect or not the trait considered. The approach of Meuwissen et al. (2001) therefore attempts to achieve the maximum estimate precision of the complete genetic value for a given dataset by including loci that may have too small of an effect to achieve statistical significance, and, thus, reduces the overestimation of allele effects (Goring *et al.* 2001).

Wray et al. (2007) computed the precision of the individual genetic risk estimates by simulation. While simulation studies are useful in getting initial results on the number of phenotypic records needed to achieve a desired level of accuracy, they are computer intensive and time consuming with large numbers of markers. Most importantly, they do not provide a deep insight on how all variables that affect accuracy interact. Therefore, it is desirable to develop deterministic equations that are responsive to all variables that influence accuracy.

Here we present simple expressions for the genome-wide accuracy of prediction of genetic disease risk. We derive general expressions for continuous traits and the necessary extensions for dichotomous disease traits with data obtained either from population studies or case control studies. The predictions are tested by computer simulation under a variety of parameters influencing accuracy, such as, for example, disease prevalence, heritability and distributions of allele effects and frequencies

MATERIAL AND METHODS

Derivation of Equations. The predicted accuracy that is derived below represents the upper bound that can be achieved when estimating effects in one population sample and then predicting individual genetic risk in another sample from the same population. Throughout this article the accuracy of predicted genetic risk ($r_{g\hat{g}}$) is defined as the correlation between true and predicted genetic values. One advantage of using $r_{g\hat{g}}$ is that the factors influencing it can be clearly derived using the principles of population genetics, as we show below. We will first derive equations that are predictive of $r_{g\hat{g}}$ for a genome-wide approach with a continuous phenotype, such as height, assuming a population study where individuals are sampled at random. These will then be adapted to predict disease risk for a dichotomous phenotype ('affected' or 'unaffected') with an underlying continuous liability. The equations are then further adapted to the situation of case control data.

Continuous phenotype. We will assume that there are n_G potential loci affecting a trait which are independent, biallelic and acting additively, where n_G may be large. These loci may be candidate genes or genetic markers of which a significant proportion may have zero effects. For locus $j, j = 1 \ldots n_G$, let a randomly chosen reference allele for that locus have frequency p_j and true allelic substitution effect β_j. We shall assume without loss of generality that the distribution of allele frequencies p_j is symmetric about $p = {}^1/_2$, and likewise that allelic effects β_j are symmetric about $\beta = 0$. No further distributional assumptions will be made here on p_j and β_j, so for example, many of the allele segregating may have negligible or zero effect. No assumptions are made concerning the covariance between p_j and β_j in the populations sampled. We intend to derive the accuracy of the prediction of the additive genetic value ($r_{g\hat{g}}$) of an individual that can be achieved after the measurement of n_P phenotypes.

An estimate of the effect of each allele may be obtained by regression of the phenotypic records on the genotypes one locus at a time because the loci are independently segregating. Assume the population variance of the phenotypes is 1. The estimated allele substitution effect will be $\hat{\beta}_j$ with expectation $E[\hat{\beta}_j] = \beta_j$, and is obtained by regressing the phenotypes on the observed number of reference alleles in the genotype, denoted x_{ij} for individual i and locus j (i.e. x_{ij} = 0, 1, or 2). The sampling variance of the allele estimate is $var(\hat{\beta}_j - \beta_j) = \sigma_e^2/S_{xx,j}$ where σ_e^2 is the residual variance after regression on x_{ij} and $S_{xx,j} = n_P var(x_{ij})$ is the adjusted sums of squares for x_{ij}. Although not assumed here, when the population is in Hardy-Weinberg equilibrium $S_{xx,j}$ is given by $2n_P\, p_j(1-p_j)$. For the present, we shall conservatively take $\sigma_e^2 = 1$, which underestimates the accuracy of the prediction.

Our aim is to predict the accuracy of a new population sample, so we apply the original estimates to a new sample of the same population. Values referring to the second sample will be 'dashed', hence individual i from the second sample has x'_{ij} alleles at locus j. The additive genetic value of *i* is given by $g_i = \sum_{loci\, j} x'_{ij}\beta_j$ with estimate $\hat{g}_i = \sum_{loci\, j} x'_{ij}\hat{\beta}_j$. Then $r_{g\hat{g}}^2 = [cov(g_i, \hat{g}_i)]^2/[var(g_i)var(\hat{g}_i)]$. Noting that $\hat{g}_i$ can be re-written as $\sum_{loci\, j} x'_{ij}[\beta_j + (\hat{\beta}_j - \beta_j)]$ with $cov(\beta_j, \hat{\beta}_j - \beta_j) = 0$, it is seen that $cov(g_i, \hat{g}_i) = var(g_i)$ and that $r_{g\hat{g}}^2 = var(g_i)/var(\hat{g}_i)$. Of these remaining terms, $var(g_i) = h_o^2$, where h_o^2 is the observed heritability for the trait, assuming the phenotypic variance is 1. Again using the decomposition $\hat{\beta}_j = \beta_j + (\hat{\beta}_j - \beta_j)$, it can be shown that $var(\hat{g}_i) = h_o^2 + \sum_{loci\, j} var(x'_{ij})[n_P\, var(x_{ij})]^{-1}$, following from (i) the independence of the loci and (ii) the sampling variance of $\hat{\beta}_j$ derived earlier. Finally $var(x'_{ij}) = var(x_{ij})$, since the second sample comes from the same population, so $r^2_{g\hat{g}} = h^2_o[h^2_o + n_G/n_P]^{-1}$, and substituting $\lambda = n_P/n_G$ gives

$$r_{g\hat{g}} = \sqrt{\frac{\lambda h_o^2}{\lambda h_o^2 + 1}}. \qquad (1)$$

Therefore accuracy is seen to be a function of the product of the observed heritability h^2_o and the ratio of the number of phenotypes recorded to the number of loci involved, λ. A second order correction to relax the assumption $\sigma^2_e = 1$ is given in Appendix S1, where it is shown to result in an upward correction to $r_{g\hat{g}}$ of fractional magnitude $\approx {}^1/_2\, r^4_{g\hat{g}}\, \lambda^{-1}$.

Dichotomous disease phenotype. We shall now derive the accuracy of predicting individual genetic risk to disease ($r_{g\hat{g}}$) in a random population sample by considering disease prevalence in a liability model [9]. For a disease with prevalence q, phenotypes are defined as $s_i = 0$ for unaffected, and $s_i = 1$ for affected, so $E[s_i] = q$ and $var(s_i) = q(1-q)$. Individuals with the highest liability are affected by the disease. Let liability be y_i, scaled so $E[y_i] = 0$ and $var(y_i) = 1$, and β_j is the regression of liability on the number of reference alleles at locus *j*. The linear predictor of s_i on y_i is given by $s_i = q + qi_q y_i$ (Robertson 1961), where i_q equals the mean liability of affected individuals, which we will term the selection intensity (Falconer & Mackay 1996) corresponding to the prevalence of the disease in the population. Let the slope of the regression of s_i on x_{ij} be $\hat{\pi}_j$, then $E[\hat{\pi}_j] = qi_q\beta_j$, with sampling variance, estimated conservatively using the phenotypic variance $q(1-q)$

$$var(\hat{\pi}_j) = q(1-q)\,[n_P\, var(x_{ij})]^{-1}. \qquad (2)$$

The coefficients $\hat{\pi}_j$ may be rescaled to give estimates $\hat{\beta}_j = \hat{\pi}_j/(qi_q)$, with sampling variance

$$var(\hat{\beta}_j) = (1-q)\,[n_P\, var(x_{ij})\, q\, i^2_q]^{-1}. \qquad (3)$$

Repeating the argument outlined above for a continuous phenotype with $var(g_i) = cov(g_i, \hat{g}_i) = h^2_l$, and

$var(\hat{g}_i) = h_l^2 + [n_G q(1-q) var(x_{ij}^{'})].(1-q)/[n_P var(x_{ij}) q i_q^2]^{-1}$, where h_l^2 is the heritability on the liability scale. Simplifying terms results in:

$$r_{g\hat{g}}^2 = \frac{n_P h_l^2 q\, i_q^2}{n_P h_l^2 q\, i_q^2 + n_G(1-q)} \tag{4}$$

Robertson and Lerner (1949) show that the relationship between additive heritability on the observed scale and the heritability on the liability scale satisfies

$$h_o^2 \approx h_l^2 q^2 i_q^2 [q(1-q)]^{-1}. \tag{5}$$

Substitution then results in Equation (1) with h_l^2 being replaced by h_o^2:

$$r_{g\hat{g}} = \sqrt{\frac{\lambda h_o^2}{\lambda h_o^2 + 1}}. \tag{6}$$

Therefore the dichotomous phenotype study of disease results in an identical formula for $r_{g\hat{g}}$ as the continuous phenotype provided the heritability used is that for the observed dichotomous scale.

Case Control Disease Study. The formulae will now be extended to derive the accuracy $r_{g\hat{g}}$ of a genetic risk prediction when applying a case control design to a dichotomous phenotype. The need for modification of the equations for a case control design comes from the selection of individuals from within the population to achieve a prevalence within the sample of cases and controls of w, and where typically $w = 1/2$ with equal numbers of cases and controls. Parameter values post-selection will be 'starred'. It is assumed in the following without loss of generality that cases are less common than controls in the population so $q \leq w \leq 1/2$. Two parameters in particular need to be re-estimated because of the selection practiced: (i) $S_{xx,j}^* \neq n_P\, var(x_{ij})$; and (ii) the regression of s_i on x_{ij}, $E[\hat{\pi}_j^*] \neq q i_q \beta_j$. Both these corrections can be made as shown in detail in Appendix S2.

Briefly, assuming no covariance between p_j and β_j, $E[var(x_{ij})] = var(g_i)/(n_G\, E[\beta_j^2])$. $S_{xx,j}^*$ is $n_P var^*(x_{ij})$ and so since n_G and $E[\beta_j^2]$ over loci are unaffected by the sampling of cases and controls, $E[var^*(x_{ij})] = E[var(x_{ij})]\, var^*(g_i)/var(g_i)$. Appendix S2 shows that using Normal theory $var^*(g_i) = var(g_i)(1 - h_l^2 \bar{i}(\bar{i} - x))$. Further

$E[\hat{\pi}_j^*] = w(i_q - \bar{\imath})\beta_j(1 - h_l^2\bar{\imath}(\bar{\imath} - x))^{-1}$, where x is the truncation point of a Normal distribution for upper-tail probability q, $\bar{\imath} = wi_q - (1-w)i_{(1-q)}$.

Approximating $\sigma_e^2 = 0.25$ for a binomial trait with probability ½, appropriate for equal numbers of cases and controls, gives $var(\hat{\beta}_j) = (1-w)(1 - h_l^2\bar{\imath}(\bar{\imath} - x))[w(n_P\, var(x_{ij})(i_q - \bar{\imath})^2)]^{-1}$, and substituting λ results in

$$r_{g\hat{g}}^2 = \frac{\lambda w h_l^2 (i_q - \bar{\imath})^2}{\lambda w h_l^2 (i_q - \bar{\imath})^2 + (1-w)(1 - h_l^2\bar{\imath}(\bar{\imath} - x))}. \tag{7}$$

Changing the heritability from the liability scale for a population sample to the observed scale for a population sample using Equation (5) produces

$$r_{g\hat{g}}^2 = \frac{\lambda h_o^2}{\lambda h_o^2 + q(1-q)(1 - h_l^2\bar{\imath}(\bar{\imath} - x))w^{-1}(1-w)^{-1}}. \tag{8}$$

Finally, substituting $q(1-q)(1 - h_l^2\bar{\imath}(\bar{\imath} - x))w^{-1}(1-w)^{-1} = c$, gives

$$r_{g\hat{g}}^2 = \frac{\lambda h_o^2}{\lambda h_o^2 + c}. \tag{9}$$

Thus the form of $r_{g\hat{g}}$ for a case control study shows equivalence to the $r_{g\hat{g}}$ of continuous and dichotomous phenotypes provided heritability is on the observed scale and the appropriate changes are made in c to account for the selection of cases and controls. The value of c is 1 in population studies (Equation (6)), where $w = q$ (and, hence, $\bar{\imath} = 0$). When $q < w < 1/2$, $c < 1$ and there is an increase in $r_{g\hat{g}}$ compared to a population study with the same λ.

Simulations. Stochastic computer simulations were used to test the deterministic predictions of $r_{g\hat{g}}$ for a number of parameters affecting the continuous and dichotomous phenotypes. We describe the full simulation method for the continuous trait and then state additional steps that were needed for the dichotomous phenotypes (random population sample and case control). In all scenarios (i) individuals were unrelated; (ii) loci were independent; (iii) all genetic action was additive; (iv) for simplicity, loci were assumed to be in Hardy-Weinberg equilibrium; and (v) each scenario was replicated 100 times, except for case control scenarios with $\lambda = 0.02$ where 500 replicates were run. Furthermore for initial simulations (vi) allele frequencies were sampled from a uniform distribution corresponding to a common-disease-common-variant hypothesis (CDCV) (Reich & Lander 2001); and (vii) allele effects were drawn from a reflected exponential distribution which was made symmetric about $x = 0$. Items (vi) and (vii) were modified as described below.

For the continuous phenotypes, the phenotypic variance was 1. True additive genetic values for n_P individuals were calculated as $(1 - p_j)\beta_j$ and $-p_j\beta_j$ for the minor and major alleles, respectively, for each of n_G simulated loci, and summing over loci. The value of n_G used in most scenarios was 1000 and n_P varied accordingly, depending on λ. Two exceptions were λ = 0.02, where n_G = 20,000, and the scenarios in which λ was kept constant with n_G = 100. The scale factor of the exponential distribution was chosen to obtain the required additive heritability (h_o^2). Phenotypic records were simulated by adding independent environmental terms to the true genetic effects drawn from a Normal distribution with mean zero and variance $1 - h_o^2$. Allele substitution effects ($\hat{\beta}_j$) were estimated by regression of n_P phenotypic records on genotypes one locus at a time. A second sample of individuals was then simulated with genotypes based on the same allele frequencies and effects as the original population. The estimated additive genetic values were then computed according to the following model: $\hat{g}_i = \sum_{loci,j} x'_{ij}\hat{\beta}_j$, as described above. Finally, $r_{g\hat{g}}$ was calculated as the correlation between true and estimated additive genetic values. Bias was also assessed by the slope of the regression of g_i on $\hat{g}_i$.

The continuous phenotype case was tested for robustness to different distributions of allele frequency and effects, and their correlation. The allele frequencies were also drawn from a beta (U-shape) distribution, consistent with a neutral allele model (Pritchard 2001), with parameters alpha = 0.3, and theta = 0.3. Allele effects were also sampled from a normal distribution with mean zero. The effect of having a percentage of loci with zero effects was investigated by setting a proportion of the effects to zero while keeping the overall genetic variance constant. In all cases, the scale factor for the distribution of allele effects was modified to maintain the desired h_o^2.

Further testing of the predictions was done by introducing a correlation between the heterozygosity at a locus and the squared magnitude of the allele substitution effect at a locus. This was done for a uniform distribution of allele frequencies and the reflected exponential distribution of allele effects. This was achieved empirically: if the randomly drawn frequency had heterozygosity greater than the median (i.e. $2p(1-p) > 0.375$) then the magnitude of the allele effect was drawn to be less than the median of the distribution of the magnitudes.

The simulation of a random population sample for the dichotomous disease phenotype followed the same structure as above but contained the additional step of treating the underlying continuous phenotype distribution as a liability for the disease with heritability h_l^2 on the liability scale (Robertson & Lerner 1949). Therefore, with prevalence q, the fraction q of the population with the greatest liability were considered to be affected. Therefore allele effects were estimated from the dichotomous phenotype and the accuracy, $r_{g\hat{g}}$, was calculated as the correlation between the true and estimated genetic liability for the disease estimated in an independent population sample.

Case control studies were simulated with an equal number of cases and controls (i.e. $w = 1/2$). A dichotomous disease phenotype with sample size n_P was simulated by including an additional selection step which expanded the population size to $n_P[2q_d]^{-1}$. The liabilities were constructed as for the population study of a dichotomous disease, the $n_P/2$ individuals with the greatest phenotypic liability were considered to be affected cases, and a further $n_P/2$ were randomly chosen from those remaining as control phenotypes. Allele effects were estimated as for the population studies, and the accuracy was estimated from a randomly-drawn independent population sample of size n_P.

RESULTS

Population-wide studies of continuous phenotypes. When allele effects were drawn from an exponential distribution and frequencies were from the uniform, the deterministic formula for $r_{g\hat{g}}$ was found to predict the simulated data reliably across the wide range of parameters used (Table 1). The prediction errors across all parameters studied were in the range of -1.3 to 4.0% (Table 1).

The close agreement between the predicted and achieved accuracies is also seen in Table 2 and was maintained when: (i) allele frequencies were drawn from a beta-distribution (% error -0.9 to 0.7); (ii) allele effects were drawn from a normal distribution (% error -0.8 to 5.0); (iii) exponential allele effects were mixed with varying proportions of alleles with no effects, ranging from 0 to 95% (% error 0.1 to 26.6, Table 3); (iv) λ's ranging from 0.02 to 5 were investigated (% error -20.0 to 4.0, Table 1); and (v) the genetic architecture was varied by keeping λ constant and changing n_G ($n_G = 100$, % error 0.1 to 7.6; and $n_G = 1000$, % error -0.5 to 0.0). It should be noted that the large percentage errors seen when λ =0.02 are due to low $r_{g\hat{g}}$, where the absolute difference between the expected and simulated $r_{g\hat{g}}$ was still less than 0.02. The introduced correlation between heterozygosity and squared substitution effect was tested with λ = 1 and $n_G = 1000$ using the empirical procedure described in the Materials and Methods. With an achieved correlation of -0.36 and an observed h_o^2 = 0.39, the predicted accuracy from Equation (1) was 0.53, with an error of 1.1% when compared to simulation. In conclusion, it is clear that the deterministic $r_{g\hat{g}}$ is robust to wide distributional assumptions on the joint distribution of frequency and effect of allele substitution, as predicted from the derivation.

Therefore the predictions of genome-wide accuracy shown in Figure 1 based on Equation (1) for different values of observed h^2 and λ have wide applicability. For all λ, the accuracy was most sensitive to h^2 when h^2 was low and this sensitivity was potentiated by higher numbers of phenotypes per genotype tested. The accuracies are functions of λh^2, so the required λ to achieve a given accuracy is proportional to $1/h^2$. Thus, the numbers of phenotypes per genotype need to be twice as high for half the heritability. To obtain accuracies of 0.71, corresponding to predicting half the genetic variance, $\lambda = 1/h^2$, and therefore λ must be $\geq$ 1 because $h^2 \leq 1$.

Table 1. Predicted accuracy and percentage prediction error assessed by simulation with disease prevalence = 0.1 (SE range 0.0004 - 0.0065)

		$\lambda^{a} = 0.02$		$\lambda = 0.50$		$\lambda = 1.00$		$\lambda = 5.00$	
	h^{2b}	P[c]	%error[d]	P	%error	P	%error	P	%error
C[e]	0.1	0.045	4.0	0.218	3.6	0.301	2.2	0.577	0.4
	0.5	0.100	2.1	0.447	-0.5	0.577	-0.2	0.845	-0.1
	0.9	0.133	-1.3	0.557	0.2	0.688	-0.2	0.905	-0.1
D_P[f]	0.1	0.026	-14.1	0.130	-6.6	0.182	-2.2	0.382	-1.6
	0.5	0.058	-1.1	0.281	0.6	0.382	-1.1	0.679	0.2
	0.9	0.078	-9.8	0.365	1.6	0.485	0.8	0.779	0.2
D_C[g]	0.1	0.043	-0.6	0.209	2.4	0.290	3.5	0.560	-1.9
	0.5	0.089	-4.3	0.407	3.0	0.533	0.8	0.816	-2.9
	0.9	0.112	-20.0	0.490	-0.4	0.622	-0.4	0.872	-3.3

[a] λ = number of phenotypes per number of loci
[b] h^2 = heritability (observed scale for C and D_P, liability scale for D_C)
[c] P = predicted accuracy of estimated additive genetic value
[d] % error = percentage prediction error = 100(P–accuracy from simulation)/P
[e] C = continuous phenotype
[f] D_P = dichotomous phenotype, population study
[g] D_C = dichotomous phenotype, case control study

Population-wide studies on dichotomous disease phenotypes. The form of the predicted accuracy ($r_{g\hat{g}}$) is very similar to that for a quantitative trait. Again the prediction of $r_{g\hat{g}}$ was very good (% error -14.1 to 1.6; see Table 1). The validity of the prediction resulting from Equation (6) was robust to varying disease prevalence over the range of 0.01 to 0.5 (% error -1.9 to 1.4, Table 4). The form of the prediction in Equation (6) is a function of λ and the observed additive heritability on a (0,1) scale, but this can be achieved with varied combinations of disease prevalence and underlying heritability of liability. This is shown in Table 5, which also demonstrates that, as predicted from Equation (6), $r_{g\hat{g}}$ is a function of only h_o^2 as accuracy remains constant with varied disease prevalence and h_l^2.

Table 2. The effects of different distributions of allele frequency and effects on accuracy in a continuous phenotype with observed heritability = 0.5 (SE range 0.0004 - 0.0057)

		Simulated			
λ[a]	Predicted	Beta[b]/Nrm[c]	Beta/Exp[d]	Uni[f]/Nrm	Uni/Exp
0.02	0.100	0.095	0.093	0.100	0.097
0.50	0.447	0.442	0.436	0.451	0.450
1.00	0.577	0.577	0.579	0.576	0.578
2.00	0.707	0.709	0.714	0.704	0.709
5.00	0.845	0.849	0.848	0.846	0.846
10.00	0.913	0.914	0.914	0.913	0.912

[a] λ = number of phenotypes per number of loci
[b] Beta = beta distribution (alpha = 0.3, theta = 0.3) of allele frequencies
[c] Nrm = normal distribution of allele effects
[d] Exp = exponential distribution of allele effects
[f] Uni = uniform distribution of allele frequencies

The predicted $r_{g\hat{g}}$ of population studies of continuous phenotypes and dichotomous disease phenotypes with an underlying continuous liability follow the same functional form as seen in Equation (6). Therefore, Figure 1 can be used to derive predicted $r_{g\hat{g}}$ for dichotomous phenotypes as well as continuous phenotypes. However, note that in the liability model, even if liability was fully determined genetically, the additive heritability on the observed scale will never exceed 0.64 (i.e. $4\theta(0)^2$, where $\theta(x)$ is the standardized normal density function) with the remaining genetic variation appearing non-additive. The corresponding maximum $r_{g\hat{g}}$ achievable will be reduced and this will be most serious for low λ. Even with the most favorable circumstances of $q = 1/2$ and liability $h_l^2 = 1$, the accuracy will never exceed 0.71 if $\lambda < 1.56$, and it should be expected that λ needs to be much greater than this to explain half the genetic variance. This circumstance should not be expected to change when using other disease models than the liability, since the loss of $r_{g\hat{g}}$ arises from the loss of quantitative information when moving from a continuous genetic value (however defined) to the categorical observation of affected or not.

Table 3. Accuracy for continuous phenotype when setting 0.95 of n_G[a] loci to zero (λ = 0.02 = 400n_P[b]/20,000n_G, SE range 0.0042 - 0.0057)

h_o^2[c]	0.95 of n_G zero	0.0 of n_G zero	Predicted
0.1	0.057	0.043	0.045
0.5	0.101	0.097	0.100
0.9	0.129	0.135	0.133

[a] n_G = number of loci
[b] n_P = number of phenotypes
[c] h_o^2 = observed heritability

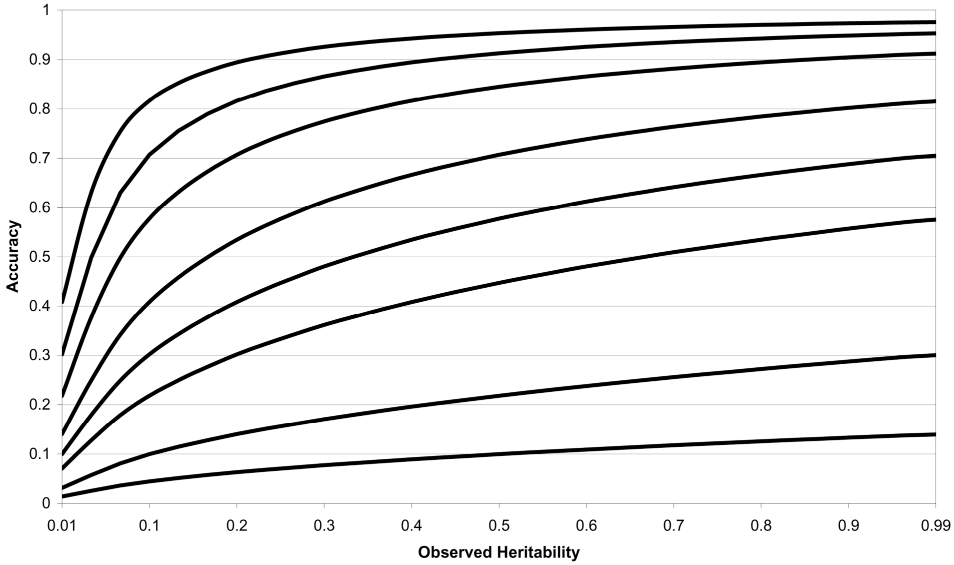


Figure 1. Predicted accuracy of estimated genetic values of a continuous phenotype. Predicted accuracy of estimated additive genetic values of a continuous phenotype as a function of observed heritability and number of phenotypes per genotype tested, λ = 0.02, 0.1, 0.5, 1, 2, 5, 10 and 20 from minimum to maximum accuracy respectively.

Case control studies of dichotomous disease phenotypes. The prediction formula for accuracy of case control studies ($r_{g\hat{g}}$) is not a simple function of λ and the observed h_o^2, but also depends on both the heritability on the liability scale and the disease prevalence, as seen from Equation (8). Therefore, comparisons require consideration of how c in Equation (9) varies. The simulations assumed $w = 1/2$, with equal numbers of cases and controls. Although, as seen in Table 1, the predictions are generally good (% error -20.0 to 3.5), where the large error deviations are again due to low λ, there is a trend towards the underestimation of $r_{g\hat{g}}$ as prevalence becomes low (Table 4).

Table 4. Accuracy for a dichotomous disease trait as prevalence varies ([a]h_l^2 = 0.5, [b]λ = 1, SE range 0.0026 - 0.0048).

Prevalence	Study Type D_P[c]		Study Type D_C[d]	
	P[e]	% Error[f]	P	% Error
0.01	0.186	-0.8	0.593	-11.1
0.03	0.271	-1.9	0.568	-6.8
0.05	0.317	0.3	0.554	-3.5
0.10	0.382	-0.6	0.533	0.6
0.20	0.444	1.4	0.511	-2.5
0.30	0.473	1.2	0.499	-0.2
0.40	0.487	-0.6	0.493	1.2
0.50	0.491	0.0	0.491	1.4

[a] h_l^2 = heritability on liability scale
[b] λ = number of phenotypes per number of loci
[c] D_P = population study of dichotomous phenotypes
[d] D_C = case control study of dichotomous phenotypes
[e] P = predicted accuracy of additive genetic values
[f] % error = percentage prediction error = 100(P–accuracy from simulation)/P

Table 5. Simulated accuracy of a population study for a dichotomous phenotype as prevalence and $h_l^{2\mathrm{a}}$ varies and $h_o^{2\mathrm{b}}$ stays constant (λ^{c} = 10, h_o^2 = 0.2, predicted accuracy = 0.816, Equation (4), SE range 0.0025 - 0.0038)

Prevalence	h_l^2	Accuracy
0.05	0.893	0.810
0.10	0.584	0.814
0.20	0.408	0.814
0.30	0.347	0.813
0.40	0.322	0.813
0.50	0.314	0.813

[a]h_l^2 = heritability on liability scale
[b]h_o^2 = heritability on observed scale
[c]λ = number of phenotypes per number of loci

The value of $r_{g\hat{g}}$ for case control studies is best illustrated by comparison with population studies of dichotomous disease traits. Figure 2 integrates this information and shows the relationship of prevalence and observed heritability in population and case control studies. Values of $r_{g\hat{g}}$ below the narrowly dashed line derived from Equation (5) are not possible under the liability model, for example, an observed additive heritability of 0.5 and a prevalence of 0.1 could not exist in the same dataset. Each contour represents an level of constant $r_{g\hat{g}}$, where the dashed lines represent a population study and the solid lines denote a case control design with $w = 1/2$. As described above the contours are vertical for population studies as, given h_o^2, the accuracy is independent of q, but for case control studies move towards lower h_o^2 as prevalence decreases. Several clear conclusions on case control studies can be drawn: (i) the overall trend of $r_{g\hat{g}}$ increasing with more phenotypes per number of genotype holds true for case control studies (Table 1); (ii) population studies and case control studies are equivalent when the prevalence is 0.5 (Figure 2) ; (iii) a case control study is always more accurate than a population study with the same number of individuals genotyped (Figure 2); (iv) for a constant h_l^2, $r_{g\hat{g}}$ increases as the disease prevalence increases in population studies, since this increases h_o^2, but in case control studies $r_{g\hat{g}}$ increases as the disease prevalence decreases because of the more intense selection induced by the less prevalent disease (Table 4).

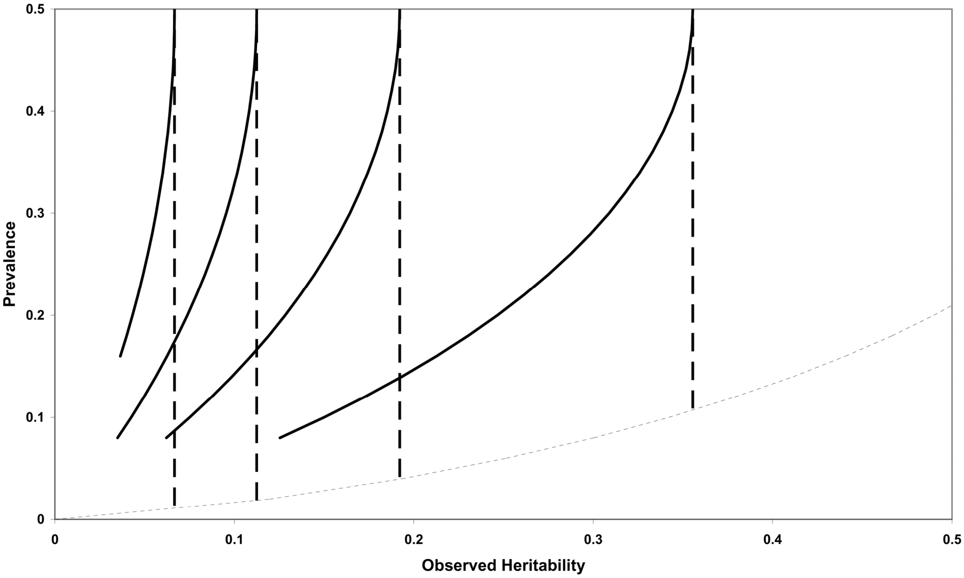


Figure 2. Predicted accuracy of estimated genetic risk from population and case control designs of a dichotomous phenotype. Contour plot of predicted accuracy for varied prevalence and additive heritability on the observed scale, in population studies (dashed vertical line) and case control studies (solid line) of dichotomous phenotypes. Each contour represents a line of constant accuracy, starting from the right 0.9, 0.8, 0.7, and 0.6. The narrowly dashed line is derived from Equation (5) with $h_l^2 = 1$, so values below this line are not possible under the liability model.

DISCUSSION

We have derived simple deterministic predictions of $r_{g\hat{g}}$ in continuous and dichotomous phenotypes using either a population or a case control study and we have shown them to be appropriately responsive to changes in disease prevalence, heritability, and the number of phenotypic records per number of risk loci to be estimated. In addition, the equations have proven robust to changes in allele effect distributions, including different fractions of loci with zero effect and differing allele frequency distributions. Population studies are also robust to covariances between the magnitude of allele effects and heterozygosity, although, in principle, this robustness does not hold for case control studies. This advance in understanding has been used to summarize the influence of critical parameters such as heritability and numbers of phenotypes and risk loci on accuracy of prediction, and also to show the degree to which case control designs can add power to studies.

The approach taken here has been to assume the potential loci affecting the trait are known, and this has an impact that is double edged. First, it allows for a clear quantification of the limitations imposed on $r_{g\hat{g}}$ by the number of phenotypes obtained, irrespective of marker densities. The information gained by doing so is of equal importance to knowing the number of markers needed for a certain $r_{g\hat{g}}$ but seems to have received less attention recently. Second, it implies that the predicted $r_{g\hat{g}}$ are upper bounds for the data obtained, since some loss of $r_{g\hat{g}}$ will occur through the use of markers which are potentially in imperfect linkage disequilibrium (LD) with loci with effect (Dekkers 2004), and the inclusion of candidate loci that may have no effect within the population.

The impact of including these loci with no true effect may be explained by two applications of our formulae. The first application assumed the loci affecting a disease trait are known and thus $r_{g\hat{g}}$ demonstrates an upper bound on the accuracy; for example, consider n_G = 1000 loci with effects greater than 0, n_P = 10,000 phenotypes and h_o^2 = 0.1, then the predicted accuracy is obtained with λ = 10, and will be 0.71. Now consider if those 1000 loci are contained with a set of n_G = 100,000 marker loci, with 99% having zero effect so that now the accuracy is obtained with λ = 0.1; our predictive equations remain valid and predict an accuracy of 0.10. From these applications of our formulae it is clear that the approach of estimating loci effects one at a time will inevitably result in low accuracies, and further, adding more marker loci with zero effects while using the same approach will reduce the expected accuracy. The low accuracies predicted accord with the empirical findings from large scale studies of human data that have recently been reported (Weedon *et al.* 2008). It is clear that alternative approaches to prediction will be needed to bridge the gap and raise accuracies towards the potential placed by the phenotype collection.

Nevertheless, potential alternative approaches are available and evidence already exists that these approaches may significantly increase predictive accuracy. One approach is to implement model selection approaches. Similarly, improvements in $r_{g\hat{g}}$ can be achieved by implementing model selection least squares procedures to identify a subset of SNP from which to predict effects (Meuwissen *et al.* 2001; Habier *et al.* 2007), or by using more complex procedures to identify a subset to set to zero (Yi & Xu 2008). Some of these studies (Meuwissen *et al.* 2001; Habier *et al.* 2007; Yi & Xu 2008) also incorporate the use of prior information within Bayesian procedures and demonstrate significant increases in accuracy over least squares. Increasing the number of markers when using priors can increase accuracy because the size of the marker subset chosen stays the same due to the prior but the portion of the genetic variance captured by the markers subset increases (Solberg *et al.* 2008). However the use of Bayesian approaches will demand reliable distributions for incorporation into models. Literature estimates informing priors on n_G and the distributions of the effects will become more widely available as GWA studies become more powerful (Hayes & Goddard 2001; Chamberlain *et al.* 2007). Full genome-wide methods (Meuwissen *et al.* 2001; Xu 2003), where genetic risk or additive genetic values are estimated in one step, using all loci simultaneously particularly if they are correlated, might be expected to approach the upper bound of $r_{g\hat{g}}$ faster than methods which impose significance thresholds and, thus, do not capture all the genetic variation. From the results presented here it may be argued that priors on the numbers of loci positively contributing to the genetic variance will be more critical than those describing the distribution of gene effects.

In this paper we have used a liability model for disease instead of the commonly used log genetic risk model and the impact of doing so is likely to be small for large datasets. For a set of h_o^2 and q, an underlying log-risk can be approximated well by a liability (Lynch & Walsh 1998; Wray *et al.* 2007) and the distribution of effects on the log-risk scale will be transformed to a distribution on the liability scale, and the predictions developed here are not dependent on the distribution of effects. However there is evidence that distinctions may be larger when q is very close to zero or one (Cox 1970).

A critical assumption of the genetic models studied was that the loci acted independently. In humans, most LD stretches for 10 to 30kb, while some linkage disequilibrium blocks may be >100kb (Ardlie *et al.* 2002). The human genome contains 3.1 billion bases (Venter *et al.* 2001) and, assuming 2000 known loci contribute to the additive genetic variance, each genomic segment between them would be 1550kb. This confirms that this model is viable in human. One could apply our formulae by interpreting n_G as the number of independent chromosome segments (i.e. haplotype blocks). The length and, thus, the number of these segments would depend on the amount of LD present in the genome. The number of such segments have been estimated directly from pair-wise LD between markers (Barrett *et al.* 2005) and closely related measures, such as the number of independent tests on the genome, have been estimated using principle component analysis (Shriner *et al.* 2008) and have been derived analytically for specific experimental designs (Risch 1991). When LD exists, either between markers and risk loci or between risk loci, the predictive efficiency of our equations will be reduced. Modeling the pattern of LD by extension of our formulae would thus be important when many loci are used, as with dense SNP marker maps, or when predicting additive genetic values in other species, such as some livestock populations where the extent of LD is large compared to human (Mcrae *et al.* 2002; Sargolzaei *et al.* 2008).

An attraction of molecular predictors of genetic risk compared to pedigree predictors is the potential to apply the predictions more widely within populations and across populations. Obtaining sufficient accuracy within populations can be achieved by the quality and size of sampling, but there are additional factors in play when transfer across populations is being considered. For example, one benefit of genome-wide prediction is that individual allele effects are estimated with a precision that is related to the molecular variation observed at the locus, $var(x_{ij})$, which determines the contribution of genetic variance when combined with the squared magnitude of effect. This benefit may break down when predictions are transferred across populations. As an illustration, consider a rare allele of large effect which will be relatively imprecisely estimated in the estimation sample, but because the contribution of the locus to total variance is small there is only a small impact upon the accuracy of further predictions within the same population. In a different population, such an allele may have a greater frequency and contribute a greater part of the genetic variance, and, consequently, the predictive accuracy will suffer. Specifically, the ability to transfer predictions will depend on $var(x_{ij})$ in each of the two populations used for estimation and application, and this in turn depends on both the allele frequency (p_j) and the degree of admixture present in the population. Furthermore, an additional risk of transferability across populations is the presence of epistasis which may differentially influence β_j.

Any directional selection present in the population is likely to introduce a covariance between the magnitude of allelic effect and heterozygosity, since selection promotes the movement of alleles of large effect quickly through intermediate frequencies, where they create large genetic variance, towards extreme frequencies. The predictions of $r_{g\hat{g}}$ developed make no assumption of the covariance, and hence are robust to such selection in the population prior to estimation in population studies. In contrast, the derivation for the case control study does assume independence of heterozygosity and magnitude (as described in Appendix S2). However, in the limited simulations carried out with such covariances in case control studies, the impact of the breaking this assumption appeared small (results not shown).

Our derivations show that $r_{g\hat{g}}$ can be reduced to very similar forms for population and case-control studies of continuous and dichotomous phenotypes (c.f. Equations (1), (6) and (9)). The common element affecting $r_{g\hat{g}}$ for all three equations is the term λh_o^2, describing the joint effect of λ, the number of phenotypic records per locus associated with the trait, and the observed heritability. Increasing either of these improves $r_{g\hat{g}}$, but the study shows that the major determinant of the trade-off between these two factors is their product. For a population study λh_o^2 is completely sufficient to determine accuracy, independent of prevalence (q) and heritability (h_l^2) of liability for a dichotomous trait, but for a case control study both q and h_l^2 retain some influence on $r_{g\hat{g}}$ over and above their impact upon h_o^2. This is because, in a case control study, the term C in Equation (9) is adjusting for the selection of the cases and controls, and the strength of selection will depend upon q, and its impact on genetic variance will depend on h_l^2.

The predictive equations are a good fit to the simulated values and we have demonstrated, by theory and simulation, that they are independent of allele frequency and effect distributions. The formulae have increased the understanding of the relative differences between predicting $r_{g\hat{g}}$ in a random sample of a population and in case control studies. The expressions for $r_{g\hat{g}}$ derived will help researchers design experiments of appropriate size to estimate genetic risk to disease.

ACKNOWLEDGEMENTS

We are grateful to Bill Hill, Piter Bijma and two anonymous reviewers for their helpful and constructive comments. HDD was supported by the SABRETRAIN Project, which is funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission. BV received support from the Scottish Executive Environment and Rural Affairs Department (SEERAD) and INIA, and JAW received funding from the Biotechnology and Biological Sciences Research Council (BBSRC).

REFERENCES

Ardlie, K. G., L. Kruglyak, and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**: 299-309.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly, 2005 Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**: 263-265.

Bijma, P., and J. A. Woolliams, 1999 Prediction of genetic contributions and generation intervals in populations with overlapping generations under selection. *Genetics* **151**: 1197-1210.

Chamberlain, A. J., H. C. McPartlan, and M. E. Goddard, 2007 The number of loci that affect milk production traits in dairy cattle. *Genetics* **177**: 1117-1123.

Cox D. R., 1970 *Analysis of Binary Data*. Methuen & Co Ltd, London.

Dekkers, J. C., 2004 Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J.Anim Sci.* **82 E-Suppl**: E313-E328.

Falconer D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman, Harlow, UK.

Goring, H. H. H., J. D. Terwilliger, and J. Blangero, 2001 Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* **69**: 1357-1369.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**: 2389-2397.

Hayes, B. J., and M. E. Goddard, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33**: 209-229.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**: 95-108.

Janssens, A. C. J. W., Y. S. Aulchenko, S. Elefante, G. J. J. M. Borsboom, E. W. Steyerberg *et al.* 2006 Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genet. in Med.* **8**: 395-400.

Lynch M., and B. Walsh, 1998 *Genetics and the analysis of quantitative traits*. Sinauer Associates Inc., Sunderland, MA.

Mcrae, A. F., J. C. Mcewan, K. G. Dodds, T. Wilson, A. M. Crawford *et al.* 2002 Linkage disequilibrium in domestic sheep. *Genetics* **160**: 1113-1122.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Pharoah, P. D. P., A. Antoniou, M. Bobrow, R. L. Zimmern, D. F. Easton *et al.* 2002 Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**: 33-36.

Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**: 124-137.

Reich, D. E., and E. S. Lander, 2001 On the allelic spectrum of human disease. *Trends in Genetics* **17**: 502-510.

Risch, N., 1991 A Note on Multiple Testing Procedures in Linkage Analysis. *American J. Hum. Gene.* **48**: 1058-1064.

Robertson, A., 1961 Inbreeding in Artificial Selection Programmes. *Genet. Res.* **2**: 189-&.

Robertson, A., and I. M. Lerner, 1949 The Heritability of All-Or-None Traits - Viability of Poultry. *Genetics* **34**: 395-411.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer, 2008 Extent of linkage disequilibrium in Holstein cattle in North America. *J.Dairy Sci.* **5**: 2106-2117.

Shriner, D., T. M. Baye, M. A. Padilla, S. Zhang, L. K. Vaughan *et al.* 2008 Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies. *Nucleic Acids Res.* **36**.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. *J. Anim. Sci.* **86**:2447-2454.

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman *et al.* 2006 Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* **38**: 879-887.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.* 2001 The sequence of the human genome. *Science* **291**: 1304-+.

Weedon, M. N., H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans *et al.* 2008 Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Gene.* **40**: 575-583.

Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661-678.

Wray, N. R., M. E. Goddard, and P. M. Visscher, 2007 Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**: 1520-1528.
Xu, S. Z., 2003 Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789-801.
Yi, N. J., and S. H. Xu, 2008 Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045-1055.

APPENDIX S1

Equation (1) has been derived assuming $\sigma_e^2 = 1$. In practice, if $\lambda \gg 1$, σ_e^2 would be derived from a multiple regression and a better approximation would be $\sigma_e^2 = (1 - h_o^2) + h_o^2(1 - r_{g\hat{g}}^2)$, where the first term is the environmental variance and in the second the genetic variance is unaccounted for. By replacing n_G in the derivation of Equation (1) with $n_G\,\sigma_e^2$ and rearranging the terms using the substitution $\lambda = n_P/n_G$ gives:

$$r_{g\hat{g}}^2 = \frac{\lambda h_o^2}{\lambda h_o^2 + (1 - h_o^2\, r_{g\hat{g}}^2)} \tag{10}$$

This gives a quadratic equation in $x = r_{g\hat{g}}^2$, and x is the solution of $x^2 - (\lambda + 1/h_o^2)x + \lambda = 0$ which allows for a second order correction of the accuracy. By re-arranging the denominator of Equation (10) so that $\lambda h_o^2 + (1 - h_o^2\, r_{g\hat{g}}^2) = (\lambda h_o^2 + 1)(1 - h_o^2\, r_{g\hat{g}}^2 (\lambda h_o^2 + 1)^{-1})$, and noting the last term may be approximated by $(1 - \lambda^{-1} r_{g\hat{g}}^4)$, the fractional magnitude of this upward correction to $r_{g\hat{g}}$ is seen to be $\approx {}^1/_2\, r_{g\hat{g}}^4\, \lambda^{-1}$. For example if λ = 10 and h^2 = 0.1, then Equation (1) gives $r_{g\hat{g}} = 0.71$ and the fractional underestimate is of the order of 0.0125 (i.e. 1.25% error). The same formula can be shown to apply for population studies of dichotomous traits and, analogously, ${}^1/_2\, c\, r_{g\hat{g}}^4\, \lambda^{-1}$ for case control studies.

APPENDIX S2

Consider the impact on $var(g_i)$ of selection of cases and controls. The selection is equivalent to setting a truncation point x of a Normal distribution on the liability scale corresponding to the proportion of affected individuals q. This requires sampling the required number of cases with liabilities $y_i > x$ and the controls with liabilities $y_i < x$. In this appendix we will not assume the cases and controls are equally sampled but consider the more general case where wn_P are cases and $(1-w)n_P$ are controls. Prior to setting the truncation point, $g_i = h^2 y_i + e_i$ where y_i is the phenotypic value and $var(e_i) = var(g_i)(1-h^2)$. With n_G not small, then we may assume constancy of regression, a property of multivariate normal distributions [31], giving $var^*(g_i) = h^4 var^*(y_i) + var(e_i)$. The $var^*(y_i)$ can be calculated directly as $var^*(y_i) = var(y_i)(1 - \bar{i}(\bar{i} - x))$, where $\bar{i} = wi_q - (1-w)i_{(1-q)})$. Therefore $var^*(g_i) = var(g_i)(1 - h^2\bar{i}(\bar{i} - x))$ and assuming no covariance between p_j and β_j gives the result $var^*(x_{ij}) = var(x_{ij})(1 - h^2\bar{i}(\bar{i} - x))$.

There are three traits to consider, the disease score s_i, liability y_i, and the allele number at locus j, x_{ij}. Prior to selection of the cases and controls $var(y_i) = 1$ and the following regression equation holds:

$$x_{ij} = \beta_j var(x_{ij}) y_i + e_{1,i}, \tag{11}$$

with $var(e_{1,i}) = var(x_{ij})(1 - \beta_j^2 var(x_{ij}))$. After selection of cases and controls it is assumed that using a normal approximation the validity of (11) remains, and:

$$s_i = w(i_q - \bar{i}) var^*(y_i)^{-1}(y_i - \bar{i}) + e_{2,i} \tag{12}$$

with $cov^*(e_{1,i}, e_{2,i}) = 0$. Therefore using (11) and (12) gives:

$$E[\hat{\pi}_j^*] = w(i_q - \bar{i})\beta_j var(x_{ij})/var^*(x_{ij}) = w(i_q - \bar{i})\beta_j/(1 - h^2\bar{i}(\bar{i} - x)). \tag{13}$$

Note when $w = q$, there is no selective sampling and $\bar{i} = qi_q - (1-q)i_{1-q} = 0$, and $E[\hat{\pi}_j^*] = qi_q\beta_j$, which is identical to $E[\hat{\pi}_j]$ in the population study.

Chapter 4

The Impact of Genetic Architecture on Genome-wide Evaluation Methods

Hans D. Daetwyler[*,§], **Ricardo Pong-Wong**[*], **Beatriz Villanueva**[**,§§],
and John A. Woolliams[*]

[*]The Roslin Institute and Royal (Dick) School Veterinary Studies, The University of Edinburgh, Roslin, UK, EH25 9PS; [§]Wageningen Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, NL; [**]Scottish Agriculture College, Edinburgh, UK, EH9 3JG; [§§]Departamento de Mejora Genética Animal, INIA (Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria), 28040 Madrid, Spain

Submitted to Genetics.

ABSTRACT

Genome-wide evaluation combines statistical methods with genomic data to predict genetic values for complex traits. Considerable uncertainty currently exists in determining which genome-wide evaluation method is the most appropriate. We hypothesise that genome-wide methods deal differently with the genetic architecture of complex traits. We compared a genomic best linear unbiased prediction method (GBLUP), and a genomic non-linear Bayesian variable selection method (BayesB) using stochastic simulation across three effective population sizes and a wide range of numbers of quantitative trait loci (N_{QTL}). GBLUP had a constant accuracy, for a given heritability and number of phenotypes, regardless of N_{QTL}. BayesB had a higher accuracy than GBLUP when N_{QTL} was low, but this advantage diminished as N_{QTL} increased and when N_{QTL} became large, GBLUP slightly outperformed BayesB. In addition, deterministic equations are extended to predict the accuracy of both methods and to estimate the number of independent chromosome segments (Me) and N_{QTL}. The predictions of accuracy and estimates of Me and N_{QTL} were generally in good agreement with results from simulated data. We conclude that GBLUP and BayesB (at high N_{QTL}) accuracy are highly dependent on Me. We propose a decision rule to choose a genome-wide method: when N_{QTL} < Me choose a variable selection method such as BayesB and when N_{QTL} > Me choose GBLUP.

INTRODUCTION

Genome-wide evaluation combines traditional approaches to the prediction of genetic values with the use of high throughput genotype data such as single nucleotide polymorphisms (SNP) (Meuwissen *et al.* 2001). The accuracy of predicted genetic values from genome-wide evaluation can be substantially higher than that of traditional methods provided that enough phenotypic records are available for estimating marker effects (Daetwyler *et al.* 2008; Goddard 2008; Hayes *et al.* 2009c). Genome-wide selection, i.e. selection based on genomic predicted genetic values, also has the potential to reduce the generational inbreeding rates in animal breeding programs (Dekkers 2007; Daetwyler *et al.* 2007). Furthermore, the application of genome-wide evaluation approaches can significantly aid our understanding of complex trait genetic architecture in a similar way as quantitative trait loci (QTL) analysis (Hayes & Goddard 2001).

The genome-wide evaluation methods suggested to date can be broadly categorized into groups according to whether or not there is assortment of the SNP by magnitude of effect or contribution to the variance. One group treats SNP homogeneously and includes variants of genomic best linear unbiased prediction (GBLUP). This group includes a form of ridge regression (Whittaker *et al.* 2000) proposed by Meuwissen et al. (2001), where individual loci effects are regressed towards zero, and the use of a realised relationship matrix in GBLUP. In the latter, the relationships may be based upon either identity-by-descent probabilities (Villanueva *et al.* 2005) or, more commonly, identity-by-state probabilities (NejatiJavaremi *et al.* 1997; Hayes *et al.* 2009c) averaged over all loci. The ridge-regression and the identity-by-state approaches have been shown to be equivalent (Habier *et al.* 2007; Goddard 2008) as the number of SNP become large. A second group provides for heterogeneity among SNP contributions, with some contributions permitted to be large whilst the remainder are small, possibly zero. This assortment is aided by Bayesian approaches placing priors on numbers assumed to have a major contribution (e.g. BayesA, BayesB (Meuwissen *et al.* 2001; Meuwissen *et al.* 2009), among others (Lee *et al.* 2008)), or some penalty based on functions of the magnitude of effect for each SNP (e.g. Lasso (Tibshirani 1996; Yi & Xu 2008)) or other smoothing metric (Long *et al.* 2007). A third group attempts to reduce dimensionality by using principal components or partial least squares (e.g. Raadsma *et al.* 2008; Solberg *et al.* 2009) to identify an informative subset contrasts among SNP genotypes. The main two methods currently used in real datasets are a linear prediction method, GBLUP, and variants of non-linear Bayesian variable selection approaches such as BayesB.

As described above, Bayesian methods can accommodate prior assumptions where loci effect variances can differ across loci, in contrast to GBLUP. In addition, BayesB (Meuwissen *et al.* 2001) is a variable selection method because it incorporates priors on the numbers of loci with effect, while others are set to zero. According to Meuwissen et al. (2001), when the number of loci is large, the proportion of them actually having an effect may become small. Thus, only fitting a subset of loci with which to estimate the effects will reduce the dimensionality of the model and can be advantageous in terms of accuracy.

In most simulated published data, the accuracy of BayesB outperformed that of GBLUP (e.g. Meuwissen *et al.* 2001; Habier *et al.* 2007; Lund *et al.* 2009). However, real data results have not consistently supported this conclusion. Two reviews of empirical results in dairy cattle to date have shown that GBLUP and BayesB result in very similar accuracies for most traits (VanRaden *et al.* 2009; Hayes *et al.* 2009a). One reason for the disagreement between simulated and real data results could be that the genetic architecture simulated is significantly different from what is found in real populations.

Studies published to date that compare methods using simulated architectures have considered only 50 or fewer QTL affecting the trait (Meuwissen *et al.* 2001; Habier *et al.* 2007; Lund *et al.* 2009), which displays a rather narrow range. We hypothesise in this paper that the relative utility of genome-wide evaluation methods depends significantly on both the genomic structure of the population and the genetic trait architecture. The accuracy of some genome-wide evaluation approaches can be accurately predicted from the number of individuals in the training population (N_P), number of loci and heritability (h^2) (Daetwyler *et al.* 2008). The derivation of the number of independent chromosome segments (Me) has further advanced accuracy predictions (Goddard 2008). However, the relative efficiency of different methods is uncertain when the number of QTL (N_{QTL}) is high. Thus, the main objective of this study was to compare a linear method, GBLUP, and a non-linear variable selection method, BayesB, using simulated data across a range of population and trait genetic architectures to further understand the mechanics of genome-wide evaluation methods. An important secondary objective was to extend deterministic prediction models to predict the accuracy of both methods and to propose methodology to estimate the Me of populations and N_{QTL} of complex traits. Theoretical models complement stochastic simulation by helping the understanding of the factors involved in genome-wide evaluation performance and, in return, stochastic simulation is used to confirm theoretical derivations.

METHODS

Theoretical development: Deterministic prediction of genome-wide accuracy can help us understand why the relative performance of methods may differ. Daetwyler et al. (2008) derived equations for predicting the accuracy of a simple least-squares genome-wide evaluation approach for continuous and dichotomous traits. The original formula for genome-wide accuracy for a continuous trait is $r_{g\hat{g}} = \sqrt{(N_P h^2)/(N_P h^2 + n_G)}$, where $r_{g\hat{g}}$ is the correlation between true and estimated additive genetic values (i.e. accuracy) and n_G is the number of independent loci (Daetwyler *et al.* 2008). The accuracy was independent of how large the subset of loci was that make non-zero contributions. Thus, it did not matter whether there were many non-zero loci effects of small magnitude or only a few non-zero loci effects of large magnitude. In Daetwyler et al. (2008), the formulae were derived by considering the regression of phenotypes on one locus at a time and naturally extend to multiple loci if they are independent. Therefore the formula will work for small numbers of dispersed loci in a genome and will tend to zero as n_G becomes large; erroneously, because loci cannot be added independently in a finite genome due to linkage. Daetwyler et al. (2008) discussed that an empirical value for Me could be used in place of n_G, because n_G was assumed independent and this dependence of the accuracy of GBLUP on a concept of predicted Me was also proposed by Goddard (2008). His derivation builds on work by Visscher et al. (2006) in which the variance of identical-by-descent sharing for full sibs was developed, and provides a prediction $Me = 2NeL/\log(4NeL)$, where L is the genome length in Morgans (Goddard 2008). Substituting Me in place of n_G results in:

$$r_{g\hat{g}} = \sqrt{\frac{N_P h^2}{N_P h^2 + Me}}, \qquad (1)$$

which predicts the accuracy of GBLUP. At no time does the argument moving from original formula of Daetwyler et al. (2008) to Equation (1) depend on the distribution of marker effects, so we come to the first testable hypothesis in this study which states that GBLUP accuracy is independent of N_{QTL}.

Our second testable hypothesis was that the accuracy of BayesB when N_{QTL} is high would tend to that of GBLUP. If our first hypothesis is confirmed, then the dependence of GBLUP on Me is an advantage at high N_{QTL}, even though N_{QTL} may be higher than Me.

Heuristically, if GBLUP delivers accuracy as if there are a Me number of QTL, the benefit from prior information that there are approximately Me (or more) QTL is unclear given equation (1). On the other hand, it is a clear disadvantage if $N_{QTL} < Me$ because GBLUP cannot adapt the model to suit the data. In contrast, BayesB is a variable selection method which attempts to determine the 'optimum dimensionality' given the data and prior information. When N_{QTL} is high this optimum is likely to be Me in both methods. Hence, the accuracy of BayesB at high N_{QTL} can be predicted in the same way as GBLUP: if $N_{QTL} < Me$, then variable selection may deliver an advantage in accuracy because choosing a subset of variables will reduce the dimensionality of the model, and substituting N_{QTL} for Me is likely to better predict the accuracy of BayesB. This results in the following equation,

$$r_{g\hat{g}} = \sqrt{\frac{N_P h^2}{N_P h^2 + min(N_{QTL}, Me)}}. \tag{2}$$

Further rearrangement of Equations (1) and (2) allows for empirical estimates of Me ($\widehat{Me}$) to be made in the following way,

$$\widehat{Me} = (N_P h^2)(1 - r^2_{g\hat{g}})/r^2_{g\hat{g}}, \tag{3}$$

where $r^2_{g\hat{g}}$ is the squared accuracy of estimates of genetic values using GBLUP or BayesB (when $N_{QTL} \geq Me$) for individuals without phenotypes. Predicting $\widehat{Me}$ with GBLUP requires molecular relatedness to be known, whereas this is not required when using BayesB. This result gives a further sub-hypothesis that the empirical Me is predicted by the formula for independent segments given by Goddard (2008). Also, if $N_{QTL} < Me$, additional information on N_{QTL} can be gathered using BayesB accuracy because it can choose a subset of loci or variables, by applying the following formula,

$$\widehat{N}_{QTL} = (N_P h^2)(1 - r^2_{g\hat{g}})/r^2_{g\hat{g}}, \tag{4}$$

where, in this case, $r^2_{g\hat{g}}$ is the squared accuracy of genetic values resulting from BayesB. Hence, additional insight into complex traits can be gained by combining genome-wide evaluation and deterministic prediction.

Simulations

Our study consisted of three main steps. First, populations of individuals were simulated to be in mutation drift equilibrium. Second, effects were assigned to a number of QTL that were randomly selected from the whole set of segregating loci and true genetic values and phenotypes were generated for each individual. The third step consisted of the genetic evaluations of the individuals generated with both GBLUP and BayesB.

Populations and Genome: Populations in mutation drift equilibrium were simulated by random mating individuals for many generations with recombination and mutation. The number of male and female parents was ½ Ne across generations. One male and one female offspring were produced per mating. The number of generations needed to attain mutation drift equilibrium was approximately $5Ne$. Thus, a total of 1,000, 5,000 and 10,000 generations were simulated until linkage disequilibrium and heterozygosity values were stable for Ne = 200, Ne = 1,000 and Ne = 2,000, respectively. In the final generation, a set of training individuals (of variable size) in which the loci effects were to be estimated was generated by random mating. Using the same parents with dams re-randomised, a set of validation individuals of size equal to the training set was produced whose genetic values were to be predicted. This limited the impact of relationships on the accuracy of predicted genetic values as the validation set was not the offspring of the training set but it ensured that both sets were from the same gene pool. In scenarios where the size of the training sets (N_P) was larger than Ne, population size was increased by increasing the number of offspring per mating in the final generation.

The total genome size was 10 Morgans and 10 chromosomes of 1Morgan each were simulated. In generation zero all individuals were completely homozygous for the same allele and mutations were applied at a rate of 2.5 $*10^{-5}$ per locus per meiosis in the following generations. Mutations switched allele one to two and vice versa. The number of mutations per chromosome was sampled from a Poisson distribution with mean corresponding to the product of the number of loci per chromosome and the mutation rate and they were then randomly distributed across the chromosome. Similarly, recombinations per chromosome were sampled from a Poisson distribution with a mean of one per M and were then randomly placed along the chromosome. Linkage disequilibrium

(LD, R^2) statistics (Hill & Robertson 1968) between adjacent segregating loci were averaged among all pairs exceeding a minor allele frequency of 0.05 and matched expected R^2 values (Sved 1971; Tenesa *et al.* 2007). When all segregating loci are included, achieved LD will not match expected R^2 closely (Hudson 1983). Heterozygosity of segregating loci at mutation drift equilibrium followed expectations, $He = (4\ Ne\ u)[4\ Ne\ u + 1]^{-1}$, where u is the mutation rate (Sved 1971). Allele frequency distributions were found to follow a U-shaped distribution.

Table 1. Number of QTL simulated for each proportion of independent chromosome segments (Me) for three values of effective population size (Ne).

Ne	0.03 *Me*	0.05 *Me*	0.15 *Me*	0.3 *Me*	0.5 *Me*	0.75 *Me*	1 *Me*
200	12	24	73	146	243	365	486
1000	51	101	303	606	1010	1515	2020
2000	95	189	567	1134	1890	2835	3780

The number of loci at the start of the simulation (generation zero) required several considerations concerned with obtaining appropriate number of segregating loci (N_L) and N_{QTL} in the final generation. The realised relationship matrix used in GBLUP can be singular if N_L is less than the number of individuals in the matrix (VanRaden 2008), preventing the inversion needed to compute solutions. Thus, N_L at mutation drift equilibrium was made larger than the maximum sum of training and validation individuals to be used, and a similar N_L was used across all scenarios to reduce variability. However, as Ne increased the proportion of segregating loci (N_L) in the last generation also increased and for Ne = 200, Ne = 1,000, and Ne = 2,000, approximately 0.04, 0.28 and 0.52 of initial loci were segregating at mutation drift equilibrium, respectively. This required the adjustment of the number of initial loci. The achieved value of N_L at equilibrium for each scenario is shown in Table 2. To obtain $N_{QTL,}$ Me in a random mating population, as derived by Goddard (2008) was used as a guide to allow comparisons to be made across Ne. The following N_{QTL} scenarios were simulated: 0.03, 0.05, 0.15, 0.30, 0.50, 0.75 and $1Me$. Table 1 outlines the corresponding N_{QTL} for these proportions of Me for the three Ne. Note that throughout this study our use of the terms 'low' or 'high' N_{QTL} may refer to different actual N_{QTL} across the three Ne because N_{QTL} was scaled to be proportional to Me (Table 1).

Table 2. Parameter values for the simulated scenarios, where Ne = effective population size; N_P = number of individuals in the training set; h^2 = heritability; Prior = prior used in BayesB; Me= number of independent chromosome segments; N_L = number of segregating loci in last generation (SE < 9.5).

Scenario	Ne	N_P	h^2	Prior	Me	N_L
1	200	200	0.3	Exact	445	4576
2	200	1000	0.3	Exact	445	4646
3	1000	1000	0.5	Exact	1890	4696
4	1000	1000	0.3	Exact	1890	4696
5	1000	1000	0.1	Exact	1890	4696
6	1000	500	0.3	Exact	1890	4599
7	1000	2000	0.3	Exact	1890	4721
8	1000	1000	0.3	51 QTL	1890	4696
9	2000	2000	0.3	Exact	3774	4632

The desired N_{QTL} were randomly chosen from N_L. True allele substitution effects (β_j) were sampled from $N(0,1)$. True genetic values for $2N_P$ (i.e. training and validation set) individuals were calculated as $(1-p_j)\beta_j$ and $-p_j\beta_j$ (where p_j is the major allele frequency at locus j) for the minor and major alleles, respectively, for each QTL. These were summed over N_{QTL} and scaled to have the variance of h^2 (Falconer & Mackay 1996). Phenotypic records were simulated for N_P (training set) animals by adding independent environmental terms drawn from $N(0,1-h^2)$ to true genetic values. The accuracy of both genome-wide evaluation methods was computed as the correlation between true and estimated genetic values.

GBLUP analysis: The evaluation with GBLUP applied the following model which was fit in ASReml (Gilmour *et al.* 1995): $\mathbf{y} = \mu\mathbf{1} + \mathbf{Z}\mathbf{a} + \mathbf{e}$, where $\mathbf{y}$ is the vector of phenotypic values, μ is the population mean, $\mathbf{Z}$ is an incidence matrix for random individual effects, $\mathbf{a}$ is a vector of random individual additive genetic values and $\mathbf{e}$ is the residual. Random effects $\mathbf{a}$ and $\mathbf{e}$ were assumed normally distributed as $N(0, \mathbf{G}\sigma_a^2)$ and $N(0, \mathbf{I}\sigma_e^2)$, respectively, where $\mathbf{G}$ was the realised relationship matrix computed using the N_L loci. In $\mathbf{G}$, the relationship between a pair of individuals was based on identical-by-state probabilities and included all training individuals with phenotypes and validation individuals without phenotypes. The total allelic relationship at a locus between a pair of individuals was calculated as $0.5\sum_{i=1}^{2}\sum_{j=1}^{2}\delta_{ij}$, where δ_{ij} is 1 if allele i in the first individual is identical to allele j in the second individual and 0 otherwise. Averaging over loci as $0.5\sum_{i=1}^{2}\sum_{j=1}^{2}\delta_{ij}[N_L]^{-1}$ yields the numerator relationship between all individual pairs required for $\mathbf{G}$ (NejatiJavaremi *et al.* 1997; Hayes *et al.* 2009c). Genetic values were then predicted by solving $(\hat{\mathbf{a}}) = [\mathbf{Z}'\mathbf{1}\ \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\sigma_a^2]^{-1}(\mathbf{Z}'\mathbf{y})$, where σ_a^2 was the estimated additive genetic variance.

BayesB analysis: We implemented a variant of the original BayesB (Meuwissen *et al.* 2001). The model applied was $\mathbf{y} = \mu\mathbf{1} + \sum_{N_L=1}^{N_L}\mathbf{X}'\hat{\beta} + \mathbf{e}$, where $\mathbf{X}$ was an incidence matrix relating the number of favourable alleles an individual carries at a locus to the estimated loci effect $\hat{\beta}$ (i.e. 11 = -1, 12 = 0, 22 = 1). A flat prior was used for $\mathbf{e}$ and for $\hat{\beta}$ the prior was assumed to follow $N(0, \sigma_a^2)$. BayesB can account for the probability that a proportion of loci have no effect (π). The prior applied for the proportion of non-zero effect loci was exact, so $1 - \pi = N_{QTL}[N_L]^{-1}$, except for scenario 8 where 51 QTL was used regardless of the actual N_{QTL}. A weak prior for the genetic variance was chosen from a scaled inverted chi square distribution and it was found to have an inconsequential influence on results. Our implementation of sampling loci variances was slightly different from Meuwissen et al. (2001). Instead of sampling a variance at each locus, we calculated a variance based on all loci at the end of each iteration and then drew locus variances from that distribution in the subsequent iteration. The length of the Gibbs chain was 105,000 iterations and the first 5,000 iterations were discarded as warm up. Estimates at every 20th iteration were stored as a sample resulting in a total of 5,000 samples. Autocorrelations of sampled effects were found to be close to zero which showed that they were almost independent (Wang *et al.* 1994). This allowed for shortening of the chain length to 45,000 iterations (2000 samples) for scenarios with N_P = 2,000 to reduce running time.

RESULTS

GBLUP accuracy: A clear trend is apparent in GBLUP. The accuracy of GBLUP for a given set of values for N_P and h^2 stays constant regardless of N_{QTL} and this constant accuracy is observed across all scenarios simulated (Figure 1 and Table 3). This confirmed our first hypothesis. The constant accuracy results from the unique Me of a population which, in turn, depends on Ne and L in a random mating population (Goddard 2008). The plateau of GBLUP accuracy increased when more phenotypic records were used in the estimation of genetic values and when h^2 increased (Figure 1).

BayesB accuracy: In contrast to GBLUP, with BayesB the accuracy was highest at low N_{QTL} and then decreased as N_{QTL} increased (Figure 1 and Table 3). This can be explained by the increase in dimensionality in the analysis as the number of variables with effects to be estimated increases. Once N_{QTL} was high BayesB reaches a plateau where the accuracy does not decrease anymore despite increasing N_{QTL}. This plateau is observed in all BayesB scenarios and the value of the accuracy at this plateau depended on Ne, h^2 and N_P (Tables 3 and 4). The plateau decreased when Ne increased. An increase in h^2 and N_P influenced the accuracy in two ways, firstly, it raised the overall accuracy in all N_{QTL} scenarios. Secondly, it slightly shifted the onset of the accuracy plateau to higher N_{QTL}.

Table 3. Accuracy of GBLUP and BayesB (exact priors) for different effective population sizes (Ne), numbers of QTL expressed as proportions of Me, numbers of individuals in the training set (N_P) when the heritability is 0.3. SE < 0.023 in all scenarios.

Method	Ne	N_P	0.03Me	0.05Me	0.15Me	0.3Me	0.50Me	0.75Me	1 Me
	200	200	0.405	0.450	0.429	0.414	0.444	0.416	0.398
GBLUP	1000	1000	0.505	0.501	0.508	0.502	0.507	0.501	0.511
	2000	2000	0.575	0.579	0.571	0.568	0.571	0.571	0.568
	200	200	0.739	0.649	0.463	0.400	0.398	0.365	0.344
BayesB	1000	1000	0.865	0.772	0.601	0.516	0.480	0.451	0.445
	2000	2000	0.886	0.812	0.646	0.573	0.544	0.522	0.506

Under scenario 8 (Table 2) an incorrect low prior was applied and results demonstrate the need to use accurate priors for N_{QTL} in Bayesian analyses. The use of incorrect low priors for $1 - \pi$ yielded a lower accuracy than exact priors as seen in Figure 2. The gap between the accuracy of exact and low priors increased as N_{QTL} increased because proportion of the genetic variance explained by the low prior became smaller.

Comparison of GBLUP and BayesB: The comparison of GBLUP and BayesB leads to several key observations. BayesB always performed better than GBLUP at low N_{QTL}. However, as N_{QTL} increased, the difference between the two methods became smaller and eventually both approaches achieved very similar accuracy. The N_{QTL} at which this equivalence occurred was increased with increasing Ne, N_P and h^2 (Figure 1, Tables 3 and 4). Once N_{QTL} increased past the equivalence point, BayesB had a slightly lower accuracy than GBLUP and settled at a constant accuracy (Table 3). The difference between GBLUP and BayesB at high N_{QTL} decreased when N_P was increased.

Table 4. Accuracy of GBLUP and BayesB (exact priors) in training (T) and validation (V) individuals for different effective population sizes (Ne), numbers of QTL (N_{QTL}) expressed as

proportions of Me and numbers of individuals in the training set (N_P) when the heritability is 0.3. SE < 0.023 in all scenarios.

			$Ne = 200$		$Ne = 1000$		
	N_{QTL}	Pop.	$N_P = 200$	$N_P = 1000$	$N_P = 500$	$N_P = 1000$	$N_P = 2000$
BayesB	0.03 Me	T	0.794	0.958	0.801	0.883	0.933
		V	0.739	0.952	0.757	0.865	0.924
	0.05 Me	V	0.649	0.905	0.602	0.772	0.870
	0.15 Me	V	0.463	0.803	0.421	0.601	0.744
	0.30 Me	V	0.400	0.709	0.371	0.516	0.656
	0.50 Me	T	0.599	0.778	0.583	0.657	0.741
		V	0.398	0.654	0.373	0.480	0.613
	1 Me	V	0.344	0.591	0.342	0.445	0.567
GBLUP	All	T	0.656	0.771	0.625	0.682	0.747
		V	0.444	0.622	0.407	0.507	0.612

In Figure 2, the maximum x-value of N_{QTL} plotted is equal to the predicted Me from Goddard (2008) and it is clear that BayesB approaches the plateau and approximates the equivalence to GBLUP well below this value. A first inspection therefore suggests that the second hypothesis does not hold. However the argument for the second hypothesis is based upon the empirical $\widehat{Me}$, and the values $\widehat{Me}$ may be calculated for h^2 = 0.1, 0.3 and 0.5 by averaging over the values of N_{QTL} and using Equation (3) gives values of 890, 900 and 700. In this context, hypothesis two is shown to be broadly valid, in that superiority of BayesB over GBLUP disappears when N_{QTL} approaches $\widehat{Me}$, although there is trend for this superiority to disappear slightly sooner than Me. These observations held for other scenarios (excluding the use of the incorrect prior). The comparison between Me and $\widehat{Me}$ is addressed in more detail below.

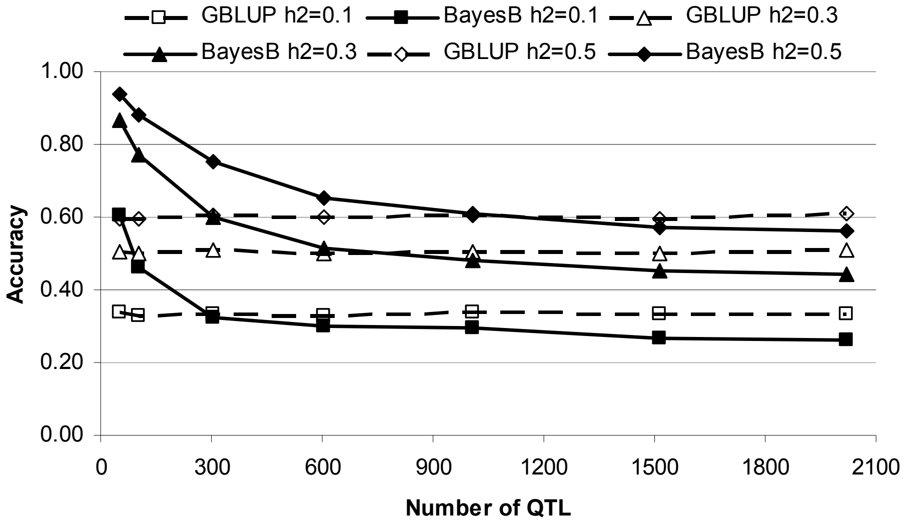


Figure 1. Accuracy of GBLUP and BayesB (exact priors) in validation individuals for different numbers of QTL and heritabilities (h^2) when the effective population size is 1,000 and the number of individuals in the training set is 1,000. SE < 0.018 in all scenarios.

Decay in accuracy: A non-trivial issue in genome-wide evaluation is the decay in accuracy observed when effects are estimated in a population sample (training set) with individuals genotyped and phenotyped and then these estimates are used to obtain genetic values in another population sample (validation set) with individuals only genotyped. Habier et al. (2007) have shown that this decay in accuracy is, in part, due to a decay in genetic relationships between individuals and was greater in GBLUP than in BayesB when 50 QTL were simulated. A similar trend can be seen in Table 4 where the decay in accuracy between training and validation individuals was also much greater in GBLUP than that of BayesB at low N_{QTL}. However, this trend diminished as N_{QTL} increased and the decay of accuracy reached similar levels in both methods at high N_{QTL}.

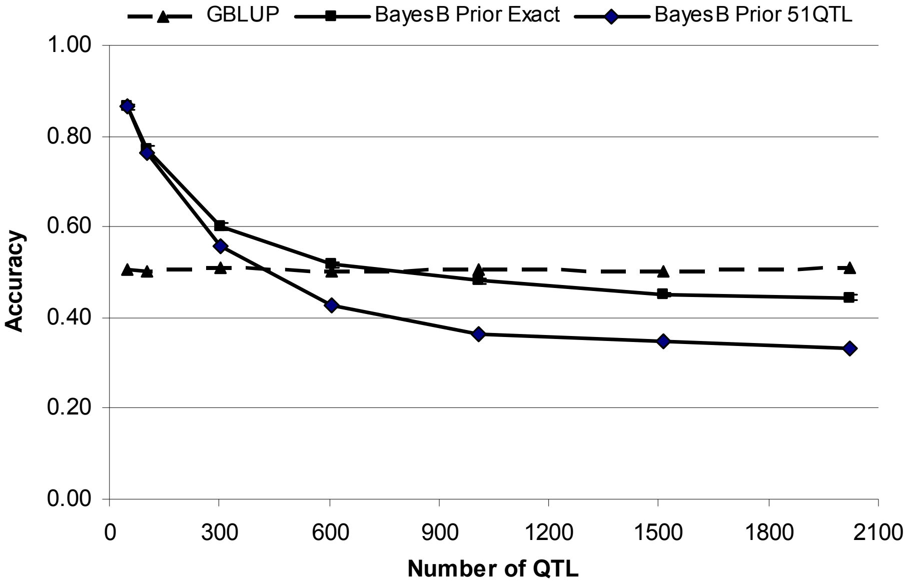


Figure 2. Accuracy of BayesB in validation individuals with exact priors (BayesB Prior Exact) and a low prior of 51 QTL regardless of the actual number of QTL (BayesB Prior 51 QTL = 0.05Me) when the effective population size is 1,000, the number of individuals in the training set is 1,000, and heritability is 0.3. Accuracy of GBLUP is included for reference. SE < 0.009 in all scenarios.

Predictions of accuracy: Figure 3 shows the accuracies of GBLUP and BayesB predicted with Equations (1) and (2), respectively and the accuracies from simulations in the validation set. Predictions of GBLUP and BayesB (at high N_{QTL}) accuracy were generally accurate. The accuracy of the predictions were highly dependent on Me. In BayesB, the drop in accuracy as N_{QTL} increased was predicted well. Equation (2) tended to over-predict BayesB accuracy, particularly in scenarios with low proportions of Me (using Goddard (Goddard 2008)) and low h^2 and N_P. Overall, predictions became more accurate as N_P increased (Figure 3).

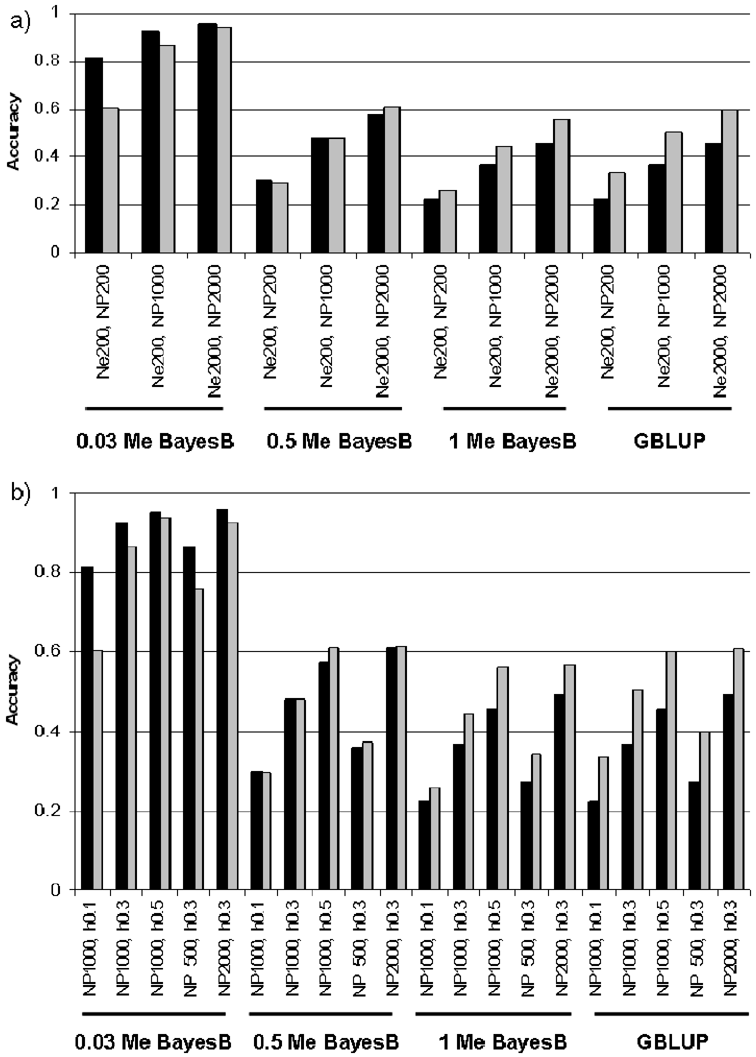


Figure 3. Predicted (black bars) and simulated (grey bars) accuracy of GBLUP and BayesB for a) a heritability (h^2) of 0.3 and varying effective population size (Ne) and number of individuals in the training set (N_P) and for b) Ne = 1000 and varying h^2 and N_P. Different numbers of QTL expressed as proportions of Me were considered for BayesB.

Table 5. Estimated ($\widehat{Me}$) and predicted (Me and Me_H) number of independent chromosome segments. Estimates were obtained from Equation (3) with mean squared accuracy of GBLUP or BayesB from 50 replicates of simulated data (±SE). The number of QTL is $1Me$ and predictions are $Me = 2NeL/log(4NeL)$ as in Goddard (2008) and $Me_H = 2NeL$ as in Hayes et al. (2009c).

			$\widehat{Me}$			
Ne	N_P	h^2	GBLUP	BayesB	Me	Me_H
200	200	0.3	294 ±29	579 ±73	445	2,000
200	1000	0.3	487 ±17	584 ±25	445	2,000
1000	1000	0.5	883 ±24	1103 ±27	1890	10,000
1000	1000	0.3	890 ±28	1243 ±33	1890	10,000
1000	1000	0.1	822 ±41	1551 ±111	1890	10,000
1000	500	0.3	803 ±39	1187 ±48	1890	10,000
1000	2000	0.3	1014 ±21	1280 ±26	1890	10,000
2000	2000	0.3	1253 ±24	1769 ±42	3774	20,000

Empirical $\widehat{Me}$ and $\hat{N}_{QTL}$: The accuracy of GBLUP or BayesB (when $N_{QTL} = Me$) is required for estimating Me using Equation (3). When GBLUP accuracy was used, we averaged the accuracy across all N_{QTL} scenarios simulated for a given set of values for h^2 and N_P. We also used the BayesB accuracy when $N_{QTL} = Me$. This was done for each population replicate to obtain a standard error. It is a sub-hypothesis that Me as predicted by Goddard (2008) approximates $\widehat{Me}$. Empirical estimates of Me using GBLUP were always lower than those using BayesB (Table 5) due to the higher GBLUP accuracy when N_{QTL} is high. The estimates using BayesB accuracy were more variable than GBLUP as shown by the larger SE of BayesB in Table 5. A general trend was apparent showing that $\widehat{Me}$ increased as N_P increased which suggests that $\hat{M}e$ has not reached its true value, however the change in $\widehat{Me}$ is small in relation to the difference from Me. Furthermore,

$\widehat{Me}$ does not increase linearly with N_P and this may indicate that it may be approaching asymptotic values.

The number of QTL controlling the trait (N_{QTL}) was estimated using Equation (4) with reliability values from BayesB when $N_{QTL} < Me$. As shown in Figure 4 for Scenario 7, the estimated N_{QTL} do follow the actual N_{QTL} well and are predictive of the trend. Empirical $\widehat{N}_{QTL}$ were better estimated with higher N_P following the same trends as deterministic predictions of BayesB accuracy. Note that incorrect priors will reduce $\widehat{N}_{QTL}$ accuracy.

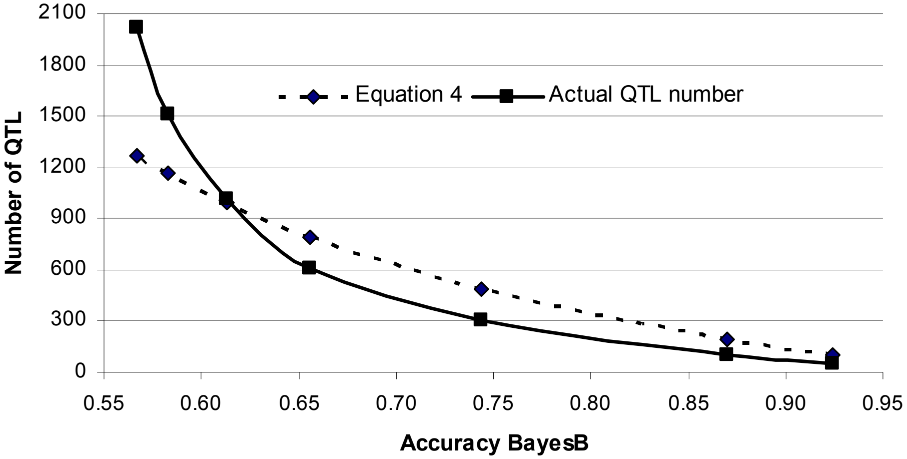


Figure 4. Actual number of QTL simulated and number of QTL predicted with Equation (4) using BayesB accuracy when the effective population size is 1,000 and the number of individuals in the training set is 2,000 and the heritability is 0.3.

DISCUSSION

We have compared GBLUP and BayesB at various population and trait genetic architectures and at various N_P. We demonstrated that GBLUP had a constant accuracy, for a given N_P and h^2, regardless of N_{QTL}. The accuracy of BayesB was greatest at low N_{QTL}, decreased with increasing N_{QTL} and eventually reached a lower accuracy plateau below which the accuracy did not fall even when N_{QTL} was further increased. BayesB has an advantage over GBLUP at low N_{QTL}, but this advantage decreased as N_{QTL} increased and it finally diminished completely or, in some cases, the advantage switched to GBLUP depending on Ne and N_P. The value of near equivalence was related to the empirical number of independent segments estimated from the GBLUP accuracy, $\widehat{Me}$, which was less than the theoretical prediction of Me provided by Goddard (2008). It is clear from this study that quantifying the superiority of GBLUP over BayesB or vice versa depends upon three sets of attributes: the population genome structure (e.g. Ne, L), the trait genetic architecture (e.g. N_{QTL}, h^2) and the experimental design (e.g. N_P). Superiority is, therefore, not a property of the method and general statements to that effect should be avoided. Furthermore, we have proposed and tested equations for the prediction of GBLUP and BayesB accuracy and the estimation of $\widehat{Me}$ and $\widehat{N}_{QTL}$. Our predictions follow achieved GBLUP and BayesB accuracy well. Empirical $\widehat{Me}$ values seem to be approaching an asymptote with increasing N_P and our estimates of $\widehat{N}_{QTL}$ follow the trend of true N_{QTL}.

The constant accuracy of GBLUP, for a given h^2 and N_P, confirmed our first hypothesis and clearly shows that this accuracy depends crucially on genomic properties, and not on properties of the trait, and summarised in the concept of Me, the number of independent segments. In turn, Me will depend on Ne and L, which can be viewed, in the short term, as constants in a random mating population (Goddard 2008). In a wider sense, Me and the more commonly known haplotype blocks are comparable measures resulting from population history. Haplotype blocks are segments of the genome within which haplotype diversity is low, bounded by areas where evidence for historical recombination exists (e.g. Goldstein 2001; Gabriel *et al.* 2002; Frazer *et al.* 2007). While haplotype blocks are an empirical measure, Me is theoretically derived and results from variation in relationship between relatives and from variation in linkage disequilibrium across the genome (Goddard 2008). Both the number of haplotype blocks and Me increase with increasing Ne and in close relatives haplotype blocks will be long and Me will be small (Hayes *et al.* 2009c). It should be noted that the dependence of GBLUP on Me shown in this study does not support the conclusion that GBLUP assumes an infinitesimal model in which there are a very large number of genes each contributing a small portion to the genetic variance. In fact, GBLUP is indifferent to N_{QTL}, unless N_{QTL} is very small (unpublished results), as demonstrated in this study.

While it is clear that in GBLUP the accuracy is a due to Me regardless of N_{QTL}, in BayesB the accuracy depends on the interplay of two features of genetic architecture, namely N_{QTL} and Me. Our results follow our second hypothesis that the behaviour of BayesB accuracy at high N_{QTL} is similar to that of GBLUP. The accuracy of BayesB declines as N_{QTL} increases due to increasing model dimensionality. Eventually, BayesB reaches a N_{QTL} at which there is no advantage to variable selection and the accuracy becomes very similar to GBLUP accuracy. The point at which this occurs approaches Me with increasing N_P. Therefore, we propose that the accuracy of BayesB at high N_{QTL} is also dependent on Me just like in GBLUP. This is also supported by the accuracy plateau being observed across at similar proportions of Me and that near equivalence is approximated closely by $N_{QTL} = \widehat{Me},$ where $\widehat{Me}$ is the empirical estimate of Me obtained from GBLUP analysis of the data. Therefore, the plateau is not function of actual N_{QTL} but of Me. Another argument for Me to be a major determinant in BayesB accuracy at high N_{QTL} is that it can be accurately predicted with Equation (3).

The difference in accuracy at high N_{QTL} between GBLUP and BayesB may be explained by the way both methods process and model information. With GBLUP, each independent segment is estimated irrespective of whether it has an effect or not whereas with BayesB, an additional step is involved in which for each locus it is estimated if the locus has an effect or not (i.e. determination of π). This comes at a cost in efficiency because there is an error associated with this process. We can view this as a balance of two factors. The first is the rise in BayesB accuracy achieved by choosing a near correct subset of loci with effect when $N_{QTL} < Me$ and the second is the error associated with determining π which depends on N_L. When $N_{QTL} \geq Me$ the advantage of choosing a subset diminishes (heuristically, it is likely that each independent segment contains QTL) and, with BayesB, the balance shifts to the second factor. Thus, under this scenario, GBLUP performs slightly better than BayesB. This argument is further supported by the decreasing difference in accuracy between GBLUP and the BayesB accuracy plateau at high N_{QTL} when N_P is increased, because with more information the error associated with π decreases in BayesB. The findings that both GBLUP and BayesB depend significantly on Me are given more weight by the fact that the accuracy of both methods can be predicted with Equations (1) and (2), respectively. The predictions were generally accurate but limitations have also been highlighted, especially in predicting BayesB accuracy. Extensions to the formulae may be needed to predict BayesB more accurately at low proportions of Me when h^2 or N_P are also low, and there is also a need to review whether Me as formulated by Goddard (2008) is a good predictor of $\widehat{Me}$. However, being able to predict the trend in BayesB accuracy is a significant step forward (Figure 3). One of the assumptions in the original derivation (Daetwyler *et al.* 2008) was that all of the genetic variance was tagged by the loci used in the analysis. This represents a complication when applying our equations to predict the accuracy using a commercially available SNP chip, because the current chips are likely to miss a portion of the genetic variance. This is due to several reasons. Firstly, it is likely that the number of SNP on current chips is not high enough to tag all the genetic variance and variation not associated with SNP (e.g. copy number variation) will also be missed. Secondly, SNP with higher than average heterozygosity are selected for developing the chips and therefore loci with low minor allele frequency are proportionally underrepresented (i.e. ascertainment bias). The problem is amplified by researchers discarding low minor allele frequency SNP as part of quality control measures. This leads to substantial difficulties when estimating effects for rare QTL alleles of which there are likely a large number. The result of this missing genetic variance in the analysis of real populations is that our deterministic equations are likely to over-predict the accuracy in both methods. Extensions to the formulae are needed which take into account this missing genetic variance.

The fact that our equations account for the entire genetic variance will, however, be a clear advantage as the scientific community moves towards the analysis of sequence data for which our formulae are appropriate in their current form. In sequence data analysis, all base-pairs are included and therefore no rare alleles would be missing. Thus, all the genetic variance is contained in the sequence and the prediction does not rely on capturing LD with the true mutation. Analysis of sequence data will, however, be challenging because the number of variables (i.e. all base-pairs) will be extremely large underlining a further need to develop rapid variable selection methods (Meuwissen *et al.* 2009).

Additional insight into complex traits can be gained by combining genome-wide evaluation and deterministic prediction. We have shown that Me can be estimated with Equation (3) if the accuracy of GBLUP or BayesB is known. Two theoretical values for Me have been proposed to date, $Me = 2NeL/log(4NeL)$ (Goddard 2008) and $2NeL$ (Hayes *et al.* 2009c). In addition, $Me = 2NeL/log(4NeL)$ could be calculated per chromosome and summed over chromosomes, which yields a larger value than considering the full genome length at once (Meuwissen 2009). Our estimates of Me ($\widehat{Me}$), even though still increasing with increasing N_P, remain lower than either but were of the right order of magnitude when using Goddard (2008) rather than $2NeL$ (Table 5). In real data using $2NeL$ in Equation (1) appears to predict GBLUP accuracy well (Hayes *et al.* 2009b). However, this may be due to the artifacts of SNP arrays missing a significant proportion of the genetic variance (as described above) leading to lower accuracies and an upward bias in estimating Me. Once more of the genetic variance is captured with new technology we would expect that estimates of $\widehat{Me}$ from real data would likely tend towards the derivation of Goddard (2008), or possibly lower. In addition to $\widehat{Me}$, N_{QTL} can be estimated with BayesB accuracy if $N_{QTL} < Me$. As Figure 4 shows, this can be a coarse measure of N_{QTL}, because small changes in accuracy can cause relatively large fluctuations in $\widehat{N}_{QTL}$. A complication in estimating N_{QTL} in data with markers in LD with QTL is that several SNP may be in LD with a particular QTL and this could lead to overestimates of N_{QTL}. In addition, BayesB requires knowledge of the true N_{QTL} in its prior. Nevertheless, estimates of N_{QTL} could aid investigations into complex trait architectures, perhaps through examining the correspondence between the assumed prior on N_{QTL} and the resulting estimate and they represent an additional piece of information for choosing a genome-wide evaluation method.

The trends observed in this study are supported by experiences in real data. Results in dairy cattle genotyped with a 50K SNP chip show that GBLUP and BayesB lead to very similar accuracies in most traits (VanRaden *et al.* 2009; Hayes *et al.* 2009a). VanRaden et al. (2009) report correlations between linear and nonlinear methods of >0.99 in a vast majority of traits. This is consistent with the findings of this study and it suggests that in real animal populations quantitative traits are controlled by a large number of QTL and for most traits $N_{QTL} \geq Me$. There are of course always exceptions to the rule and, for example, in dairy cattle BayesB performed better than GBLUP in milk fat content (VanRaden *et al.* 2009). This is likely due to a significant portion of the variation being explained by genes of large effect such as DGAT (Grisart *et al.* 2004). Hence, in this trait it is likely that $N_{QTL} < Me$, or that a relatively small number of QTL explain the majority of the genetic variance in the trait. We have demonstrated in this study the reasons for the methods' relative performance in different traits are based on population and trait genetic architecture.

The principles established in this study should be transferable to other populations as the trends have been confirmed across three different Ne. In our view, investigators need to gather evidence to answer two questions. Firstly, what is the population's Me and, secondly, how many N_{QTL} are likely contributing to the genetic variance in a particular trait. When $N_{QTL} \geq Me$ GBLUP will result in higher accuracy than BayesB, but when $N_{QTL} < Me$ BayesB will outperform GBLUP. Therefore, in populations where Ne is very low, for example in Jersey cattle $Ne \approx 36$ (Weigel 2001), it is unlikely that applying BayesB will be advantageous because Me is also small. In contrast, in humans $Ne \approx 10,000$ (Erlich *et al.* 1996) and the resulting Me of 42,864, using Goddard's equation (2008), is very large. Thus, the N_{QTL} affecting a particular disease trait would need to be very numerous for GBLUP to be advantageous in a genome-wide evaluation of genetic risk. Reducing the dimensionality of the data by applying a variable selection approach, such as BayesB, would be more promising. While the computational requirement of a full BayesB approach would be extensive when a large number of SNP are used, new approaches based on the similar principles as BayesB have been proposed which are less computationally demanding (Meuwissen *et al.* 2009).

In summary, we have demonstrated the relative performance of a linear (GBLUP) and a non-linear variable selection (BayesB) genome-wide evaluation method under different genetic population and trait architectures, and have proposed equations for the deterministic prediction of the accuracy of both methods. Furthermore we have provided guidance on which method is appropriate for a certain population via the introduction of a decision rule: when $N_{QTL} < Me$ choose a variable selection method such as BayesB and when $N_{QTL} \geq Me$ choose GBLUP. The methodology presented here to estimate $\widehat{Me}$ and $\widehat{N}_{QTL}$ will aid in unraveling the complexity of quantitative traits.

ACKNOWLEDGEMENTS

We are grateful to Piter Bijma and Jesús Fernández for their helpful and constructive comments. HDD was supported by the SABRETRAIN Project, which is funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission. BV received support from the Scottish Executive Environment and Rural Affairs Department (SEERAD) and INIA, and JAW received funding from the Biotechnology and Biological Sciences Research Council (BBSRC).

REFERENCES

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams, 2007 Inbreeding in genome-wide selection. *J.Anim.Breed.Genet.* **124**: 369-376.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008 Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLoS ONE* **3**: e3395.

Dekkers, J. C. M., 2007 Prediction of response from marker-assisted and genomic selection using selection index theory. *J.Anim.Breed.Genet.* **124**: 331-341.

Erlich, H. A., T. F. Bergstrom, M. Stoneking, and U. Gyllensten, 1996 HLA sequence polymorphism and the origin of humans. *Science* **274**: 1552-1554.

Falconer D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman, Harlow, UK.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve *et al.* 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-8U3.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.* 2002 The structure of haplotype blocks in the human genome. *Science* **296**: 2225-2229.

Gilmour, A. R., R. Thompson, and B. R. Cullis, 1995 Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**: 1440-1450.

Goddard, M. E., 2008 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **online**.

Goldstein, D. B., 2001 Islands of linkage disequilibrium. *Nat. Genetics* **29**: 109-111.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim *et al.* 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc.Natl.Acad.Sci.U.S.A* **101**: 2398-2403.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**: 2389-2397.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009a Invited review: Genomic selection in dairy cattle: progress and challenges. *J.Dairy Sci.* **92**: 433-443.

Hayes B.J., Daetwyler H.D., Bowman P.J., Moser G., Tier B., Crump R.E., Khatkar M.S., Raadsma H.W., Pryce J. & Goddard M.E. Accuracy of genomic selection: comparing theory and results. Association for the Advancement of Animal Breeding and Genetics 30th Anniversary Conference. Proc.Assoc.Advmt.Anim.Breed. 17, 352-355. 2009b.

Hayes, B. J., and M. E. Goddard, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33**: 209-229.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009c Increased accuracy of artificial selection by using the realized relationship matrix. *Genet.Res.* **91**: 47-60.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. *Theor.Appl.Genet.* **38**: 226-231.

Hudson, R. R., 1983 Properties of A Neutral Allele Model with Intragenic Recombination. *Theor. Pop. Biol.* **23**: 183-201.

Lee, S. H., J. H. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher, 2008 Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* **4**: e1000231.

Long, N., D. Gianola, G. J. Rosa, K. A. Weigel, and S. Avendano, 2007 Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J.Anim Breed.Genet.* **124**: 377-389.

Lund, M. S., G. Sahana, D. J. de Koning, G. Su, and O. Carlborg, 2009 Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. *Bmc Proc.* **3 Suppl 1**.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Meuwissen, T. H. E., 2009 Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* **41**.

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams, 2009 A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.* **41**.

NejatiJavaremi, A., C. Smith, and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* **75**: 1738-1745.

Raadsma, H. W., G. Moser, R. E. Crump, M. S. Khatkar, K. R. Zenger *et al.* 2008 Predicting Genetic Merit for Mastitis and Fertility in Dairy Cattle using Genome Wide Selection and High Density SNP Screens. *Animal Genomics for Animal Health* **132**: 219-223.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen, 2009 Reducing dimensionality for prediction of genome-wide breeding values. *Genet. Sel. Evol.* **41**.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Pop. Biol.* **2**: 125-141.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.* 2007 Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* **17**: 520-526.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. Ser. B-Method.* **58**: 267-288.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* **91**: 4414-4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.* 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **92**: 16-24.

Villanueva, B., R. Pong-Wong, J. Fernandez, and M. A. Toro, 2005 Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* **83**: 1747-1752.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu *et al.* 2006 Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *Plos Genetics* **2**: 316-325.

Wang, C. S., J. J. Rutledge, and D. Gianola, 1994 Bayesian-Analysis of Mixed Linear-Models Via Gibbs Sampling with An Application to Litter Size in Iberian Pigs. *Genet. Sel. Evol.* **26**: 91-115.

Weigel, K. A., 2001 Controlling inbreeding in modern breeding programs. *J. Dairy Sci.* **84**: E177-E184.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. *Genet. Res.* **75**: 249-252.

Yi, N. J., and S. H. Xu, 2008 Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045-1055.

Chapter 5

Accuracy of Genomic Selection: Comparing Theory and Results

B. J. Hayes[1], H. D. Daetwyler[2,3], P. Bowman[1], G. Moser[3], B. Tier[4], R. Crump[4], M. Khatkar[5], H. Raadsma[5], and M. E. Goddard[1,6]

[1]Biosciences Research Division, Department of Primary Industries Victoria, 1 Park Drive, Bundoora 3083, Australia, [2]The Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, EH25 9PS, United Kingdom, [3]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands, [5]Centre for Advanced Technologies in Animal Genetics and Reproduction (ReproGen), University of Sydney, Camden NSW 2570, Australia, [4]Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351, Australia, [6]Faculty of Land and Food Resources, University of Melbourne, Parkville 3010, Australia.

Published in Proc. of Assoc. Advmt. Anim. Breed., 2009, 17:352-355.

ABSTRACT

Deterministic predictions of the accuracy of genomic breeding values in selection candidates with no phenotypes have been derived based on the heritability of the trait, number of phenotyped and genotyped animals in the reference population where the marker effects are estimated, the effective population size and the length of the genome. We assessed the value of these deterministic predictions given the results that have been achieved in Holstein and Jersey dairy cattle. We conclude that the deterministic predictions are useful guide for establishing the size of the reference populations which must be assembled in order to predict genomic breeding values at a desired level of accuracy in selection candidates.

INTRODUCTION

Genomic selection refers to the selection of animals for breeding based on genomic breeding values. Meuwissen *et al.* (2001) demonstrated using simulation that the accuracy of genomic breeding values can be very high if they are predicted from a large number of DNA markers. Provided the markers are dense enough, the accuracy of genomic breeding values will depend on the number of individuals genotyped and phenotyped in the reference population where the effect of the markers are predicted, the heritability of the trait, and the number of independent loci or chromosome segments in the population (Daetwyler *et al.* 2008; Goddard 2008). Goddard (2008) and Hayes et al (2009) further derived deterministic predictions of the number of independent chromosome segments based on the effective population size and the length of the genome of the species in question. These deterministic predictions would have great value in guiding the design of experiments to implement genomic selection if the accuracy they predicted agreed with that observed in real data. Such data is now available; recently, VanRaden *et al.* (2009) reported accuracies of genomic breeding values as high as 0.75 for total merit index in Holstein Friesian dairy cattle using 38416 single nucleotide polymorphism (SNP) markers genotyped in 3576 progeny tested bulls. Accuracies of genomic selection are also available for Australian Holstein Friesian and Jersey cattle, using a similar number of SNPs.

The aim of this paper was to assess the value of the deterministic predictions of accuracy of genomic breeding values given results that have been achieved in Holstein and Jersey dairy cattle.

MATERIALS AND METHODS

In Daetwyler et al. (2008) the accuracy of genomic breeding values was predicted as $r = \sqrt{Nh^2 /(Nh^2 + q)}$ where N = number of individuals genotyped and phenotyped in the reference population, h^2 = heritability of trait or reliability of breeding values in the reference population, q = number of independent chromosome segments in the population. Daetwyler et al. (2008) also proposed a corrected for their prediction when $N \geq q$. The correction was to add $r^4 q/(2N)$ to the above prediction to get the final accuracy. As $N \geq q$ for most of the situations we will investigate, we will use the accuracy from the above equation with the correction.

In Goddard (2008), the accuracy of genomic breeding values was predicted as $r = \sqrt{\left[1 - \lambda /(2N\sqrt{a}) * \ln((1 + a + 2\sqrt{a})/(1 + a - 2\sqrt{a})\right]}$ where a= 1+2 λ/N , and λ= qk/h^2 , with $k = 1/log(2N_e)$, where N_e is the effective population size. Note that this derivation assumes that σ_e^2 is close to the phenotypic variance. For both predictions, the value of q used was the number of independent chromosome segments, $2N_eL$, where L is the length of the genome in Morgans (Hayes *et al.* 2009). The difference between the formula of Daetwyler et al. (2008) and Goddard (2008) potentially arises because Goddard (2008) assumed that that the effect estimate for common QTL is more accurate for QTL with intermediate allele frequency, because they explain more of the genetic variance than QTL with extreme allele frequency. In contrast, Daetwyler et al. (2008) assumed the accuracy of estimating QTL effects was equal regardless of their frequency. The accuracy of genomic breeding values for the two deterministic predictions were compared for a range of heritabilities, N=5000 and N_e=100.

Accuracy of breeding values from the two predictions were also compared to accuracies of genomic breeding values reported by VanRaden et al. (2009) and United States Department of Agriculture results (http://aipl.arsusda.gov/reference/genomic_comparison_yng_0901.htm) for total net merit in Holstein Friesian cattle and Jersey cattle, and our own results in Australian data for these two breeds. The phenotypic records in the reference population were daughter yield deviations (DYD) for total merit index for the US data or de-regressed breeding values for Australian Profit Ranking (APR) in the Australian data. The average reliability of the DYD in the reference population was 0.9. In order to deterministically predict the accuracy that these experiments could have achieved, an assumption of the N_e in each breed was required. Young & Seykora (1996) gave an estimate of 100 for the N_e of US Holsteins. The N_e in Australian Holsteins is similar (de Roos *et al.* 2008). For US Jerseys, the effective population size is smaller, with a recent estimate of 30 (Weigel 2001). The N_e of Jersey's in Australia is likely to be similar given the large contribution of US Jersey bulls to the Australian population. Given these estimates of N_e in the two breeds, we used N_e=100 in the predictions for Holsteins and 30 in Jersey's. A genome length of 30 Morgans was assumed.

RESULTS AND DISCUSSION

The accuracies of genomic breeding value predicted by Goddard (2008) and Daetwyler et al. (2008) are similar, though Daetwyler et al. (2008) would predict a lower accuracy of breeding value at low to moderate heritabilities given the same number of independent chromosome segments and number of phenotypic records, Figure 1. Both deterministic predictions agreed fairly well with the accuracies of genomic breeding value reported for US and Australian Holstein Friesian and Jersey dairy cattle, Figure 2. The % error was low for the Goddard prediction vs the US Holstein data at 3%.. However in the Australian Holstein data the observed accuracies were somewhat higher than the predictions. This may just reflect a small validation sample used in the Australian data leading to a large standard error for the estimate of reliability.

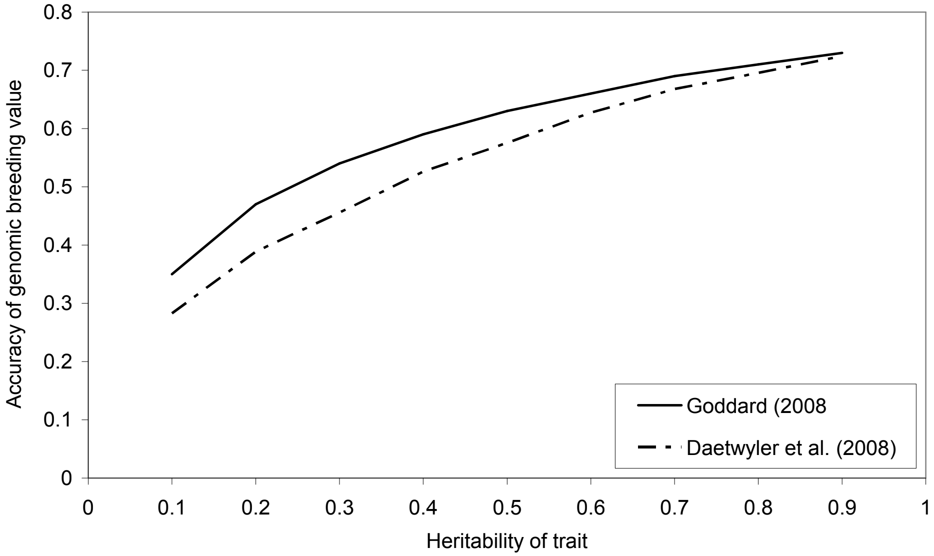


Figure 1. Accuracy of genomic breeding values with 5000 phenotypic records, effective population size of 100 and increasing heritability, predicted by the deterministic formula of Goddard (2008) or Daetwyler et al. (2008).

Another contributing factor may be that the deterministic predictions assume that the accuracy of breeding values is a result of the SNPs capturing the effect of QTL, whereas some of the accuracy of genomic breeding values in livestock populations may be a result of the SNPs capturing the effect of relationship, particularly if there are large half sib families in the population (eg. Habier *et al.* 2007). For comparison, the accuracy of parent average breeding values for net merit available for young bulls in the US data was 0.37 (VanRaden et al. 2009).

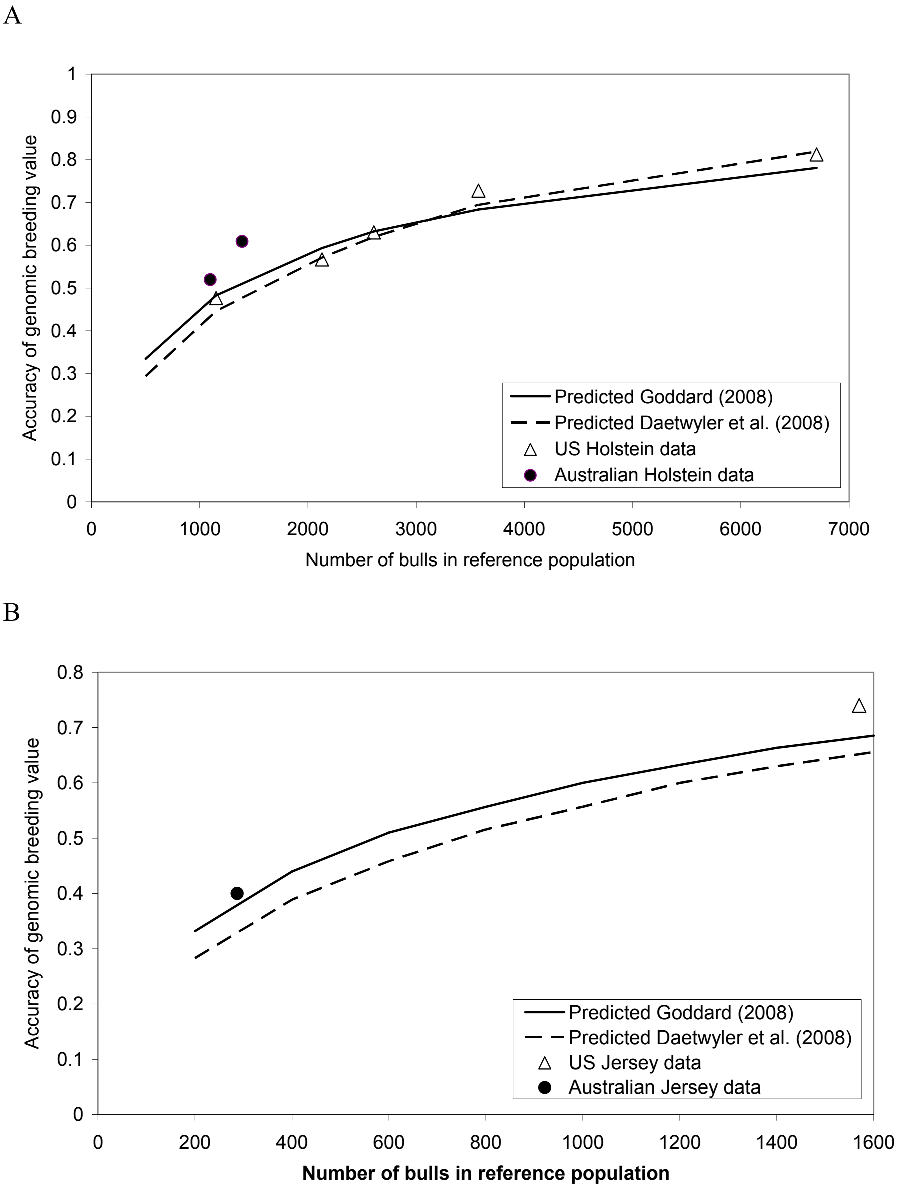


Figure 2. A. Accuracy of genomic breeding values from the deterministic prediction of Goddard (2008) and Daetwyler et al. (2008) with N_e=100, and accuracy of genomic breeding value for total merit index or Australian Profit ranking in US or Australian Holstein Friesian cattle. B. Accuracy of genomic breeding values from deterministic predictions with N_e=30, and accuracy of genomic breeding value for total merit index or Australian Profit ranking in US or Australian Jersey cattle respectively.

The deterministic method of Goddard (2008) used here assumes a normal distribution of QTL effects. For the majority of traits studied by Van Raden et al (2008), methods for predicting genomic breeding values which assumed a normal distribution of quantitative trait loci (QTL) effects performed almost as well as methods assuming a exponential distribution of QTL effects. The exception was traits with a QTL of known large effect, eg. fat percentage (Grisart *et al.* 2004). For such traits, the deterministic prediction of Goddard (2008) would under-predict accuracy of genomic selection. The accuracies of prediction also depend on N_e. The values of N_e used here are estimates of N_e in the recent past, however N_e in cattle has been much larger historically. It is not clear how the change in historical N_e should affect accuracy of genomic breeding values. Nevertheless, using current N_e gave good agreement between predictions and observed results.

CONCLUSIONS

The deterministic predictions of accuracy of genomic selection presented by Goddard (2008) extended by Hayes et al. (2009), and that of Daetwyler et al. (2008) agree well with observed accuracies of genomic selection in US and Australian Holstein Friesians and Jerseys. We can conclude that these deterministic predictions are a useful tool to guide design of genomic selection experiments, for example how large should the reference population be to achieve a desired level of accuracy. It must be noted we have compared predicted and observed accuracies of genomic breeding value for a situation where phenotypes were very accurate predictors of breeding value. The performance of the deterministic predictions of both Daetwyler et al. (2008) and Goddard (2008) should be also evaluated in other situations where the heritability of the trait is lower, as the difference predicted accuracy of genomic selection is greater at lower heritabilities.

ACKNOWLEDGEMENTS

The authors are grateful to Curt van Tassell and Tad Sonstegard from the USDA for providing genotypes under a collaborative agreement between USDA and Department of

Primary Industries Victoria. We are also grateful to Paul Van Raden and George Wiggans of USDA for providing some of the information required to assemble Figure 2.

REFERENCES

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008 Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLoS ONE* **3**: e3395.

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

Goddard, M. E., 2008 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**: 245-252.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim *et al.* 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc.Natl.Acad.Sci.U.S.A* **101**: 2398-2403.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**: 2389-2397.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. *Genet.Res.* **91**: 47-60.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.* 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **92**: 16-24.

Weigel, K. A., 2001 Controlling inbreeding in modern breeding programs. *J. Dairy Sci.* **84**: E177-E184.

Young, C. W., and A. J. Seykora, 1996 Estimates of inbreeding and relationship among registered Holstein females in the United States. *J. Dairy Sci.* **79**: 502-505.

Chapter 6

Inbreeding in Genome-wide Selection

Hans D. Daetwyler[*#], Beatriz Villanueva[†], Piter Bijma[#], and John A. Woolliams[*]

[*]Genetics and Genomics, Roslin Institute (Edinburgh), Roslin, Midlothian, EH25 9PS, UK; [#]Animal Breeding and Genomics Centre, Wageningen University and Research Centre, 6700 AH Wageningen, The Netherlands; [†]Sustainable Livestock Systems, Scottish Agriculture College, West Mains Road, Edinburgh, EH9 3JG, UK

Published in Journal of Animal Breeding and Genetics, 2007, 124: 369-376.

ABSTRACT

Traditional selection methods, such as sib and best linear unbiased prediction (BLUP) selection, which increased genetic gain by increasing accuracy of evaluation have also led to an increased rate of inbreeding per generation (ΔF_G). This is not necessarily the case with genome-wide selection, which also increases genetic gain by increasing accuracy. This paper explains why genome-wide selection reduces ΔF_G when compared to sib and BLUP selection. Genome-wide selection achieves high accuracies of estimated breeding values through better prediction of the Mendelian sampling term component of breeding values. This increases differentiation between sibs and reduces coselection of sibs and ΔF_G. The high accuracy of genome-wide selection is expected to reduce the between family variance and reweigh the emphasis of estimated breeding values of individuals towards the Mendelian sampling term. Moreover, estimation induced intraclass correlations of sibs are expected to be lower in genome-wide selection leading to a further decrease of coselection of sibs when compared to BLUP. Genome-wide prediction of breeding values, therefore, enables increased genetic gain while at the same time reducing ΔF_G when compared to sib and BLUP selection.

INTRODUCTION

Meuwissen et al. (2001) described genome-wide prediction (GWP) methods to estimate haplotype effects, assuming a high density genetic marker map across the entire genome. Their methods yielded high accuracies of estimated breeding values (EBV) based on genotypic information in newborn individuals without phenotypic records. Moreover, they showed that this high accuracy could then be maintained, with only minor loss, over subsequent generations when neither offspring nor parent had records.

In the past, methods proposed to increase accuracies of EBVs have resulted not only in accelerated rates of genetic gain (ΔG) but also in increased inbreeding rates per generation (ΔF_G). This was particularly true for methods that include information on relatives such as Best Linear Unbiased Prediction (BLUP) (Henderson 1975). When EBVs derived from

BLUP were used in a traditional way, namely ranking the candidates on these EBVs and truncating the distribution to choose those with the highest values, ΔG was increased but so was ΔF_G (Belonsky & Kennedy 1988). This meant that short term gain was greater at a cost to long term gain (Quinton *et al.* 1992). While the long-term consequences of genetic variance reduction are often ignored in commercial breeding schemes, high ΔF_G also has more immediate effects. Monogenic recessive alleles can drift to high frequencies due to high usage of one superior individual (e.g. CVM in Holsteins; Agerholm *et al.* 2001; Kearney *et al.* 2005) and inbreeding depression can have increased impact because the degree of depression is empirically associated with ΔF_G (Wiener *et al.* 1992).

This experience with BLUP, coupled with the increased ΔF_G observed when selection intensity is increased, has led to an empirical association being perceived between gain and inbreeding. However, this association is much weaker in GWP. This paper has the objective of explaining why the increased accuracy of genome-wide methods leads to decreased ΔF_G when compared to sib and BLUP selection. Thus, GWP provides a method for achieving both the short term goal of increased and sustained ΔG and the long-term needs for maintaining genetic variation. The approach taken will be to examine the existing quantitative genetic theory related to inbreeding and selection both for truncation selection and for methods using optimum contributions (e.g. Meuwissen 1997; Grundy *et al.* 1998).

Inbreeding with mass and BLUP truncation selection. It is useful to discuss in terms of the breeders equation, $\Delta G = i\rho\sigma_A$, how increasing ΔG has led to increased ΔF_G in mass and BLUP truncation selection. The additive genetic standard deviation (σ_A) is a constant for a trait in the short term and, therefore, advances in ΔG come from increasing the selection intensity (i) or the accuracy of EBVs (ρ).

The first option of increasing ΔG is by increasing i. However, reducing the proportion of individuals selected decreases the number of parents and invariably leads to increased

ΔF_G. This is true for both mass and BLUP truncation selection as shown in Figure 1 for different values of heritability (h^2).

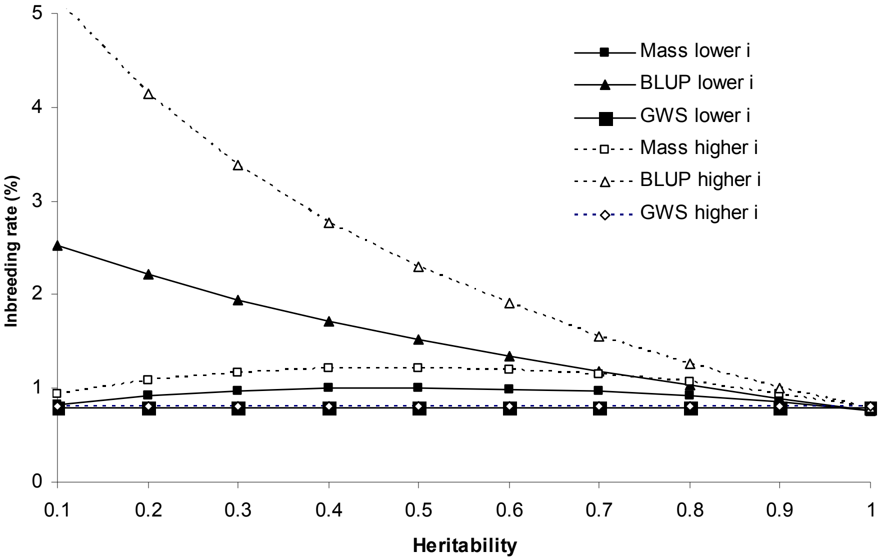


Figure 1. Inbreeding rates per generation from mass and best linear unbiased prediction (BLUP), and genome-wide selection (GWS) at two selection intensities (i) withheritability ranging from 0.1 to 1.0, predicted with SelAction (Rutten *et al.* 2002). SelAction input parameters: 20 males, 200 females, 4 male and 4 female offspring per dam, proportion selected = 0.05 males, 0.20 females (lower i), and 0.01 male, 0.1 female (higher i), mass used own performance, BLUP included information on own performance, full-sibs and half-sibs, GWS used only information on phenotypes in the marker trait, and GWS accuracy assumed was 0.85.

The second way to increase ΔG is to increase accuracy. Consider mass selection with a simple model of additive and independent environmental effects. Here both accuracy and intraclass correlation among sibs are determined entirely by h^2 and there is a balance between two effects. Correlations among sibs increase as h^2 increases, leading to increased coselection of sibs and higher ΔF_G. In contrast, at higher h^2 the Bulmer effect reduces the between family genetic variance (σ_B^2) relatively more, reducing co-selection of sibs and ΔF_G. When h^2 is lower than intermediate values, ΔF increases because of the first effect, but when h^2 increases beyond intermediate values the balance shifts to the second effect and ΔF_G is decreased (Figure 1).

In contrast to mass selection, BLUP makes use of information from all relatives, appropriately weighted to maximise accuracy. The higher accuracy leads to a stronger Bulmer effect, which reduces the σ_B^2. The Bulmer effect is less dependent on h^2 in BLUP than in mass selection and, therefore, has a relatively small impact on intraclass correlations. However, intraclass correlations are increased due to inclusion of sib information because of additional induced correlations which are due to using common information (i.e. residual terms averaged among relatives) (Wray *et al.* 1990). The high intraclass correlations increase coselection of sibs and ΔF_G. The emphasis on sib information is high at lower h^2 but decreases when h^2 increases and so the coselection of relatives always decreases as h^2 increases, in contrast to mass selection. Therefore, both elements (co-selection and Bulmer effect) combine to produce the downward trend of ΔF_G as h^2 increases (Figure 1). As h^2 approaches 1 the use of sib information becomes unimportant when the phenotype is observed and the ΔF_G approaches that achieved with mass selection.

Three components of a breeding value. The breeding value of an individual can be conceived as having three components (Woolliams 2007): (i) the breeding value of the sire; (ii) the breeding value of the dam; and (iii) the Mendelian sampling term, which is the aggregate deviation arising from sampling the segregation of alleles within the sire and within the dam (See Figure 2 for an illustration). Information on ancestors and collateral relatives increases accuracy through directly adding precision on the first two of these components. The accuracy of the Mendelian sampling term can be increased by using an individual's phenotypic record or progeny information. In practice, most BLUP selection schemes increase accuracy by capturing additional information on ancestors and collateral relatives, because progeny information is often not available at the time of selection. It becomes clear that, at the time of selection, BLUP relies heavily on increasing accuracy of σ_B^2 to increase ΔG (Figure 2). In contrast, GWP utilises the Mendelian sampling term more heavily and the consequences of this feature on ΔF_G will now be discussed further.

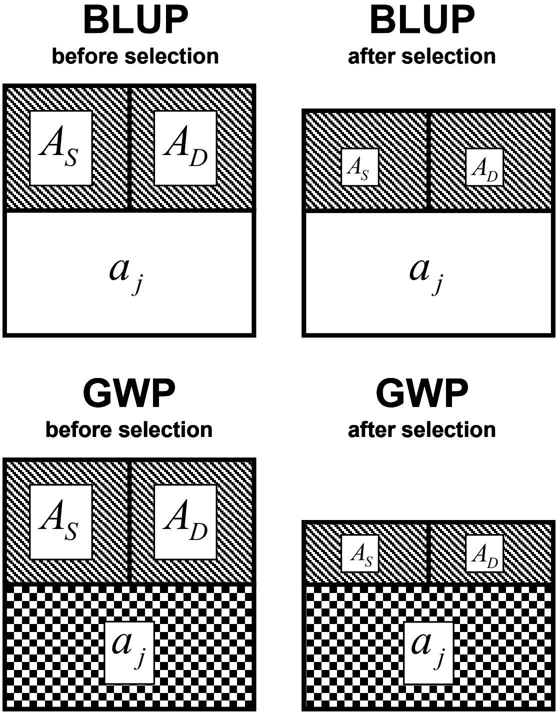


Figure 2. Representation of the sources of information utilised (shaded areas) and their proportions before and after selection (i.e. selection reduces the between family variance) when using best linear unbiased prediction (BLUP) and genome-wide prediction (GWP) to predict the estimated breeding value of a newborn with no phenotypic record. A_S is the sire breeding value, A_D is the dam breeding value, and a_j is the Mendelian sampling term.

Genome-wide prediction of breeding values. Meuwissen et al. (2001) demonstrate that GWP increases accuracy of EBV prediction. The important issue is how the increased accuracy is achieved, namely using the markers to explain the Mendelian sampling terms. In the past, physiological indicator traits, which were genetically correlated to a particular trait of interest, were used to select young animals and increased ΔG by giving an early indication of an animal's Mendelian sampling term (Woolliams & Smith 1988). Genotyping technology provided another tool that could be used to gain insight into an animal's unique ability, as individuals could be genotyped at birth or even as an embryo. Marker assisted selection (MAS) was found to increase ΔG because each genetic marker explained a part of the within family variance (Mendelian sampling variance, σ_M^2) (Meuwissen & Van Arendonk 1992). Preselecting young dairy sires with MAS increased ΔG and offered a method to select within families (Mackinnon & Georges 1998). While the number of markers is dramatically increased with dense marker maps, the principle is the same. Thus GWP offers the possibility that an individual's Mendelian sampling term can be estimated with great accuracy early in its life.

As an example of the potential of GWP consider the EBV accuracy achieved by Meuwissen et al. (2001). An individual with only parent information and no record has an EBV ($\hat{A}$) that is equal to $\hat{A}_i = (0.5)\hat{A}_S + (0.5)\hat{A}_D$, where $\hat{A}_S$ and $\hat{A}_D$ are the sire and dam EBVs, respectively. The accuracy of the EBV ($\rho_{A\hat{A}}$) is $\rho_{A\hat{A}} = \sigma_{A\hat{A}}\left[\sigma_A\sigma_{\hat{A}}\right]^{-1}$, where $\sigma_{A\hat{A}}$ is the covariance between true breeding value (TBV) and EBV and $\sigma_{\hat{A}}$ is the EBV standard deviation. Assuming that the parent EBVs have an accuracy of 1 (i.e. $\hat{A} = A$), then $\sigma_{A\hat{A}} = (0.5)\sigma_A^2 = \sigma_{\hat{A}}^2$, and $\rho_{A\hat{A}} = (0.5)\sigma_A^2\left[(0.5)\sigma_A^2\sigma_A^2\right]^{-1/2} = 0.71$, which is the upper bound of accuracy for an animal at birth when using conventional BLUP. The GWP Bayesian method achieved an accuracy of 0.85 (Meuwissen *et al.* 2001). Hence, the difference in accuracy of 0.14 observed in GWP and the upper bound in conventional

BLUP must originate from the increased accuracy of the Mendelian sampling term estimate (Woolliams *et al.* 2002).

The accuracy of the Mendelian sampling terms ($\rho_{M\hat{M}}$) in GWP can be approximated for this example. Assuming that σ_B^2 was explained precisely (i.e. $\sigma_B^2 = (0.5)\sigma_A^2$), then the proportion of the Mendelian sampling variance explained by the GWP EBV ($\rho_{M\hat{M}}^2 \sigma_M^2$) is:

$$\rho_{M\hat{M}}^2 \sigma_M^2 = \rho_{A\hat{A}}^2 \sigma_A^2 - (0.5)\sigma_A^2,$$

where $\rho_{A\hat{A}}^2$ is the proportion of σ_A^2 explained by the EBV. When $\rho_{A\hat{A}}$ is 0.85 in GWP, then the $\rho_{M\hat{M}}$ of GWP is:

$$\rho_{M\hat{M}} = (\rho_{A\hat{A}}^2 - 0.5)^{1/2} \sigma_A \left[(0.5)\sigma_A^2\right]^{-1/2} = (0.85^2 - 0.5)^{1/2} \sigma_A \left[(0.5)\sigma_A^2\right]^{-1/2} = 0.67 .$$

The approximated increase of 0.67 in the accuracy of $\rho_{M\hat{M}}$ of GWP is very large when compared to $\rho_{M\hat{M}} = 0$ in conventional BLUP. However, it is unlikely that σ_B^2 is explained precisely. A more plausible scenario would be that $\sigma_B^2 < 0.5\sigma_A^2$ and, if overall GWP $\rho_{A\hat{A}}$ is still 0.85, this would result in $\rho_{M\hat{M}} > 0.67$. Conventional BLUP EBVs are parent averages when an animal has no record of its own, whereas GWP identifies and uses the new Mendelian sampling variation that is generated in each generation. This exploitation of new variation is the major source of increased ΔG of GWP over conventional approaches. Utilizing Mendelian variation is key to achieving sustained genetic progress (see Figure 2) and reducing ΔF_G (Woolliams & Thompson 1994; Woolliams *et al.* 1999).

Inbreeding with truncation genome-wide selection. In GWP, ΔF_G can be much lower than in mass or BLUP for comparable resources and there are several reasons why this is the case. First, GWP breeding values are less correlated between sibs because they rely more on Mendelian sampling information (Figure 2). The increased accuracy of Mendelian sampling terms in GWP allows for better differentiation within families and leads to lower coselection of sibs, which reduces ΔF_G. Second, GWP achieves higher accuracy for all values of phenotype h^2 and, therefore, a strong Bulmer effect is induced by selection and reduces σ_B^2. Due to the Bulmer effect, GWP further re-weights the offspring EBV towards the Mendelian sampling term (Figure 2), which further reduces coselection of sibs and ΔF_G. This is repeated in successive generations where an individual's breeding value has less influence on selection of descendents. The above processes decrease ΔF_G because the Mendelian sampling term arises from the random sampling of alleles carried by the parents and the variance of these terms is regenerated in each generation. In the long-term the Mendelian sampling variance is reduced by the loss of alleles due to inbreeding. Moreover, in species where only males can attain high accuracies (through progeny tests) and have a high number of selected offspring, GWP is expected to shift the selection emphasis from males towards females because males and females will have more similar accuracies. This leads to more evenly distributed long-term contributions among male ancestors and, therefore, decreases ΔF_G when BLUP and GWP are compared at the same ΔG. This would be the effect of the shift in emphasis from sires to dams in dairy cattle pointed out by Schaeffer (2006).

Inbreeding and genetic gain with optimum contribution genome-wide selection. The previous section has examined how genome-wide selection (GWS) may affect ΔG and ΔF_G when the design parameters are fixed (i.e. truncation selection). However, a more appropriate approach is to consider how to maximise ΔG with fixed resources and fixed ΔF_G by optimising long-term genetic contributions of the selection candidates (Meuwissen 1997; Grundy *et al.* 1998).

Optimum contribution selection is attempting to allocate contributions of candidates and ancestors in relation to the best estimate of the Mendelian sampling term of each individual (Avendano *et al.* 2004). The optimum solution is, beyond a threshold value, to have a linear relationship between the long-term contribution of an individual and its (true) Mendelian sampling term (Grundy *et al.* 1998). In reality, however, this optimum cannot be attained for two reasons. First, contributions of distinct individuals cannot always be changed independently, e.g. it is not possible to change the contribution of an individual without changing that of its parent. Second, because Mendelian sampling terms are estimated with limited precision, the true optimum contributions are also known with limited precision. Hence, the solution is a compromise repeated each generation as more accurate information on Mendelian sampling terms becomes available. This was confirmed by Avendano et al. (2004), who showed by simulation that the major component by which optimum contribution algorithms keep ΔF_G at a predefined level, while maximising ΔG, is the estimated Mendelian sampling term. Grundy *et al.* (1998, 2000) show that with optimum contributions, ΔG is proportional to Mendelian sampling term estimate. It therefore follows directly that a more accurate estimate of the Mendelian sampling term will lead to more ΔG while not affecting ΔF_G. Hence, the use of optimum contribution procedures and GWP together will always result in more ΔG when compared at the same ΔF_G. Quantifying the full benefit of GWP in relation to inbreeding will require further development of methods to predict the accuracy of the Mendelian sampling term (Avendano *et al.* 2005).

Implications on inbreeding of frequency of haplotype effect re-estimation. There are other considerations in GWP that reinforce why GWP is expected to reduce ΔF_G, but these may depend on how GWP is implemented. Two cases can be considered: (i) where haplotype effects are estimated in either earlier generations or, conceivably, in related but distinct populations, and (ii) where haplotype effects are re-estimated each generation or whenever new phenotypic information is available as part of a continuous process.

No updating. When GWS is used with previously estimated haplotype effects with no updating, then the EBV is a sum of haplotype values which do not change over generations. In this case the marker based genome-wide EBV can be treated as a classical trait with $h^2 = 1$ and its genetic correlation with the original phenotypic trait is equal to the accuracy of GWP (Schrooten *et al.* 2005; Dekkers 2007a; Dekkers 2007b). Thus, genome-wide truncation selection is expected to have a similar ΔF_G to those achieved by mass and BLUP selection at $h^2 = 1$. Figure 1 shows there is no distinction in this case between mass selection and BLUP, as BLUP ΔF_G tends towards mass selection ΔF_G as h^2 increases. This trend is substantiated by the fact that a lower ΔF_G can be achieved in BLUP by artificially increasing the trait h^2 which reduces the reliance on relatives (Toro & Perez-Enciso 1990; Grundy *et al.* 1994). When predicted with SelAction (Rutten *et al.* 2002), the ΔF_G of GWS is similarly low as BLUP at $h^2 = 1$ and, in addition, stays at this low and constant level regardless of the h^2 of the original phenotypic trait (Figure 1).

Another property of traits with $h^2 = 1$ is that increasing selection intensity by reducing the proportion of candidates selected, while increasing the total number of candidates, has only a small effect on ΔF_G. This scenario would be equivalent to genotyping more individuals but still selecting the same number of parents to increase selection intensity. In Figure 1, while BLUP shows a large increase in ΔF_G at lower h^2, GWS (when treated as a trait with $h^2 = 1$) results only in a small and constant increase in ΔF_G regardless of phenotypic trait h^2. Therefore, when applying GWS with no updating of haplotypes, selection intensity can be increased in this way with relatively little consequence on ΔF_G.

Continuous updating. When GWP is applied with continuous re-estimation of haplotype effects, then the process of estimation might be considered as inducing correlations due to the averaging of residual terms of relatives just like the estimation of sire and dam EBVs in BLUP. This applies particularly in a simple pedigree with only parents and offspring. In BLUP, all offspring of a parent are averaged to provide an estimate of the parent EBV, so differences between EBVs of sires (dams) are contrasts between sire (dam) family means. This is the origin of the intraclass correlation leading to coselection of sibs in BLUP that is described above. In GWP, if haplotypes are re-estimated continuously then contrasts are made across the population as a whole comparing carriers and non-carriers of particular alleles both between and within families. Thus the estimation-induced intraclass correlations act much less strongly as sources of coselection of sibs. This would reduce ΔF_G when compared to BLUP.

It should be noted that continuous re-estimation of haplotype or marker effects must be more effective in generating ΔG for a trait than not updating effects. This follows because re-estimating marker effects with additional phenotype information must result in at least as good accuracy compared to ignoring it. In each generation, novel additive genetic variation is generated which is not captured by the original estimate of the haplotype effects. This is due to the decay of linkage disequilibrium between markers and to changes in allele frequencies which are associated with mutation, dominance and epistasis. The cost of collecting some phenotypes might prohibit regular updating of haplotype effects and so allowing some loss of accuracy (and ΔG) may be a cost-effective option.

Impact of linkage on inbreeding. In this paper all comparisons of ΔF_G between different selection methods are based on inbreeding as calculated from the pedigree. Differences do exist between inbreeding calculated from pedigree information and inbreeding computed from genotypic data.

The pedigree based method is an expectation assuming neutral loci and, therefore, the two alleles of the same neutral locus on two homologous chromosomes have an equal chance of being selected. This ignores that the two alleles present in non-neutral loci on either chromosome may have different effects on a trait which leads to unequal selection probabilities between the two alleles of the neutral locus when there is linkage (Santiago & Caballero 1998). The proportion of loci that is actually neutral, when neutral is defined as not under selection directly or indirectly (i.e. linked to an allele under selection), is unknown. However, while it was found that the assumption of no linkage is violated in small genomes (< 10 Morgans), it becomes progressively more appropriate as genomes become larger (Fernandez *et al.* 2000; Villanueva *et al.* 2005). Thus in farm animal species which typically have genome sizes of 20 to 30 Morgans, accurate average inbreeding rates across the whole genome can be predicted from pedigree records.

When inbreeding is calculated from genotypic data the expectation is adjusted with identity-by-state probabilities at the marker loci to yield actual inbreeding at specific locations across the genome (Pong-Wong *et al.* 2001; Liu *et al.* 2002; Roughsedge *et al.* 2006). The increasing amount of genotypic data available will lead to new methods for calculate inbreeding which could give an indication of the effect of linkage on the accumulation of localised inbreeding across the genome. The potential exists; therefore, to get a more complete picture of inbreeding with genotypic methods than with pedigree based methods.

Practical issues of implementing genome-wide selection. This article has discussed an important benefit of GWS, namely increased gain with no cost to inbreeding. Other potential benefits that GWS offers to livestock breeders are: (i) overcome age limitations whilst offsetting additional costs through changes in structure; (ii) overcome or reduce sex limitations, or more generally limitations caused by measuring only special subsets e.g. expensive or destructive testing; (iii) use in non-pedigreed populations; (iv) a direct link between the genetic evaluation and the genome. Nevertheless, the relevance and benefits described will vary among sectors and depend on practical issues related to the implementation of GWS.

Generation Interval. Genome-wide selection is expected to increase ΔG and reduce ΔF_G due to the high accuracy of the Mendelian sampling term. However it would be expected that there are opportunities to reduce the generation interval with GWS since a substantial increase in accuracy is available in the newborn. In dairy cattle, it has the potential to reduce the generation interval of sires of bulls and dams from six to two years, as progeny tests may become unnecessary (Schaeffer 2006). This may increase the annual inbreeding rate (ΔF_A). However, the biological risks of inbreeding depression and deleterious alleles are more relevant in the context of ΔF_G, because balancing processes, such as mutation, also occur per generation. Optimum contributions with constrained ΔF_G

(Grundy *et al.* 1998; Grundy *et al.* 2000) could be used to manage the transition to shorter generation intervals. Whether or not the scheme would evolve into that of Schaeffer (2006) remains unknown. However, in a truncation scheme an increase in ΔF_A may occur, but the arguments above would be expected to remain valid and more gain achieved with GWS if compared to BLUP at same ΔF_G per generation.

The need to manage pedigrees. GWS does not fully remove the impact of pedigree on ΔF_G. Parents come as packages of haplotypes, and with truncation selection, parents with good packages will tend to have more offspring selected even though individual haplotypes are being evaluated. While GWS decreases ΔF_G when compared to BLUP, it is not inbreeding free. Breeding programs are competitive and are expected to push for more ΔG by, for example, increasing selection intensity through a reduction of the number of parents which would increase ΔF_G. Therefore, the need to manage inbreeding using tools such as optimum contributions to maximise ΔG in relation to ΔF_G remains.

CONCLUSION

This paper has outlined why GWS is expected to result in lower ΔF_G than BLUP selection. The main reason for this reduced ΔF_G is that GWP will result in an increased estimation accuracy of the Mendelian sampling term. This allows for better differentiation within families and leads to lower coselection of sibs, which reduces ΔF_G. The between family portion of the additive genetic variance in GWS is reduced quickly due to the high EBV accuracy and shifts the emphasis of selection in favour of the Mendelian sampling term which has no effect on inbreeding as it is regenerated in each generation. Haplotype effects which are used for several generations without re-estimation will resemble a trait with $h^2 = 1$ and result in low and constant ΔF_G regardless of the original trait h^2. When haplotype effects are re-estimated in each generation, contrasts between haplotypes are made both between and within families, thereby reducing coselection and through reduced estimation induced correlations between sib EBVs. Mendelian sampling terms are also used in optimum contribution procedures which could be used to maximise ΔG at a preset rate of ΔF_G.

ACKNOWLEDGEMENTS

HDD is supported by the SABRETRAIN Project, which is funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission. BV receives support from the Scottish Executive Environment and Rural Affairs Department (SEERAD), and JAW receives funding from the Biotechnology and Biological Sciences Research Council (BBSRC). We thank the two anonymous reviewers for their comments.

REFERENCES

Agerholm, J. S., C. Bendixen, O. Andersen, and J. Arnbjerg, 2001 Complex vertebral malformation in Holstein calves. *J. Vet. Diag. Invest.* **13**: 283-289.

Avendano, S., J. A. Woolliams, and B. Villanueva, 2004 Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. *Genet. Res.* **83**: 55-64.

Avendano, S., J. A. Woolliams, and B. Villanueva, 2005 Prediction of accuracy of estimated Mendelian sampling terms. *J. Anim Breed.Genet.* **122**: 302-308.

Belonsky, G. M., and B. W. Kennedy, 1988 Selection on Individual Phenotype and Best Linear Unbiased Predictor of Breeding Value in A Closed Swine Herd. *J. Anim. Sci.* **66**: 1124-1131.

Dekkers, J. C. M., 2007a Marker-assisted selection for commercial crossbred performance. *J. Anim Sci.* **85**: 2104-2114.

Dekkers, J. C. M., 2007b Prediction of response from marker-assisted and genomic selection using selection index theory. *J.Anim.Breed.Genet.* **124**: 331-341.

Fernandez, B., E. Santiago, M. A. Toro, and A. Caballero, 2000 Effect of linkage on the control of inbreeding in selection programmes. *Genet. Sel. Evol.* **32**: 249-264.

Grundy, B., A. Caballero, E. Santiago, and W. G. Hill, 1994 A Note on Using Biased Parameter Values and Nonrandom Mating to Reduce Rates of Inbreeding in Selection Programs. *Anim. Prod.* **59**: 465-468.

Grundy, B., B. Villanueva, and J. A. Woolliams, 1998 Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. *Genet. Res.* **72**: 159-168.

Grundy, B., B. Villanueva, and J. A. Woolliams, 2000 Dynamic selection for maximizing response with constrained inbreeding in schemes with overlapping generations. *Anim. Sci.* **70**: 373-382.

Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**: 423-447.

Kearney, J. F., P. R. Amer, and B. Villanueva, 2005 Cumulative discounted expressions of sire genotypes for the complex vertebral malformation and beta-casein loci in commercial dairy herds. *J.Dairy Sci.* **88**: 4426-4433.

Liu, Y., G. B. Jansen, and C. Y. Lin, 2002 The covariance between relatives conditional on genetic markers. *Genet. Sel Evol.* **34**: 657-678.

Mackinnon, M. J., and M. A. J. Georges, 1998 Marker-assisted preselection of young dairy sires prior to progeny-testing. *Livest. Prod. Sci.* **54**: 229-250.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Meuwissen, T. H., and J. A. Van Arendonk, 1992 Potential improvements in rate of genetic gain from marker-assisted selection in dairy cattle breeding schemes. *J. Dairy Sci.* **75**: 1651-1659.

Meuwissen, T. H. E., 1997 Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* **75**: 934-940.

Pong-Wong, R., A. W. George, J. A. Woolliams, and C. S. Haley, 2001 A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genet. Sel. Evol.* **33**: 453-471.

Quinton, M., C. Smith, and M. E. Goddard, 1992 Comparison of Selection Methods at the Same Level of Inbreeding. *J. Anim. Sci.* **70**: 1060-1067.

Roughsedge T., Pong-Wong R. & Villanueva B. Optimised selection restricting coancestry at a specific location on the genome. Proc.of the 8th WCGALP. 2006.

Rutten, M. J. M., P. Bijma, J. A. Woolliams, and J. A. M. Van Arendonk, 2002 SelAction: Software to predict selection response and rate of inbreeding in livestock breeding programs. *J. Hered.* **93**: 456-458.

Santiago, E., and A. Caballero, 1998 Effective size and polymorphism of linked neutral loci in populations under directional selection. *Genetics* **149**: 2105-2117.

Schaeffer, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. *J.Anim. Breed.Genet.* **123**: 218-223.

Schrooten, C., H. Bovenhuis, J. A. Van Arendonk, and P. Bijma, 2005 Genetic progress in multistage dairy cattle breeding schemes using genetic markers. *J. Dairy Sci.* **88**: 1569-1581.

Toro, M., and M. Perez-Enciso, 1990 Optimization of Selection Response Under Restricted Inbreeding. *Genet. Sel. Evol.* **22**: 93-107.

Villanueva, B., R. Pong-Wong, J. Fernandez, and M. A. Toro, 2005 Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* **83**: 1747-1752.

Wiener, G., G. J. Lee, and J. A. Woolliams, 1992 Effects of Rapid Inbreeding and of Crossing of Inbred Lines on Conception Rate, Prolificacy and Ewe Survival in Sheep. *Anim. Prod.* **55**: 115-121.

Woolliams, J. A., 2007 Genetic contributions and inbreeding, pp. 147-165 in *Utilisation and conservation of farm animal genetic resources*, edited by K. Oldenbroek. Wageningen Academic Publishing, AE Wageningen.

Woolliams, J. A., P. Bijma, and B. Villanueva, 1999 Expected genetic contributions and their impact on gene flow and genetic gain. *Genetics* **153**: 1009-1020.

Woolliams J.A., Pong-Wong R. & Villanueva B. Strategic optimisation of short- and long-term gain and inbreeding in MAS and non-MAS schemes. 7th world congress of genetics applied to livestock production. Proc.7th WCGALP, 2002.

Woolliams, J. A., and C. Smith, 1988 The Value of Indicator Traits in the Genetic-Improvement of Dairy-Cattle. *Anim. Prod.* **46**: 333-345.

Woolliams J.A. & Thompson R. A theory of genetic contributions. 5th World Congress on Genetics Applied to Livestock Production. Proc. of the 5th WCGALP,19, 127-133. 1994.

Wray, N. R., J. A. Woolliams, and R. Thompson, 1990 Methods for Predicting Rates of Inbreeding in Selected Populations. *Theor. Appl. Genet.* **80**: 503-512.

Chapter 7

In Silico Genotyping Using Long-Range Phasing in a Complex Pedigree

Hans D. Daetwyler[1,2], George R. Wiggans[3], Ben J. Hayes[4], John A. Woolliams[1] and Mike E. Goddard[4,5]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin, Midlothian, EH25 9PS, United Kingdom; [2]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands; [3]Animal Improvement Programs Laboratory Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA; [4]Biosciences Research Division, Department of Primary Industries Victoria, Bundoora 3083, Australia; [5]Faculty of Land and Food Resources, University of Melbourne, Parkville 3010, Australia.

ABSTRACT

Related individuals share potentially long chromosome segments which trace to a common ancestor. We describe a chromosomal phasing algorithm (ChromoPhase) which utilises this characteristic of finite populations to phase large sections of a chromosome. In addition to phasing, our method can impute missing genotypes in individuals genotyped at lower marker density when more densely genotyped relatives are available. ChromoPhase uses a pedigree to collect an individual's surrogate parents and offspring and then cycles through these relatives one at a time to find shared chromosome segments. Once a segment has been identified, any missing information in the proband can be filled in with information from the relative. We tested ChromoPhase in a simulated population consisting of 5 generations and 600 individuals across generations at a marker density of 5.5 times the effective population size per Morgan. The percentage of correctly phased loci was high and ChromoPhase correctly imputed a high percentage of missing genotypes. Performance was marginally reduced when the proportion of genotypes missing increased and considerably reduced as the number of generations available in the pedigree decreased. Our results demonstrate that imputation of missing genotypes, and potentially full genome sequence, using long-range phasing is feasible.

INTRODUCTION

High density single nucleotide polymorphism (SNP) arrays are now available for many species. The genotypes resulting from high throughput methods are unphased and, therefore, the paternal or maternal source of each allele is unknown. Knowledge of parental origin or haplotype information can be useful in the analysis of complex traits, such as quantitative trait loci (QTL) detection (e.g. Meuwissen & Goddard 2000) and genomic selection (e.g. Meuwissen *et al.* 2001; Calus *et al.* 2008; Villumsen & Janss 2009). Many methods for resolving haplotypes have been proposed and they fall into two broad categories: those that use known relationships between individuals to perform a linkage analysis and those that rely on linkage disequilibrium among the SNP in a population without known relationships.

The first category can be subdivided into those that use likelihoods (e.g. peeling (Elston & Stewart 1971; Janss *et al.* 1995) and Lander and Green algorithm (Lander & Green 1987)) and those that are rule-based or use parsimony approaches. Rule-based approaches use genotyped parents and progeny (i.e. trios) and neighbouring loci to resolve the phase (e.g. NejatiJavaremi & Smith 1996; Pong-Wong *et al.* 2001; Qian & Beckmann 2002; Baruch *et al.* 2006). In the case where no parent genotype information is available, genotyped progeny may aid in haplotyping dense (Windig & Meuwissen 2004) and sparse marker maps (Weeks *et al.* 1995). Examples of the second category include PHASE (Stephens *et al.* 2001; Stephens & Donnelly 2003), fastPHASE (Scheet & Stephens 2006), HAPLOTYPER (Niu *et al.* 2002), BEAGLE (Browning & Browning 2009) and an approach by Schouten *et al.* (2005), where the latter two accomodate a mix of relationship and population data. Non-likelihood methods also exist for unrelated individuals, for instance methods using parsimony identify unambiguous haplotypes from individuals which are homozygous at all loci considered, and then add additional haplotypes from individuals with one heterozygous genotype (Clark 1990; Tier 2006). A crucial computational issue is that the potential number of unique haplotypes increases exponentially as more loci are considered (i.e. 2^N, where N is the number of loci). This drastically slows most methods if there are many individuals and many loci, and this is exactly the type of data which is most useful and now available.

Population characteristics such as geographical proximity can result in a high probability that individuals within a given population share a common ancestor not many generations ago. Similarly, in commercial animal populations selective breeding has reduced effective population sizes by limiting the number of parents, again causing individuals to share one or more common ancestors in the last few generations. If individuals share a common ancestor n generations ago, they are likely to have a shared chromosome segments of average length $1/n$ Morgans. With dense genotyping of markers, these segments will contain many markers and so it should be possible to recognise them and distinguish them from short segments that are identical-by-state (IBS) but do not trace to the common ancestor, without complex likelihood calculations. These observations lead to new approaches to phasing haplotypes which are based on the premise that if a large section of two gametes is IBS then there is a high probability that this section originated in a common ancestor (Kong *et al.* 2008).

Kong et al. (2008) called their method long range phasing but the principle can also be used to impute and phase missing genotypes or even to impute genotypes on individuals that have not been genotyped at all. One particularly useful application might be to impute dense genotypes on individuals with sparse genotypes using dense genotype information on their relatives. In the extreme, full genome sequences could be imputed for individuals which have been genotyped at moderate density, provided they had enough relatives that had been fully sequenced (Goddard 2008).

Here we describe a computationally efficient algorithm (ChromoPhase) that can phase whole chromosomes. We use a similar approach to that of Kong et al. (2008), but whereas they focus on genotypes, we use haplotypes more explicitly. We also use pedigree to identify whether a relative is likely to share a part of an individual's paternal or maternal chromosome. Our approach should be faster than that of Kong et al. (2008) because only known relatives are compared, more accurate because runs of genotypes that are alike by chance are less likely to be accepted as true shared segment, and more flexible in dealing with missing genotypes and completely ungenotyped individuals. This includes phasing of founders in a pedigree.

METHODS

ChromoPhase relies on the same principle as Kong et al. (2008) in that it makes use of the potentially long chromosome segments which related animals share. These segments are particularly long when individuals are closely related, as during meiosis the probability of a recombination is approximately one per Morgan. Therefore, with dense marker genotypes, the phase can be established by comparing an individual to close relatives.

We assumed bi-allelic loci with a reference allele coded 0 and an alternative allele coded 2. Genotypes were coded 0, 1, and 2, corresponding to 00, 02, and 22 respectively. Missing alleles and genotypes are assigned '5'. Loci are expected to be dense enough so the risk of double recombinations between adjacent loci may be neglected. The algorithm consists of three stages. In the first stage, potential sources of shared chromosome segments are identified using a pedigree. Secondly, rule based allele assignments are made per locus based on parents, offspring and mates. The third stage consists of an iterative process in which each individual is phased or imputed (i.e. each individual is considered to be the proband once per iteration starting with the oldest animals) and compared to related individuals to find unbroken strings of matching alleles on their respective chromosomes. Missing alleles in the proband within such shared chromosome segment are filled in with the information from the relative. We will describe all processes for the paternal side of the pedigree but the maternal side is treated in the same manner.

Stage 1, Information Sources: Pedigree and genotype data is read and ungenotyped individuals are removed unless they have at least one genotyped progeny, because ungenotyped individuals add no information unless they connect genotyped individuals. Molecular genotyping errors are checked at each locus by identifying where the proband i genotype (g_i) is inconsistent with the father genotype (g_f) (e.g. g_i = 0 and g_f = 2). For each individual, considered in turn as the proband, three sets of relatives are defined. The first set (o_i) consists of all offspring of the proband and these are collected starting with the youngest individual. The second set, called surrogate fathers (sf_i), consists of individuals related to the proband through his or her father. If the father is genotyped, then only the father is included in sf_i because the information from more distant relatives will pass through the father to the proband during iteration. In a proband with an ungenotyped father, then the set of surrogate fathers of the proband comprises its father's sets of offspring (except the proband), surrogate fathers and surrogate mothers.

Stage 2, Single locus, rule-based allele assignment: ChromoPhase applies rule-based allele assignment to the paternal or maternal gamete if they can be unambiguously resolved based on an individual's own known genotype, parental alleles or offspring alleles (e.g. Pong-Wong *et al.* 2001; Baruch *et al.* 2006). All paternal (pa) and maternal alleles (ma) in all individuals are set to 5 (missing) at the start. Then the following rules are applied starting with the oldest individual. If the proband genotype (g_i) is homozygous, then both its paternal pa_i and maternal alleles ma_i equal g_i. If both alleles of the father are known and equal (i.e. g_f is homozygous) then pa_i is the same as the paternal allele of the father pa_f. If a proband genotype, g_i, is missing, but its paternal allele is known and its maternal allele is missing, then if an offspring paternal allele, pa_o, is known and opposite to the proband's known allele (pa_i), the proband maternal allele (ma_i) equals the offspring's paternal allele (pa_o).

Table 1. Rules for matching alleles one locus at a time between proband and surrogate father, where d is a distinguishing match, ? is a inconclusive match, + is a conclusive match, E is a definite non-match which ends a segment, alleles coded 0, 2 and 5 for missing.

Proband	Surrogate Father				
Paternal Allele	Paternal Allele	Maternal Allele	Match	Conclusive	Distinguishing
0	0	0	both	+	no
0	0	2	paternal	+	d
0	2	2	no	no	E
5	0	0	?	?	no

Stage 3, Comparison of relatives: An iterative process follows in which each individual is considered as the proband once per iteration starting at the top of the pedigree. Single locus, rule-based filling of alleles (stage 2) continues at the start of each iteration and proband as more information becomes available. ChromoPhase then compares each proband to each of its relatives in the three sets (sf_i, sm_i and o_i) one locus at a time to identify shared chromosome segments consisting of a consecutive string of matching loci. We will describe the different types of matches by using an example where the proband is compared to a surrogate father. Consider proband i whose pa_i is compared to both alleles of a surrogate father (pa_{sf} and ma_{sf}) at one locus, as illustrated in Figure 1. This comparison yields one of four outcomes defined in Table 1. A conclusive match (+ in Figure and Table 1) occurs when pa_i is not missing and $pa_i = pa_{sf}$ or $pa_i = ma_{sf}$ or both. A distinguishing match (d in Figure and Table 1) is defined as $pa_i = pa_{sf}$ or $pa_i = ma_{sf}$ but not both. Thus, a distinguishing match is also a conclusive match, but in a distinguishing match the source of pa_i can be clearly determined and it is used to define the start and end of a shared segment to reduce errors. Missing information in pa_i, pa_{sf} or ma_{sf} counts as an inconclusive match (? in Figure and Table 1) which is not allowed to end a shared segment. Information is stored on which surrogate father allele (pa_{sf} or ma_{sf}) was a distinguishing match with the proband at the last locus. A definite non-match (E in Figure and Table 1) occurs when pa_i is not equal to the surrogate father's allele found on the chromosome which matched at the last distinguishing locus. A minimum length of 50 consecutive matching (i.e. conclusive, inconclusive, and distinguishing) loci between two distinguishing matches was required to accept a shared segment. Within that run, the number of loci with conclusive matches needed to exceed 40. The minimum number of conclusive matches guards against too many missing loci being counted as matches within a considered segment. Requiring longer segments will reduce errors but it will also result in fewer phased or imputed loci. The minimum length required for segments can be adapted to suit a dataset and will depend on marker density.

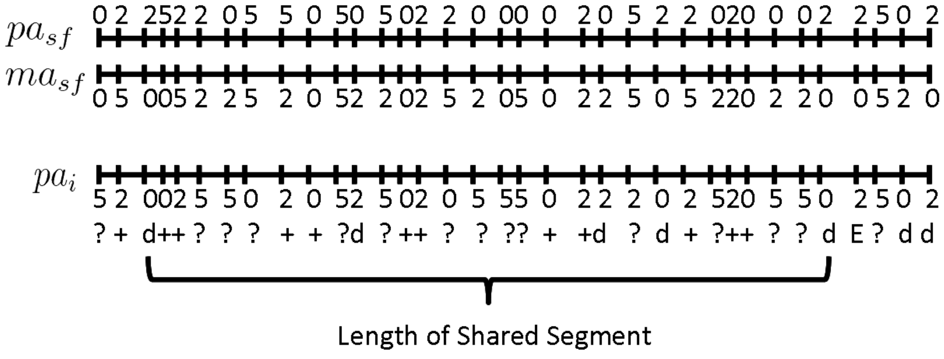


Figure 1. An example comparison of alleles on the proband's paternal gamete pa_i to alleles on both surrogate father gametes (pa_{sf} and ma_{sf}) to identify a shared chromosome segment, where d is a distinguishing match, ? is a inconclusive match, + is a conclusive match and E signifies a definite non-match which ends a segment. There is a shared segments between pa_i and ma_{sf}.

We will now describe the comparison at consecutive loci of each proband with its two different groups of relatives; surrogates, sf_i and sm_i, and offspring, o_i. First, the proband is compared to all of its surrogates. If the proband is genotyped, pa_i is compared to pa_{sf} and ma_{sf}. Otherwise, if the proband is not genotyped, proband offspring paternal allele pa_o is compared to pa_{sf} and ma_{sf}. Whenever a shared segment has been identified and exceeds the minimum number of matches required, the information from all surrogate comparisons is collectively stored as a count of allele 2 per locus. Once all the comparisons to surrogate fathers are completed pa_i is filled in based on this information. Surrogates do not always agree on a particular allele at a locus. Therefore, proband alleles are filled in based on the collective information from all surrogates exceeding a threshold. Allele 2 is assigned if the ratio of allele 2 counts over the total number of counts from all surrogates is 0.9 at a particular locus, and assigned the 0 allele if this ratio is less than 0.1. Equal weight is given to information from different surrogate fathers, if there are more than one. Hence, the 10th surrogate father may contribute as much information as the first, irrespective of degree of relationship.

The proband is then compared to its offspring (e.g. pa_i and ma_i to offspring pa_o) to fill in remaining missing alleles. Here, a distinction is made between genotyped and ungenotyped probands. In genotyped individuals, threshold criteria for shared segments (e.g. number of conclusive matches) are as in surrogate father comparisons. In ungenotyped individuals and genotyped founders, filling in of pa_i or ma_i is more liberal and any missing proband alleles within a shared segment are filled in with information from pa_o once a string of more than 50 matches has been identified (i.e. no threshold is specified for the number of missing alleles). In addition, recombinations are mapped on the chromosome. In founders, pa_i is arbitrarily filled in first with information from pa_o because chromosomes cannot be differentiated.

Iterations end when no more alleles have been changed or when the maximum number of iterations specified has been reached.

Simulations for Testing

Populations and Genome: Populations in mutation drift equilibrium were simulated by randomly mating individuals for 1000 generations with recombination and mutation. Effective population size (N_e) was 200 and the number of male and female parents was equal across generations. Previous work established that with this N_e mutation drift equilibrium was achieved with 1000 generations. One male and one female offspring were produced per mating. Pedigree and genotype information was retained for individuals in the last five generations. In generation 996 through 999, 100 individuals were simulated and generation 1000 consisted of 200 individuals for a total of 600 individuals.

One chromosome was simulated measuring one Morgan. In generation zero all individuals were completely homozygous for the same allele and mutations were applied at a rate of 2.5 $*10^{-5}$ per locus per meiosis in the following generations. Mutations switched allele one to two and vice versa. The number of mutations and recombinations per chromosome were sampled from a Poisson distribution. The mean for the number of mutations corresponded to the product of the number of loci per chromosome (both monomorphic and polymorphic) and the mutation rate, and the mean for recombinations was one per Morgan. The number of sampled mutations and recombinations were then randomly placed on the chromosome. Approximately 1100 segregating bi-allelic loci were present at generation 1000, which is equivalent to a density of $5.5Ne$ per Morgan. Two cases were considered, one where all loci were included and another where loci were selected to exceed 0.02 minor allele frequency (MAF). Discarding low MAF loci is expected to make our tests more conservative because high homozygosity is favourable to our approach and it is also meant to mimic application in real data where low MAF loci are often discarded as part of quality control measures. Linkage disequilibrium (LD, R^2) statistics (Hill & Robertson 1968) between adjacent segregating loci were averaged among all pairs exceeding a MAF of 0.05 and matched expected R^2 values (Sved 1971; Tenesa *et al.* 2007). Allele frequency was found to follow a U-shaped distribution as expected.

Testing: The utility of ChromoPhase was evaluated in the simulated data described above. Phasing was tested in five replicates of data with unselected loci and in five replicates of the data where loci were selected to have MAF > 0.02. Phasing utility was checked within each replicate in the whole five generation dataset, and in three subsets of the data consisting of the last four, three and two generations in the data set. The pedigree used by the program was restricted to the generations being tested. Hence, no additional information was available on ancestors beyond the animals in dataset. Inferred alleles were compared to true alleles and this yielded the following test parameters for both paternal and maternal alleles, i) percent correct, ii) percent missing, iii) percent wrong. In addition, alleles which are inconsistent between offspring and parents were counted as conflicts (e.g. $pa_i \neq pa_f$ or ma_f).

Imputation of missing genotypes was evaluated in two population replicates with loci < 0.02 MAF removed by setting the genotypes at a specific locus as missing for a random proportion of individuals in the last generation, with 0.2, 0.5 and 0.99. Each locus, in turn, was assessed this way. This was repeated in the last four, three, and two generations to investigate how ChromoPhase copes with varying depths of pedigree. This resulted in nine scenarios per locus (i.e. three proportions and three pedigree depths) and for each scenario five replicates were computed to reduce variability. Initial results for imputation in the whole dataset were similar to tests in the last four generations and were therefore omitted. The same test parameters were collected for imputation as in phasing. In addition, information on imputed genotypes was enumerated as, i) percent correct, ii) percent missing and iii) percent wrong when compared to true genotypes.

RESULTS

Phasing: Phasing was evaluated in all animals including founders and results can be found in Tables 2 and 3. The percentage of alleles phased correctly in the replicates selected for MAF > 0.02 (Table 2) when compared to true alleles was high, ranging from 98.1% when all generations were available to 94.5% when only 2 generations were included. Errors decreased as the number of generations increased demonstrating that ChromoPhase makes use of information more than one generation removed. Another possible reason for decreased performance in 2 generations is that in founders, the proband’s paternal allele is arbitrarily filled in with progeny information initially because distinguishing between founder chromosomes is not possible. Once more paternal alleles are filled in, it becomes possible to distinguish between chromosomes in founders. However, whole chromosomes could be switched in founders, as paternal or maternal origin cannot be assigned. It is therefore expected that founders are the source of the majority of phasing errors. For example, when a two generation pedigree is used, errors in founder are expected to be approximately 3.7% due to arbitrary assignment to paternal alleles. Such differences are not a problem as they do not reduce the accuracy of phasing progeny, and, as founders, they have no parental genotypes to conflict with. In data without selection based on MAF (Table 3) the percentage phased correctly is increased and errors are reduced as expected due to increasing locus homozygosity.

Table 2. Performance of ChromoPhase in percent in paternal (p) and maternal (m) alleles of all individuals, including founders, when the dataset consisted of the last 5, 4, 3, or 2 generations (Gen.) of genotyped animals. Conflicts refer to inconsistent genotypes between parents and offspring. Loci with minor allele frequency < 0.02 were removed, means of 5 population replicates and SE was < 2% in all scenarios.

	Correct		Missing		Wrong		Conflicts
Gen.	p	m	p	m	p	m	Genotypes
5	98.1	98.1	0.0	0.0	1.9	1.9	0.0
4	97.6	97.6	0.0	0.0	2.3	2.3	0.0
3	96.8	96.8	0.1	0.1	3.1	3.1	0.0
2	94.5	94.5	0.6	0.6	5.0	4.9	1.9

Imputation: Similar trends to phasing where observed in imputation. Table 4 shows means across all loci of imputed genotypes when a proportion of individuals in the last generation were set to missing one locus at a time. The percentage of correctly imputed genotypes was greatest when 4 generations of data was available. The number of missing genotypes increased as the number of generations decreased. However, wrongly imputed genotypes stayed constant at 0.2% in both 4 and 3 generations, but increased in the 2 generation data. In the 4 generation dataset, performance was stable as the proportion of individuals set to missing increased. A decrease in correct imputation was apparent with increasing proportions of genotypes missing only when 3 or 2 generations were included in the dataset. This suggests that when a sufficiently deep pedigree of genotyped ancestors is available the approach becomes more robust.

Table 3. Performance of ChromoPhase in percent of phasing paternal (p) and maternal (m) alleles of all individuals, including founders, when the dataset consisted of the last 5, 4, 3, or 2 generations (Gen.) of genotyped animals. Conflicts refer to inconsistent genotypes between parents and offspring. All loci included, means of five population replicates and SE < 1% in all scenarios.

	Correct		Missing		Wrong		Conflicts
Gen.	p	m	p	m	p	m	Genotypes
5	98.5	98.5	0.1	0.1	1.4	1.4	0.0
4	98.1	98.1	0.1	0.1	1.8	1.8	0.0
3	97.8	97.8	0.1	0.1	2.2	2.2	0.0
2	95.8	95.8	0.5	0.5	3.8	3.8	1.5

Means of correctly imputed loci within centiMorgan brackets are shown in Table 5 and for one scenario in Figure 2. This was to test the utility of ChromoPhase at various locations across a chromosome. Percent correctly imputed was lower at the beginning and end of the chromosome, but maximum performance was reached relatively quickly at approximately 10cM and was sustained until approximately 90cM (Figure 2). The reduced performance likely stems from the reliance on shared segments in our approach. If a proband has multiple surrogates these segments may overlap and opportunities for filling in missing information increase. However, at the beginning and end of a chromosome it is less likely that shared segments overlap because a locus with a distinguishing match must be found to start and end a segment.

Table 4. Overall performance of ChromoPhase in percent when imputing genotypes and paternal alleles in a proportion (Prop.) of individuals in the last generation one missing locus at a time in the last generation when 4, 3 or 2 generations (Gen.) are included in the dataset. Means of across all loci from two population replicates.

		Genotypes			Paternal Alleles		
Gen.	Prop.	Correct	Missing	Wrong	Correct	Missing	Wrong
4	0.20	98.7	1.2	0.2	99.3	0.6	0.2
4	0.50	98.6	1.2	0.2	99.2	0.6	0.2
4	0.99	98.5	1.3	0.2	99.2	0.6	0.2
3	0.20	98.3	1.5	0.2	99.0	0.8	0.2
3	0.50	97.7	2.1	0.2	98.7	1.1	0.2
3	0.99	94.9	5.0	0.2	97.2	2.6	0.2
2	0.20	87.1	8.0	5.0	92.6	4.7	2.7
2	0.50	83.2	12.1	4.7	89.8	7.6	2.6
2	0.99	62.2	37.6	0.2	77.1	22.8	0.1

Similarly to grouping loci within centiMorgan categories, loci were also grouped according to heterozygosity and these group means are presented in Table 6 and Figure 3. Correct imputation was very high when heterozygosity was low and showed a decreasing trend as heterozygosity increased. This trend became more pronounced as fewer generations were represented in the dataset. ChromoPhase was most robust to changes in heterozygosity and changes in proportion of individuals set to missing at a locus when 4 generations of genotype data were available. Correct imputation was significantly above imputation which could achieved via inspection of parental genotypes which is shown in Figure 3 as the probability that both parent genotypes are homozygous for a given locus heterozygosity.

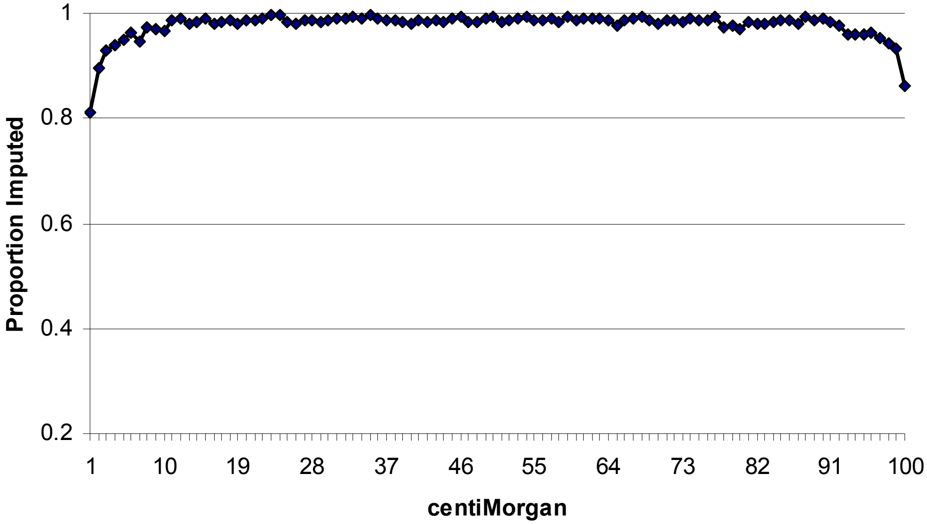


Figure 2. Proportion of genotypes imputed correctly in last generation when 0.5 of genotypes were set to missing one locus at a time and 3 generations were included in the dataset. Presented in means per centiMorgan. Mean of two population replicates.

DISCUSSION

We have described a long-range phasing and imputation algorithm, ChromoPhase. Testing of ChromoPhase has been shown to be highly accurate for both pure phasing of genotyped loci and for imputation of missing genotypes. Our approach seeks out and phases long chromosomal segments which are shared between relatives. This is a significant improvement over other approaches which phase per locus or only consider a few loci at a time (e.g. Schouten et al. 2005).

The identification of shared chromosome segments is also the key to imputing genotypes, as any missing information within a segment can potentially be filled in the proband with information from its relative. The key aspect of identifying a shared segment is the recognition that if two animals share a haplotype over many continuous loci, then the probability that this haplotype coalesces to a common ancestor becomes high.

Table 5. Performance of ChromoPhase in percent correct when imputing genotypes in a proportion (Prop.) of individuals in the last generation one missing locus at a time when 4, 3 or 2 generations (Gen.) are included in the dataset. Means of two population replicates within centiMorgan bins.

		centiMorgan						
Gen.	Prop.	1 - 2	3 – 5	6 – 10	11 – 90	91 - 95	96 - 98	99 - 100
4	0.20	86.0	96.0	98.0	99.4	97.7	95.7	93.0
4	0.50	86.3	96.0	97.9	99.4	97.6	95.9	93.2
4	0.99	85.5	95.8	97.8	99.3	97.6	95.8	93.1
3	0.20	86.2	94.6	97.3	99.2	97.3	95.4	90.1
3	0.50	85.2	93.9	96.3	98.6	96.8	95.2	89.8
3	0.99	83.0	89.7	91.9	96.0	94.5	93.0	88.6
2	0.20	80.7	86.7	88.2	88.2	80.1	81.2	75.8
2	0.50	77.4	82.8	83.6	84.3	76.7	78.8	73.2
2	0.99	59.8	61.0	58.0	63.3	59.1	63.8	57.7

Kong et al. (2008) made use of this concept in individuals of unknown relationship by searching for a sufficiently long stretch of loci with no incompatible genotypes that can therefore be assumed to have originated in a common ancestor. All potential surrogates of a proband for a genome segment of predefined length were identified and stored at the beginning of their algorithm. They then phased a proband by cycling through its surrogates to identify a homozygote at a particular locus. Our approach is similar, but operationally different, as we chose to work within a pedigree and thus we are able to compare alleles within family relationships, and to compare at the level of the allele instead of genotypes. Our algorithm compares relatives in each iteration (not just at the start, as in Kong et al. 2008) to make use of new information as it becomes available and we do not specify a maximum length for shared segments. Thus, a shared segment may potentially span the full chromosome and allows us to use all available information. In addition, we limit the list of surrogate fathers (mothers) to the nearest relative in each line of the pedigree that has been genotyped. If the father has been genotyped, no other surrogate fathers are included, because the parental genotype contains the necessary information. This reduces the number of comparisons that must be made because information from more distant relatives is transmitted through the pedigree with iteration. Consequently ChromoPhase uses all the information used by Kong et al. (2008) but uses some additional information. For instance, using distinguishing matches to define the start and end of a matching chromosome segment eliminates some errors.

Table 6. Performance ChromoPhase in percent correct when imputing genotypes in a proportion (Prop.) of individuals in the last generation one missing locus at a time when 4, 3 or 2 generations (Gen.) are included in the dataset. Means of two population replicates within 0.1 heterozygosity bins.

		Heterozygosity				
Gen.	Prop.	0.01 - 0.10	0.11 – 0.20	0.21 – 0.30	0.31 – 0.40	0.41 – 0.50
4	0.20	99.6	99.1	98.7	98.1	97.4
4	0.50	99.6	99.1	98.7	98.0	97.3
4	0.99	99.6	99.1	98.6	97.9	97.1
3	0.20	99.6	99.1	98.5	97.5	96.4
3	0.50	99.6	99.0	98.0	96.6	94.8
3	0.99	99.5	98.0	95.7	92.6	87.5
2	0.20	97.0	91.6	86.6	80.8	74.6
2	0.50	96.0	88.8	82.4	74.7	67.5
2	0.99	89.9	72.8	58.2	44.0	31.2

The program is computationally fast and one run with 600 genotyped animals required approximately 15 seconds. Approximately six iterations were required and running time was not significantly affected whether imputation was included or not.

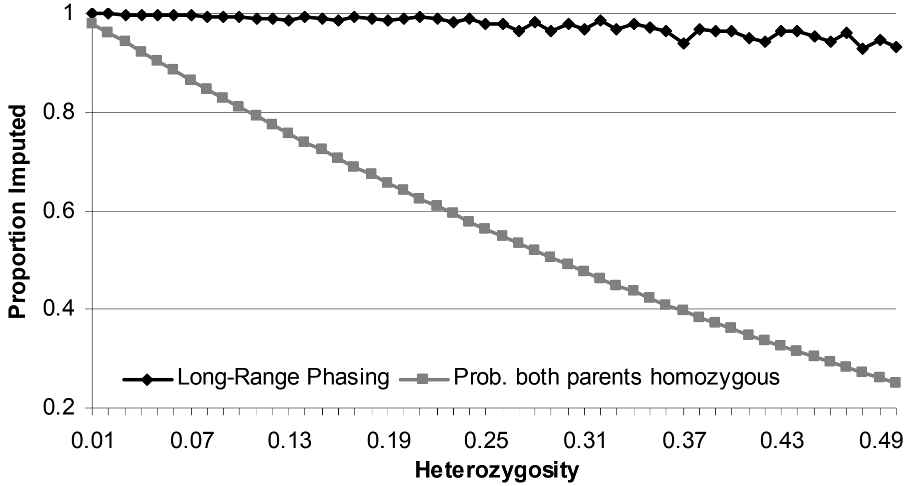


Figure 3. Proportion of genotypes imputed correctly in last generation when 0.5 of genotypes were set to missing one locus at a time and 3 generations were included in the dataset. Presented in means per 0.01 locus heterozygosity. Means of two population replicates and grey series is the probability that both parents are homozygous at a particular heterozygosity.

Our use of pedigree information to identify surrogates is limiting when such information is not available. Even though the relationship between animals without recorded pedigree relationship is unknown, they are still expected to share chromosome segments, albeit shorter than segments between close relatives, by virtue of being part of the same population. Restricting comparisons to known relatives reduces computer time and erroneous matches but also loses information, especially if the pedigree is incomplete. ChromoPhase could be modified to make comparisons among all founders if this loss of information was too great.

Currently the algorithm applies to autosomes and further modification to sex chromosomes may be necessary. Recombination occurs freely between X chromosomes hence, phasing involving females is expected to be the same as autosomes. Simplification may be possible in males since it should be straight forward to distinguish between X and Y chromosomes. Pseudoautosomal recombination between X and Y is believed to be restricted to the relatively short regions at either end of the X chromosome in human representing approximately 2% of all bases in total (Charlesworth 1991; Ross *et al.* 2005). Therefore it may possible to ignore recombination between the pseudoautosomal regions in sex chromosomes of males.

The comparison of haplotypes in our algorithm also results in computational efficiency because the same process is used for phasing and imputation. The main objective of ChromoPhase is to complete as much information as possible in a proband haplotype by using information from shared segments with relatives. It is therefore irrelevant from the method's point of view whether this is for phasing or imputation, though the algorithm benefits when genotypes are available at a locus. Furthermore, comparison at the level of the allele allows for more exact determination of what constitutes a matching segment. This results in more power than determining IBS on a genotype level (e.g. Kong et al. 2008). Furthermore, we start and end segments with distinguishing matches at a locus in which a proband allele matches one surrogate allele but not the other. The increase in power allowed for the minimum number of allele matches to be shortened substantially to 50 when compared to Kong et al. (2008) which used 1000 consecutive genotypes. Part of this reduction can be attributed to Kong et al. (2008) having a marker density which was 3 times higher, as a product of N_e, than the one used in this study. Our algorithm showed good performance with an additional reduction to the minimum length of shared segments of approximately 17 times after accounting for marker densities. In addition, our approach copes with missing alleles, but to guard against errors we have implemented a threshold for the amount of missing alleles allowed in a shared chromosome segment.

Distinguishing shared segments depends crucially on filling in as much information as possible with rule-based methods. This is especially important in early iterations as comparatively little information is available. Rule-based filling is difficult in individuals without parent information and therefore a disproportionate amount of errors are likely to occur in founders. However, in general, phasing in the last few generations is of greater interest than in founders and our method seems robust as such errors in founders do seem less likely to be transferred to younger individuals.

We have made conservative assumptions in testing ChromoPhase. First, imputation results were only reported in the dataset where loci with MAF < 0.02 removed, which resulted in a more rigorous test of the algorithm. Secondly, we used the percentage of correctly imputed genotypes as the crucial test parameter in most of our tables and figures. As can be seen in Table 4, the number of imputed alleles exceeds the number of correct genotypes and this trend becomes more substantial when fewer generations are included in the data. Partial haplotype information may however still be useful in, for example, genomic selection.

The application of our method in real data sets will require addressing several challenges, such as completely ungenotyped animals in the data, incomplete pedigrees, genotyping errors and SNP mapped to wrong genome locations. Currently, completely ungenotyped individuals are retained if they connect genotyped individuals, and their haplotypes are attempted to be phased like other individuals. This works satisfactorily if the ungenotyped individual has genotyped ancestors and descendents in the data (results not shown), but is problematic when an individual has no genotyped ancestors. It is also important that correct and as complete as possible pedigree information is available for determining surrogates. Most genotyping errors can be detected by comparing trios though if they are not detected then they may result in erroneous haplotype assignments. Map errors may cause a wrongly mapped locus to appear shared between relatives where it may only be a match by chance (i.e. it is only IBS) causing phasing errors.

A likely application of ChromoPhase is to impute haplotypes and genotypes in individuals which are genotyped at lower SNP density when relatives are available which are genotyped at higher density. The current study confirms that this would be possible when probands are genotyped at a density of $5.5N_e$ per Morgan and denser genotypes are available on relatives. Thus, it should be feasible to impute genotypes in individuals which are genotyped at 50K once information from denser SNP chips becomes available in their relatives. It may even be possible to move to sparser than $5.5N_e$ proband genotyping, but this needs further investigation. In contrast, there is no upper limit to how dense the genotypes can be for successful imputation, and even imputing full genome sequence data will eventually be feasible once sufficient ancestors have been sequenced.

The potential for ChromoPhase to increase the number of genotyped individuals while simultaneously reducing genotyping costs is very large. Key benefits will be increased sample sizes to achieve higher accuracies in genomic selection and to increase the power of QTL studies. Reducing genotyping costs through strategic genotyping of ancestors and upgrading to denser genotyping from sparser SNP chips in the current generation with ChromoPhase will allow for the application of genomic selection in species where currently this technology is not economically feasible

ACKNOWLEDGEMENTS

The authors thank Beatriz Villanueva for helpful comments on this manuscript. HDD was supported by the SABRETRAIN Project, which is funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission. Part of this work was accomplished during research stays of HDD and GRW at the Department of Primary Industries Victoria, Bundoora, Australia. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk).

REFERENCES

Baruch, E., J. I. Weller, M. Cohen-Zinder, M. Ron, and E. Seroussi, 2006 Efficient inference of haplotypes from genotypes on a large animal pedigree. *Genetics* **172**: 1757-1765.

Browning, B. L., and S. R. Browning, 2009 A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Gen.* **84**: 210-223.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**: 553-561.

Charlesworth, B., 1991 The Evolution of Sex-Chromosomes. *Science* **251**: 1030-1033.

Clark, A. G., 1990 Inference of Haplotypes from Pcr-Amplified Samples of Diploid Populations. *Mol. Biol. and Evol.* **7**: 111-122.

Elston, R. C., and J. Stewart, 1971 General Model for Genetic Analysis of Pedigree Data. *Hum. Heredity* **21**: 523-&.

Goddard, M. E., 2008 The use of high density genotyping in animal health, pp. 383-389 in *Animal Genomics for Animal Health*, edited by M.-H. Pinard, P.-P. Pastoret, and B. Dodet. Karger, Basel.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. *Theor.Appl.Genet.* **38**: 226-231.

Janss, L. L. G., J. A. M. Van Arendonk, and J. H. J. VanderWerf, 1995 Computing approximate monogenic model likelihoods in large pedigrees with loops. *Genet. Sel. Evol.* **27**: 567-579.

Kong, A., G. Masson, M. L. Frigge, A. Gylfason, P. Zusmanovich *et al.* 2008 Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics* **40**: 1068-1075.

Lander, E. S., and P. Green, 1987 Construction of Multilocus Genetic-Linkage Maps in Humans. *Proc. Nat. Acad. Sci. USA* **84**: 2363-2367.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**: 421-430.

NejatiJavaremi, A., and C. Smith, 1996 Assigning linkage haplotypes from parent and progeny genotypes. *Genetics* **142**: 1363-1367.

Niu, T. H., Z. H. S. Qin, X. P. Xu, and J. S. Liu, 2002 Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Gen.* **70**: 157-169.

Pong-Wong, R., A. W. George, J. A. Woolliams, and C. S. Haley, 2001 A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genet. Sel. Evol.* **33**: 453-471.

Qian, D. J., and L. Beckmann, 2002 Minimum-recombinant haplotyping in pedigrees. *Am. J. Hum. Gen.* **70**: 1434-1445.

Ross, M. T., D. V. Grafham, A. J. Coffey, S. Scherer, K. Mclay *et al.* 2005 The DNA sequence of the human X chromosome. *Nature* **434**: 325-337.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Gen.* **78**: 629-644.

Schouten, M. T., C. K. I. Williams, and C. S. Haley, 2005 The impact of using related individuals for haplotype reconstruction in population studies. *Genetics* **171**: 1321-1330.

Stephens, M., and P. Donnelly, 2003 A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Gen.* **73**: 1162-1169.

Stephens, M., N. J. Smith, and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Gen.* **68**: 978-989.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Pop.Biol.* **2**: 125-141.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.* 2007 Recent human effective population size estimated from linkage disequilibrium. *Gen. Research* **17**: 520-526.

Tier B. Haplotyping for linkage disequilibrium mapping. 8th World Congress of Genetics Applied to Livestock Production. Proc.of the 8th WCGALP. CD-ROM Commun. 2006.

Villumsen, T. M., and L. Janss, 2009 Bayesian genomic selection: the effect of haplotype length and priors. *Bmc Proc* **3 Suppl 1**.

Weeks, D. E., E. Sobel, J. R. Oconnell, and K. Lange, 1995 Computer-Programs for Multilocus Haplotyping of General Pedigrees. *Am. J. Hum. Gen.* **56**: 1506-1507.

Windig, J. J., and T. H. Meuwissen, 2004 Rapid haplotype reconstruction in pedigrees with dense marker maps. *J. Anim. Breed. Genet.* **121**: 26-39.

Chapter 8

General Discussion

Hans D. Daetwyler[1,2]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin, Midlothian, EH25 9PS, UK; [2]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6700 AH, The Netherlands

This thesis has primarily sought to increase the understanding on genetic evaluation of populations using genomic data. The content reflects the increasing speed with which new methods are embraced and applied in real populations in response to new data types, such as a large number of single nucleotide polymorphisms (SNP), becoming available. Currently, the animal breeding community is in the midst of a paradigm shift concerning the analysis of genomic data and incorporation of results into breeding programs. The two main drivers of this shift have been a method proposed by Meuwissen et al. (2001) called genome-wide evaluation (GWE) and the development of dense SNP chips. Genome-wide methods require large numbers of markers to be effective and, in turn, information from dense SNP can be quickly applied through GWE. Therefore, it is the synergy of the two drivers that allows for wide ranging changes in perspectives for animal breeding programs. The rate of adoption of GWE is however different in various species. The dairy cattle industry has moved quickly to implement GWE because large improvements in genetic gain are possible and cost savings in progeny testing of bulls could easily be identified. Other species, such as swine and poultry are moving more slowly towards this implementation, because the expected genetic gain from GWE in these species is perhaps less than in dairy cattle. Pre-dating this shift to GWE, the focus was on quantitative trait loci (QTL) detection and subsequent incorporation of QTL into breeding programs with marker-assisted selection. Thus, the objective pre-GWE was to mark particular DNA segments to aid in predicting part of a breeding value, whereas GWE attempts to quantify the collective contribution of all markers to predicted breeding values.

This thesis contains chapters spanning the period before and after this general shift of focus away from finding individual QTL to considering all QTL in a single evaluation step. Chapter 2 presents a genome scan for QTL detection in dairy cattle using the 10K Affymetrix bovine SNP chip, which at the time was a ‘dense’ chip. The two methods used are a variance component linkage analysis approach and an association study. In accordance with the shift to GWE of the animal breeding community, this thesis then turns its attention to GWE of populations, both animal and human. A crucial parameter used to gauge the efficacy of GWE is its accuracy, or the correlation of true and estimated breeding values, because accuracy is an important component of genetic gain. The importance of

GWE accuracy is reflected within this thesis. In fact, accuracy is the thread that ties the remaining chapters of this thesis together because it is a central component in each. In Chapter 3, the accuracy of GWE is theoretically derived for continuous and dichotomous traits in population and case control studies. The formulae are extensively tested by stochastic simulation and represent a theoretical foundation for Chapters 4 and 5. Chapter 4 investigates the impact of the genomic structure of populations and genetic trait architecture on the accuracy of GWE methods. A key conclusion of this chapter is that the relative performance of different GWE methods is heavily influenced by both population and trait genetic architecture. Equations are derived to predict the accuracy of genomic best linear prediction (GBLUP) (NejatiJavaremi *et al.* 1997; Meuwissen *et al.* 2001) and of a Bayesian variable selection approach, known as BayesB, with priors on the numbers of QTL (Meuwissen *et al.* 2001). Chapter 5 compares the accuracy predictions of Chapter 3 to those of Goddard (2008) and to empirical values from North American and Australian Holstein and Jersey data. Chapter 6 outlines the expected effect of GWE on the rate of inbreeding per generation. It concludes that selection on breeding values from GWE results in lower inbreeding levels than classical BLUP at same rate of genetic gain because of increased accuracy of Mendelian sampling terms. Chapter 7 changes focus to describe and evaluate a phasing and imputation algorithm which applies long-range phasing. A direct application of this method will be to increase sample size by imputation of genotypes to increase the accuracy of GWE per unit of genotyping cost.

This General Discussion will concentrate on four main topics, which are both theoretical and applied: i) missing genetic variance with currently available chips, ii) performance of GWE methods under different genetic architectures, iii) the impact of sequence data on GWE, and iv) challenges to implementing GWE.

The Missing Genetic Variance

A central parameter of genomic evaluation using high density SNP chips is the proportion of the genetic variance that is currently tagged by a chip. In other words, before any kind of analysis has started, what is the upper bound of accuracy that could be achieved using

GWE given a certain density of SNP? or what is the upper bound of the proportion of QTL that could be detected with a chip assuming that sample size is not a limiting factor?

The models used to investigate genomic evaluation have assumed this proportion to be one (e.g. Chapters 3 and 4 this thesis, Goddard 2008). While some acknowledge this caveat (Dekkers 2007), this issue has not been fully addressed in the literature. There are several reasons why this proportion currently does not equal one. While more and more SNP are being included on platforms, the marker density is still not high enough so that all QTL are in extensive linkage disequilibrium (LD) with at least one SNP. This is supported by the increasing accuracies demonstrated in simulated data with higher marker densities (Calus *et al.* 2008; Solberg *et al.* 2008). Furthermore, SNP with higher heterozygosity are preferentially selected for chips, introducing what is called the "ascertainment bias". While this is not a bias in the traditional breeding value estimation sense, it implicitly shifts the focus to estimating QTL with higher minor allele frequency (MAF). There were valid reasons for this selection, especially with early SNP chips developed with low genome coverage. Firstly, it can be difficult to establish if a rare allele is actually a SNP or a sequencing error. Hence, inclusion of rare SNP on chips was avoided to reduce this problem. Secondly, having higher MAF was expected to ensure that a large proportion of SNP will be segregating across breeds. Thirdly, application in QTL mapping initially involved relatively small sample sizes which only resulted in sufficient power to predict relatively common loci. In addition, SNP with low MAF are often excluded as part of extra quality control measures preceding QTL and GWE analysis to reduce genotyping errors, and this only serves to increase the problem. The effect of this difference in heterozygosity is that many rare alleles are likely missed, because rare QTL would not be in high LD with a SNP of intermediate frequency. Missing rare QTL has implications both in the short and in the long term. The short term effect is that we cannot capture all the current genetic variation, while, in the long term, genetic gain is reduced because rare alleles are not selected. Once they are at higher frequency due to selection, these rare QTL in the current generation will explain a larger proportion of the genetic variance in future generations.

The following describes theoretical ways of quantifying the proportion of the genetic variance that can currently be tagged by a SNP chip. The methodology is applicable across different SNP chips and species while the actual estimates presented are specific to the 50K Illumina SNP chip currently used in dairy cattle. The approach used is based on the accuracy of GWE in real populations and on prediction equations developed in Chapter 3.

Dekkers (2007) partitioned GWE reliability into two parts: the proportion of the total genetic variation captured by the markers (q^2) and the reliability of predicting the proportion of the variance associated with markers ($r^2_{\hat{Q}}$). This resulted in an equation for the observed reliability of a genomic breeding value (r^2_o) expressed as $r^2_o = q^2 r^2_{\hat{Q}}$. Chapter 4 proposed the following equation for the reliability of GWE, $r^2_{g\hat{g}} = (N_P h^2/Me)/(N_P h^2/Me + 1)$, where $r^2_{g\hat{g}}$ is the reliability of a genomic breeding value, N_P is the number of phenotypes, h^2 is the trait heritability, and Me is the number of independent chromosome segments. In turn, the Me used is $Me = 2NeL/ln(4NeL)$, as derived by Goddard (2008), where Ne is the effective population size and L is the genome length in Morgans. The derivation of GWE accuracy in Chapter 3 assumed that all the variation was captured by the markers and, thus, $r^2_{g\hat{g}}$ may be substituted for $r^2_{\hat{Q}}$ and rearranging gives the following expression for q^2,

$$q^2 = r^2_o / r^2_{g\hat{g}}. \tag{5}$$

Hence, we can view the proportion of the genetic variance captured by the markers as a ratio of observed and the maximum reliability achievable. Equation 1 allows for point estimates of q^2 to be made from real data.

While point estimates are valuable as a means of getting information of how much of the genetic variance is tagged at the current observed reliability, it is of interest to determine the maximum achievable q^2 (q^2_{max}) within a particular SNP chip, population and trait. In turn, this maximum will determine the upper bound of GWE accuracy achievable with a particular SNP chip.

Substituting the full formula for $r^2_{g\hat{g}}$ results in,

$$r^2_o = q^2[(N_P h^2/Me)/(N_P h^2/Me + 1)] \tag{6}$$

and inserting c for h^2/Me, gives:

$$r_o^2 = \frac{q^2 N_P c}{N_P c + 1}, \tag{7}$$

and taking its inverse results in

$$\frac{1}{r_o^2} = \frac{1}{cq^2}\frac{1}{N_P} + \frac{1}{q^2}. \tag{8}$$

This can be treated as a regression equation where $y = 1/r_o^2$, $x = 1/N_P$, with intercept $1/q_{max}^2$, and slope $1/cq^2$, where c is a constant that will depend on trait and population. Therefore by regressing $1/r_o^2$ on $1/N_P$ we can determine the q_{max}^2 possible with a particular chip and, in turn, the maximum accuracy achievable within a particular SNP chip, population and trait. When N_P is very large, the maximum achievable reliability will be equal to q_{max}^2. Note that q^2 refers to the proportion of the genetic variance captured at a particular number of phenotypes and q_{max}^2 refers to the maximum proportion of the genetic variance captured by a SNP chip. Also note that c is a constant within a trait and population and, therefore, regression estimates of q_{max}^2 do not require c to be known, whereas the point estimates using Equation (5) do require knowledge of c.

Estimation of q_{max}^2 in real data. The above equations are now used to estimate q_{max}^2 from four estimates of GWE reliability in the North American dairy cattle population. It was important that there were several estimates of reliability resulting from various numbers of phenotypes. The number of phenotypes needed to be high to ensure that the q_{max}^2 from the regression equation was not overestimated. Note that whenever the terms phenotypes or observations are used, the implicit assumption is made that phenotyped individuals are also genotyped. The USDA provided four estimates of reliability in the trait Net Merit. The data and method of GWE analysis has been described in VanRaden *et al.* (2009b) and, briefly, in Chapter 5. The method of analysis was GBLUP and SNP with a minor allele frequency of <0.05 were excluded. De-regressed estimated breeding values (EBV) of progeny tested bulls were used as observations with a reliability of 0.9, thus the h^2 used in the equations was also 0.9. The reliabilities used here included updated EBVs and one more year of predictor bulls in the estimate, hence the values for Net Merit in Table 1 slightly differ from VanRaden et al. (2009b). In addition, cows were included in estimate 3 and 4. De-regressed cow EBVs had a lower reliability than bull EBVs and, therefore, bulls and cows needed to be combined to an equivalent reliability. The contribution of the cows to the reliability was very low in estimate 3 adding 0.01 in the final value 0.59 (VanRaden *et al.* 2009a). Based on the respective bull and cow contributions, each bull contributed approximately as much as 12.4 cows to the reliability. The number of cows was therefore adjusted to "bull equivalents" by division of 12.4 to account for the difference in EBV reliability (i.e. adjusted N_P = number of bulls + number of cows/12.4).

Table 1. Description of four Net Merit GWE reliability estimates used in our calculations, resulting from GBLUP and 50K Illumina SNP chip. Adjusted cows (Adj. Cows) = cows / 12.4.

Estimate	Number of Bulls	Number of Cows	Adj. Number of Cows	Adj. Total N_P	GWE Reliability
1	2130	0	0	2130	0.48
2	3576	0	0	3576	0.58
3	4422	947	76.2	4498.2	0.59
4	7600	2711	218.3	7818.3	0.69

In trait Net Merit in North American Holsteins with 50K Illumina SNP chip, the estimate of $1/q_{max}^2$ is approximately 1.244 (SE 0.053), with 95% confidence interval (1.091, 1.398) which converts to a q_{max}^2 of 0.80, with an approximate SE of 0.034. Therefore the maximum achievable reliability that can be achieved with the current 50K Illumina chip for Net Merit in Holsteins is estimated to be 0.80. Figure 1 shows the r_o^2 which can be achieved at increasing numbers of phenotypes taking into account q^2 with the regression equation.

Point estimates of q^2 for the four reliability estimates using Equation (5) are also shown in Figure 1, in these estimates a h^2 of 0.9 and Me of 640 was used. It is clear that with the current number of observations r_o^2 has not reached asymptotic levels and, therefore, increasing sample size will yield significant benefits in increasing r_o^2. However, approaching the asymptote will require very large numbers of observations (Figure 1). The q^2_{max} estimated may decrease slightly once estimates of the reliability using more observations become available.

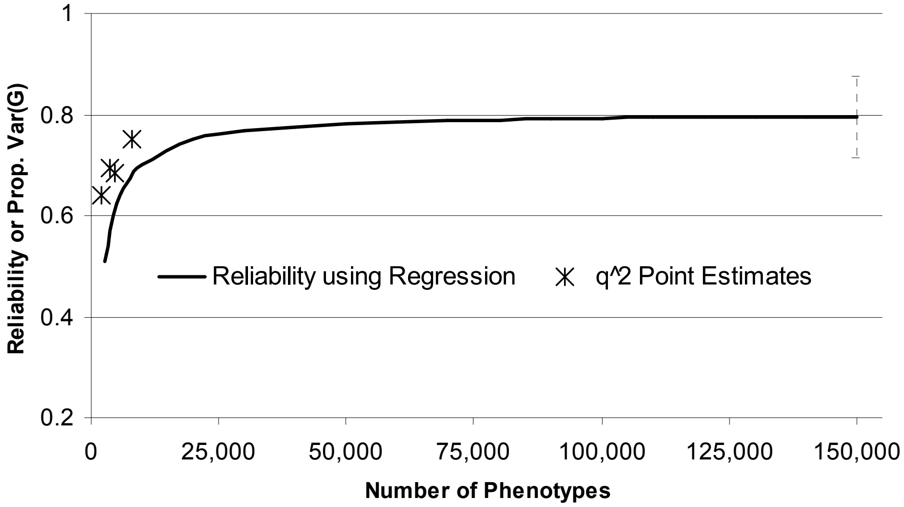


Figure 1. Reliability estimates from regression equation (8) and point estimates from equation (5) of the proportion of genetic variation tagged in North American Holstein cattle in trait Net Merit using the 50K Illumina SNP chip. Error bars represent a 95% confidence interval at 2 df (± 0.08).

Currently, the maximum proportion of the genetic variation tagged, q^2_{max}, by the 50K Illumina SNP chip is approximately 0.80. There are several ways q^2_{max} could be increased. The first step would be not to exclude SNP with low MAF. While excluding these SNP seems a convenient and effective way to eliminate genotyping errors it also results in a cost to q^2_{max} because of missed rare QTL. Similarly, more SNP with low MAF should be included in future SNP chips to capture more rare QTL alleles. Simulations for Chapter 4 suggest that up to 40% of loci can have a minor allele frequency of <0.05 at mutation drift equilibrium resulting in a u-shape distribution of allele frequencies. In human, several studies are suggesting that a very large number of loci may be contributing to genetic variation in quantitative traits (e.g. Weedon *et al.* 2008; The International Schizophrenia

Consortium 2009). Individually, low frequency alleles are not likely to explain very much of the genetic variance but, taken collectively, the proportion of the genetic variance accounted for by rare QTL could be significant. In addition, as described above, once rare QTL are selected on and their allele frequencies increase, their contribution to the genetic variance also increases. Therefore, in the long term, these QTL are an important source of genetic gain.

The second way to increase q^2_{max} would be to develop SNP chips with higher marker density. This would lead to higher LD between SNP and QTL and, therefore, more of the genetic variation would be captured. Efforts are currently underway in a number of species to increase the SNP density by discovering new SNP to include on chips.

It is currently uncertain whether increasing SNP density or including a greater proportion of rare SNP would lead to greater increases in q^2_{max}. Increases in SNP density are expected to produce significant benefits, whereas the proportion of the genetic variance accounted for by rare alleles is still of considerable debate. Consider as an example in human, where some argue for the rare allele hypothesis (Pritchard 2001) and others favour the common variant common disease theory (CVCD) (Reich & Lander 2001). However, if one considers that current sample sizes in case control studies are quite large (e.g. Wellcome Trust Case Control Consortium 2007), one would expect that the genetic variance of the sum of associations found would be much larger than observed if the CVCD hypothesis were true. Therefore, because collectively the associations detected to date do not seem to account for a larger proportion of the genetic variance, it is likely that rare alleles are an important component of the genetic variance. Given that the impact of rare QTL is potentially large, the best strategy for SNP development would be a combination of increases in SNP density and inclusion of a greater proportion of rare SNP.

The impact of a higher q^2_{max} can be quantified heuristically with the regression equation above. Assuming the slope is constant and assuming the q^2_{max} achieved by a new chip is 0.9, then r^2_o at 10,000 observations would increase to 0.774, which is an improvement of 0.072 from current levels. Therefore, large increases in reliability through advances in SNP chip technology are currently possible irrespective of increases is sample size.

I have shown that knowledge on q^2_{max} is a valuable measure to determine SNP chip performance. I will now propose how q^2_{max} could be used to gain insight into complex traits.

Me **and empirical** Me**.** This thesis has demonstrated that the number of independent chromosome segments, Me, also called the number of effective segments, is a crucial factor affecting the accuracy of GBLUP and BayesB (Chapters 4 and 5). A theoretical derivation of Me, which is an extension on the variance of identity-by-descent sharing for full-sibs (Visscher *et al.* 2006), has been proposed in the literature and that is $Me = 2NeL/ln(4NeL)$ (Goddard 2008). This formula has been extended to take into account chromosomes, resulting in $Me = \sum_{ch} 2NeL/ln(4NeL)$, where ch is chromosome and L is now the chromosome length (Meuwissen 2009). Values are now additive over chromosomes and higher than values from the original. Another predictive equation for Me is $2NeL$ (Hayes *et al.* 2009b). This prediction equation has been used in Chapter 5, though it perhaps lacks the theoretical foundation of the first two. Indeed, one of its problems in real data might be that it does not consider q^2.

Chapter 4 demonstrated that insight can be gained by combining accuracies achieved in real populations and deterministic equations. Specifically, estimates of Me ($\widehat{Me}$) can be obtained using GBLUP accuracy and $\widehat{Me} = (N_P h^2)(1 - r^2_{g\hat{g}})/r^2_{g\hat{g}}$, which is appropriate when the all the genetic variance is accounted for in the analysis. One complicating factor when estimating $\widehat{Me}$ from real data was that the original equations did not account for q^2_{max}. Equations can now be extended to consider q^2_{max}.

The origin of the missing genetic variance in current analyses can be viewed as coming from two sources. Firstly, a proportion of the genome could be missed completely because it is not marked by SNP. In the spirit of Chapter 3, the proportion not tagged by a SNP chip could appear to be part of the error variance because it cannot be tracked and, thus, h^2 would be reduced by the missing variance and approximated by $q^2_{max}h^2$. Furthermore, the genome that is being marked would appear like $q^2_{max}Me$, because some segments would be completely missed in the analysis. Amending Equation (6) by multiplying h^2 and Me by q^2_{max}, and approximating $\sigma^2_e = 1$ as in Chapter 3, results in $r^2_o = q^2_{max}(N_P q^2_{max} h^2)/(N_P q^2_{max} h^2 + q^2_{max} Me)$. Rearranging for Me gives,

$$\widehat{Me} = N_P h^2 (q^2_{max} - r^2_o)/r^2_o. \qquad (9)$$

The second source of the missing genetic variance comes from imperfect LD between markers and QTL. Here one could argue that multiplying Me by q^2_{max} is not necessary because all segments are tracked by markers, though some only weakly. Treating Equation (6) as before, but not multiplying Me by q^2_{max}, reveals

$$\widehat{Me} = q^2_{max} N_P h^2 (q^2_{max} - r^2_o)/r^2_o. \qquad (10)$$

Comparison of the two forms shows that Equation (10) can be quickly calculated by multiplying Equation (9) by q^2_{max}. These equations are expected to represent both ends of a spectrum and it is likely that $\widehat{Me}$ is a function of both sources of the missing genetic variance. Note that, while the formulae for GBLUP accuracy and Me are demonstrated here, the same equations would apply when estimating the number of QTL from Bayesian variable selection (e.g. BayesB) accuracy when the number of QTL is less than Me (Chapter 4).

$\widehat{Me}$ was calculated from the USDA Holstein and Jersey and Australian Jersey reliabilities described above and the results are shown in Table 2. Accounting for q^2_{max} decreases $\widehat{Me}$, when compared to values resulting from the formula in Chapter 4, as expected. Values from the updated formulae are greater than the prediction of Goddard (2008) but there is a general downward trend with increasing accuracy. Accuracy in US Holsteins has not reached asymptotic levels yet with current samples sizes, hence it is expected that $\widehat{Me}$ would continue to trend downward with increases in accuracy, possibly reaching the prediction of Goddard (2008). However, determination of Ne depends on a number of important assumptions and, therefore, it may deviate from actual Ne. Solving for an empirical estimate of Ne, using the formula of Goddard (2008) and $\widehat{Me}$ from Holstein reliability estimate four and Equation (10), results in an estimate of Ne of approximately 147. Thus, the discrepancy between our estimate of Ne and the Ne of 100 (Young &

Seykora 1996) used in our equations is small. Currently, however, $\widehat{Me}$ from the North American reliabilities seem to match predictions per chromosome best (Meuwissen, 2009). The estimates demonstrate clearly that $\widehat{Me}$ is smaller than the Me used in Chapter 5 (Hayes et al., 2009b).

Table 2. Empirical $\widehat{Me}$ and predictions of Me for Holstein and Jersey cattle, heritability = 0.9, assumed effective population size 100 for Holstein and 30 for Jersey, genome length = 30 Morgans, and N_P = number of observations.

	N_P	$\widehat{Me}$ Ch. 4	$\widehat{Me}$ Eq. (9)	$\widehat{Me}$ Eq. (10)	Me Goddard	Me Meuwissen	Me Hayes et al
Holstein	2130	2077	1278	1022	640	1000	6000
Holstein	3576	2331	1220	977	640	1000	6000
Holstein	4498	2813	1441	1153	640	1000	6000
Holstein	7818	3161	1122	897	640	1000	6000
Jersey	280	1323	1008	806	220	376	1800
Jersey	1560	1245	715	572	220	376	1800

Quantification of q^2_{max} is a valuable measure of a SNP chip. Furthermore, incorporating q^2 into deterministic formulae yields more precise insight into $\widehat{Me}$ and the number of QTL affecting quantitative traits.

Performance of GWE Methods Under Different Genetic Architectures

Recently, there has been a focus in the animal breeding literature on the development and evaluation of GWE methods to predict breeding values. Methods are unquestionably crucial to achieving maximum accuracy of GWE. The performance of any method significantly depends on the size and genetic structure of the dataset evaluated, whether it is simulated or real data. In GWE a pattern emerged in the literature where GBLUP and BayesB were repeatedly compared in very similar simulated genetic architectures (e.g. similar Ne and number of QTL (N_{QTL}); Meuwissen *et al.* 2001; Habier *et al.* 2007; Lund *et al.* 2009). Results consistently suggested that BayesB performed better and a conventional wisdom developed that it was superior to GBLUP. This thesis reveals that superiority is not a property of a method but that it depends significantly on both genetic population and trait architecture. Chapter 4 concentrated on GBLUP and BayesB. Here I want to generalise its findings to other GWE methods. Note that discussion focuses, for the most part, on population and trait genetic architecture and it assumes a constant number of phenotypes, genotypes and heritability. Also note that N_{QTL} refers to actual QTL with effect, whereas markers refers to the total number of genetic markers used in the analysis.

Methods of GWE can be categorised into two groups: the first group of methods considers all markers as having an effect and the second group of methods attempts to discriminate on whether markers have an effect or not. Methods in this second group are called variable selection methods. The high dimensionality of the data used in GWE may justify choosing a subset of markers for which to estimate effects. However, considering the impact of genetic architecture shows that the benefit of dimensionality reduction can be variable (Chapter 4).

Although the unifying characteristic of the first group of methods is that each marker is assumed to have an effect, there are differences in how they model the data and this leads to variation in performance. I will focus on the following three methods within the first group: least-squares without model selection, GBLUP (NejatiJavaremi *et al.* 1997; Meuwissen *et al.* 2001), and BayesA which can be regarded as a Bayesian implementation of GBLUP (Meuwissen *et al.* 2001). The least-squares without model selection method estimates marker effects by regression of phenotypes on each locus, one at a time. The implicit assumption of the method is that each marker potentially explains all of the genetic variance. With this approach, the accuracy will decrease as more markers are added. Thus, when many markers are used this method will result in low accuracy and results from Chapter 3 show that the accuracy does not depend on N_{QTL} but on the total number of independent markers.

Higher accuracies are expected with GBLUP for which two implementations have been described. The first replaces a relationship matrix based on pedigree with a realised relationship matrix calculated from identical by state or descent probabilities at all markers across the genome (Fernando & Grossman 1989; NejatiJavaremi *et al.* 1997; Villanueva *et* *al.* 2005). The second implementation of GBLUP is a form of ridge regression, which fits an incidence matrix connecting markers and phenotypes (Meuwissen *et al.* 2001; Habier *et al.* 2007). Marker effects are simultaneously estimated in one step and each marker is shrunk according to a constant variance parameter, usually the genetic variance divided by the number of markers. The two implementations of GBLUP have been shown to be equivalent (Habier *et al.* 2007; Goddard 2008). This thesis investigates GBLUP and clearly confirms that it crucially depends on what is termed the number of independent chromosome segments, Me (Chapter 4). The accuracy of GBLUP is indifferent to N_{QTL} and, in that regard, GBLUP is similar to least squares. However, in contrast to least squares, GBLUP accuracy does not decrease as the number of markers increases. This independence with regard to N_{QTL} and the number of markers is advantageous in traits where N_{QTL} is larger than Me because the dimensionality does not increase beyond Me. Furthermore, when the number of SNP used in an analysis is also large, calculating a realised relationship matrix may be a computationally more efficient way to implement GWE than using a Bayesian approach.

Bayesian estimation without variable selection has been termed BayesA by Meuwissen et al. (2001). Their model assumes that all markers have an effect just like in GBLUP. However, BayesA samples individual marker variances through Gibbs sampling, therefore, they may vary across markers. It is expected that, with appropriate priors, BayesA will perform better than GBLUP if large disparities between marker variances exist. A comparison of BayesA to GBLUP and BayesB would be expected to reveal that BayesA would perform better than GBLUP and worse than BayesB when N_{QTL} is low, which is supported by the results of Meuwissen et al. (2001). As N_{QTL} increases BayesA accuracy is expected to gradually decrease and eventually become similar to GBLUP accuracy once $N_{QTL} \approx Me$. The accuracy of BayesA could be slightly higher than that of GBLUP at high N_{QTL}, because BayesA is able to account for differences in marker variances. In addition, from results described in Chapter 4, BayesA accuracy would also be slightly above BayesB accuracy when $N_{QTL} > Me$, because BayesA performance is not reduced by the error associated with variable selection, which is applied in BayesB.

The second group of GWE methods can be broadly categorised as variable selection methods, because they aim to identify a subset of the genotypic data which explains a large proportion of the variance in phenotypes. There are differences between the methods in how this is accomplished. The following methods are discussed below: Bayesian variable selection (BayesB) (Meuwissen *et al.* 2001), least squares with model selection (Meuwissen *et al.* 2001; Wray *et al.* 2007; Habier *et al.* 2007), partial least squares (Raadsma *et al.* 2008; Solberg *et al.* 2009), principal component regression (Solberg *et al.* 2009), and Lasso (Tibshirani 1996; Hasti *et al.* 2001).

BayesB can be thought of as BayesA with an additional variable selection step. First, BayesB determines the subset of markers with effect with Gibbs sampling. It does so by determining the proportion of markers without effect (π) and then uses the same process as BayesA to sample effects and variances for proportion $1 - \pi$. In BayesB, variable selection is on the magnitude of the effect alone without significance thresholds. Chapter 4 demonstrates that BayesB achieves high accuracy when N_{QTL} is low and gradually diminishes as N_{QTL} increases. Thus, at low N_{QTL} it has an advantage over GBLUP, but eventually, as N_{QTL} increases, its accuracy becomes equal to GBLUP when $N_{QTL} \approx Me$. Once $N_{QTL} > Me$, BayesB accuracy is slightly lower than GBLUP due to error with choosing π.

Least squares with model selection chooses a subset of variables either by forward or backward stepwise selection (Hasti *et al.* 2001) and then estimates the marker effects in this subset. In this implementation all markers in the subset are estimated simultaneously. Forward selection, as applied by Meuwissen (2001) and Habier et al. (2007), builds the final model by choosing significant markers in an iterative process of individual marker linear regressions. Variables are chosen according to a significance threshold. Hence selection takes into account both the magnitude of effect and the standard error of the estimate. Applying a threshold in this way is similar to QTL analysis and therefore brings with it problems associated with the overestimation of effects (Beavis 1998). Habier et al. (2007) reported improved accuracy of least squares model selection at low N_{QTL} through relaxation of the significance thresholds used in Meuwissen et al. (2001). Least squares had a lower accuracy than BayesB and similar or lower accuracy than GBLUP depending on thresholds. It is expected that with increasing N_{QTL} least squares will become progressively worse due to a greater proportion of the genetic variance being missed, because many markers of small effect would not meet thresholds.

Partial least squares and principal component regression reduce dimensionality by choosing a set of linear combinations of the input data (i.e. marker matrix X and phenotypic data y in GWE). Thus, both methods attempt to move from estimating effects of genes to estimating contrasts between groups of genes. Solberg et al. (2009) found that both approaches resulted in very similar accuracies at varying marker densities and low N_{QTL}. In addition, their accuracy was always lower than BayesB accuracy and they are in the same range as GBLUP accuracy for similar parameters when compared to Meuwissen et al. (2001). Thus, when a variable selection method is desirable at low N_{QTL}, BayesB is expected to result in higher accuracy than partial least squares or principle component regression. Furthermore, partial least squares and principal component regression are not predicted to improve accuracy over GBLUP values as N_{QTL} increases, because the number of linear combinations necessary to capture the phenotypic variance would also increase, whereas in GBLUP the variables considered (i.e. Me) stay constant.

Lasso is a shrinkage method and is closely related to ridge regression GBLUP. However, the parameter used in Lasso to shrink back estimates differs from GBLUP. In Lasso, a portion of the marker estimates can effectively be set to zero by choosing a shrinkage parameter which is small. This can be regarded as implicit subset selection. Analogous to BayesA and GBLUP, there are also Bayesian implementations of Lasso (Yi & Xu 2008). The original Lasso, in combination with Least Angle Regression (Efron *et al.* 2004), resulted in higher accuracy than either GBLUP or BayesA when tested in the 12th QTL-MAS Workshop Uppsala data (G. Usai and B. Hayes, pers. com., 2009). In this dataset N_{QTL} was low (~ 50) and therefore it would be expected that a subset method and BayesA produce better results than GBLUP. In traits with high N_{QTL} this advantage is predicted to erode.

In summary, there are many methods of GWE and they fall in two main categories: those which choose subset of loci and those which do not. Other GWE methods exist which have not been discussed in this section, such as non-parametric analysis (Gonzalez-Recio *et al.*

2008), kernel methods (de los Campos *et al.* 2009), and machine learning approaches (Long *et al.* 2007). The general behaviour of any variable and non-variable selection methods under different population and trait genetic architectures can, however, be extrapolated from the results of this thesis. Variable selection is advantageous when N_{QTL} is low, particularly when $N_{QTL} < Me$. However, the gain in accuracy is expected to diminish as N_{QTL} increases. Indications from real data point towards many N_{QTL} contributing towards the genetic variance in most quantitative traits and only small differences between GWE methods are observed in, for example, most dairy cattle traits (Hayes *et al.* 2009a; VanRaden *et al.* 2009b). Results from this thesis suggest the following decision rule: when $N_{QTL} < Me$, choose a variable selection method (e.g. BayesB) and when $N_{QTL} > Me$ choose a non-variable selection method (e.g. GBLUP, Chapter 4). The Me used could be as in Goddard (2008) or could be empirically estimated from GBLUP as shown in this General Discussion. The evaluation of a particular GWE method must be performed in more than one trait genetic architecture (e.g. low and high N_{QTL}). However, findings may be extrapolated between population genomic structures (i.e. Ne) by considering N_{QTL} as a proportion of Me.

The Impact of Using Sequence Data on the Accuracy of GWE

Increasing numbers of markers are being included on SNP chips and the cost of SNP genotyping is still decreasing due to new and improved genotyping technology. At the same time, the speed of genome sequencing technology has improved dramatically and this also has reduced cost of sequencing. Whether and how sequencing will be used for prediction of genetic values is uncertain. Here I discuss i) the potential benefits of using sequence data in GWE, ii) how sequence data could be analysed and iii) how sample size could be increased by sequencing animals using a chromosomal phasing approach.

The main advantage of using sequence data is that all of the genetic variation would be captured and q^2_{max} would equal one. In terms of deterministic prediction of accuracy, the extensions discussed in this General Discussion to account for q^2_{max} would not be necessary and the formulae proposed in Chapter 4 would be appropriate, provided that additional sources of genetic variation would be interpreted (e.g. copy number variants). However, it may be necessary to better approximate the error associated in choosing a subset from a very large number of markers in deterministic predictions.

Currently in Holstein cattle, q^2 is approximately 0.75 with the 50K Illumina SNP chip at the current number of observations (Figure 1). Moving to a new chip would allow for additional gains in q^2. Assuming that another 0.15 improvement in q^2 is possible with a denser SNP chip and higher sample size, it still would leave an additional 0.10 to be gained by GWE in sequence. Realistically, large numbers of phenotyped individuals would be needed to achieve high accuracy even if q^2 were above 0.90. Therefore, even though sequencing costs are decreasing, sequencing a large number of training individuals and selection candidates in every generation will likely still be cost prohibitive. Thus, sequencing will only be widely used if its costs decrease to very low levels.

The first challenge of sequence data is the large amount of data that has to be processed. In terms of GWE, there is a significant amount of data editing that could be performed to reduce the computational burden. Firstly, monomorphic loci could be excluded from analysis as they do not contribute to the genetic variance. The human HapMap project has discovered approximately 3.1 million SNP (Frazer *et al.* 2007). Secondly, if there is a group of loci in very high LD (i.e. LD $\approx$ 1), then only one locus in the group could be retained. Sequencing would need to be to sufficient fold coverage so low frequency alleles could be confidently identified. These steps would reduce the dataset, but would still leave a very large number of SNP to consider. The crucial parameter is N_{QTL} and, due to the large number of SNP expected in the analysis, approaches are needed which can identify the SNP with effect. The principles for GWE method performance in different population and trait genetic architectures established in SNP data are expected to remain true for sequence data. The extreme dimensionality of the data would require variable selection type methods when N_{QTL} is less than Me. Currently, among variable selection methods, Bayesian methodology (e.g. BayesB, Meuwissen et al. 2001) seems to achieve the highest accuracies. However, MCMC methods come with a heavy computational burden when the number of markers is large and using such approaches for sequence data would be infeasible without considerable increases in computation infrastructure. New methods of variable selection are being developed to reduce computational demands (Meuwissen *et al.* 2009; Shepherd & Woolliams 2009). However, when N_{QTL} is greater than Me, non-variable selection methods are expected to perform better than variable selection approaches. A large computational load is expected and rapid methods for non-variable selection methods such as BayesA will need to be developed.

Sequencing is expected to remain too costly in the foreseeable future to be applied to the large number of individuals required for effective GWE. One way to obtain the large sample sizes needed for high GWE accuracy would be to sequence individuals *in silico*. Chapter 7 describes a method for imputing missing genotypes using a chromosomal phasing algorithm when denser genotypes are available on ancestral and collateral relatives. The approach bore good results when two ancestral generations were available in addition to the animals to be imputed (probands), but its performance was reduced when only one ancestral generation was available because phasing in founders is difficult. There is no upper limit to the density of the ancestral genotypes which can be imputed in the proband. Therefore, it would be possible to impute full genome sequence in individuals which are genotyped with a dense SNP chip. Full sequencing in ancestors would increase the proportion of correctly imputed loci because haplotypes would be fully known in sequenced founders. The number of sequenced ancestors needs to be large enough so that every chromosomal segment in the proband traces back to a sequenced ancestor. Thus, imputation could offer lower cost solutions to both genotyping and sequencing because the principle is the same.

Challenges for Implementing GWE

There are a number of barriers to widespread implementation of GWE. The feasibility of GWE in a particular species needs to be evaluated by considering both the increase in genetic gain that is achievable with GWE over traditional selection methods and the extra costs involved in GWE. The economic gain associated with genetic progress must outweigh the cost associated with genotyping and phenotyping for GWE to be profitable. Here I will discuss strategies to implement GWE in species where currently the costs of genotyping perhaps outweigh the benefits of increased genetic gain from GWE.

Increases in genetic gain from GWE are expected to come from two main sources, assuming selection intensity is constant. Annual genetic gain can be increased from decreases in generation intervals, because GEBVs can be calculated for juveniles without phenotypic records or progeny. The second source of genetic gain arises from increases in accuracy expected from applying GWE methodology. The potential to decrease generation intervals depends on the species. For example, in pigs and poultry, generation intervals are already short and GWE may not provide substantial decreases. On the other hand, in cattle, generation intervals are much longer, partly because of progeny testing of bulls. Therefore, large increases in genetic gain due to reduction of generation intervals may be possible with GWE in cattle.

The additional genetic gain from increased GWE accuracy is less variable across species and is more dependent on the number of phenotypes, h^2 and Me. Predictive formulae have shown that a large sample sizes are required to get high GWE accuracies (Chapter 3 this thesis; Goddard 2008; Hayes *et al.* 2009b). The main constraint to apply GWE is that usually only a limited number of phenotypes and genotypes are available, particularly for traits of low heritability. However, it is exactly in these traits that GWE could have the greatest benefit.

GWE could be used to select on novel traits which are difficult to phenotype in the general population. Marker estimates could be trained in a resource population and GEBVs could be calculated for selection candidates without a measured phenotype. The size of the resource population needed to train highly accurate marker estimates would also depend greatly on the h^2 of the novel trait, and in traits of low h^2 this would be a significant challenge.

Increasing samples size requires consideration of both phenotyping and genotyping. Collection of phenotypic records is being practiced in most species currently, though mostly in easily recorded traits. In species such as dairy cattle, a large number of production records have traditionally been collected to estimate accurate bull EBVs. Although the infrastructure may already exist to collect large numbers of phenotypes, in some species, it may be difficult to collect a large number of accurate phenotypic records to achieve high GWE accuracy. This is especially true for traits involving subjective scoring, such as dressage in horses, or classification of disease phenotypes, such as psychological disorders in human. Even though currently many species have systems in place to collect phenotypes, there is a need to continue and, in some cases, increase efforts to obtain large numbers of highly accurate phenotypes to reach high GWE accuracy.

Increasing the number of genotypes is greatly dependent on the cost of genotyping. While genotyping costs are decreasing, it still may be too costly to justify widespread genotyping in species where gains in genetic progress are expected to come, in large part, from increases in accuracy alone (e.g. pigs and poultry). However, lower cost solutions to benefit from the expected extra gains from GWE may exist and should be developed.

One such solution is imputation of missing genotypes through chromosomal long range phasing as described in Chapter 7. The principles for imputing missing genotypes and sequence data, as discussed above, with this approach are the same. The method makes use of ancestral and collateral relatives which are genotyped at full marker density to impute genotypes in probands which are genotyped with at lower marker density. Genotyping costs would be reduced because selection candidates would only need to be genotyped with low density chips that are less costly. Therefore imputation could provide a lower cost solution to increase the number of genotypes. Furthermore, SNP chip technology is still improving at a fast rate, and to benefit from the increases in SNP density in new chips, re-genotyping animals, which have already been genotyped at lower density, with new chips may be necessary. Imputation of new SNP in animals genotyped with the older low density chip is possible if the low density SNP are also included in the high density chip. This could provide a solution to increase sample size and to upgrade to higher density SNP chips. Similarly, imputation could be used to combine animals which are genotyped with two SNP chips with similar density, which have a large proportion of SNP in common, to create a denser combined SNP chip.

The number of ancestors with full density genotyping needed to impute a large proportion of loci with high accuracy needs further investigation. Breeding systems which have a relatively small number of founders would likely be most suitable for imputation (e.g. dairy cattle). Less hierarchical populations, such as beef cattle and sheep, would require relatively more ancestors with high density genotypes, making implementation more costly.

Strategic genotyping could be used to take full advantage of genotype imputation to increase GWE accuracy and reduce cost. Strategic genotyping designs would likely differ across species and even across breeds. However, all strategies would try to identify the smallest number of individuals with full density genotypes needed for successful imputation in animals genotyped at lower density. Pedigree analysis could reveal the key ancestors which are expected to have contributed the largest proportion of genes to the current generation. If DNA of such key ancestors is not available, then a sufficient number of their progeny with high density genotypes could be added to the sample to represent their genomes. Any set of animals with full genotyping would be added to the sample to increase the probability that shared segments could be identified. Once enough relatives were fully genotyped for feasible imputation, each subsequent generation could be imputed as long as no non-genotyped individuals from outside were introduced into the population. Nucleus schemes could be established both for intensive selection for genetic gain but also to produce breeding animals for the general population. While the best implementation strategy for imputing genotypes is currently uncertain, significant potential exists to increase sample size and reducing genotyping cost in a way that the theoretical benefits of applying GWE can be materialised in animal breeding programs.

Looking to the future. The first wave of using genomic data by incorporation of few QTL into breeding programs may have largely been a false dawn. We are now in the midst of the second wave of applying genomics in breeding programs and the two main drivers of this wave are genome-wide evaluation and the arrival of new technology. Despite all of the unknowns associated with genome-wide evaluation, there are reasons to be optimistic about its widespread use in future years. The methodology overcomes some of the problems associated with earlier genomics approaches and is easier to apply. Genome-wide evaluation can increase genetic gain in many species and there is great potential to reorganise breeding programs to exploit its benefits. Furthermore, solutions to adapt genome-wide evaluation to species where the benefits depend greatly on implementation costs are starting to emerge. It is for these reasons that we are witness to a new dawn of using genomics in the genetic evaluation of populations.

ACKNOWLEDGEMENTS

Piter Bijma, Johan van Arendonk, Beatriz Villanueva and John Woolliams have made helpful comments on this General Discussion. Paul VanRaden, USDA, has provided reliability estimates in real dairy cattle data.

REFERENCES

Beavis, W. D., 1998 QTL analyses: power, precision, and accuracy., pp. 145-162 in *Molecular Dissection of Complex Traits*, edited by A. C. Paterson. CRC Press, New York.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**: 553-561.

de los Campos, G., D. Gianola, and G. J. M. Rosa, 2009 Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. *J. Anim. Sci.* **87**: 1883-1887.

Dekkers, J. C. M., 2007 Prediction of response from marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* **124**: 331-341.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani, 2004 Least angle regression. *Annal Stat.* **32**: 407-451.

Fernando, R. L., and M. Grossman, 1989 Marker assisted selection using best linear unbiased prediction. *Genet. Sel Evol.* **21**: 467-477.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve *et al.* 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-8U3.

Goddard, M. E., 2008 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**: 245-252.

Gonzalez-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. Rosa *et al.* 2008 Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* **178**: 2305-2313.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**: 2389-2397.

Hasti T., R. Tibshirani, and J. Friedman, 2001 *The Elements of Statistical Learning*. Springer Science and Business Media, New York.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009a Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* **92**: 433-443.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009b Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91**: 47-60.

Long, N., D. Gianola, G. J. Rosa, K. A. Weigel, and S. Avendano, 2007 Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J. Anim. Breed. Genet.* **124**: 377-389.

Lund, M. S., G. Sahana, D. J. de Koning, G. Su, and O. Carlborg, 2009 Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. *Bmc Proc* **3 Suppl 1**.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Meuwissen, T. H. E., 2009 Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* **41**.

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams, 2009 A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.* **41**.

NejatiJavaremi, A., C. Smith, and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* **75**: 1738-1745.

Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Gen.* **69**: 124-137.

Raadsma, H. W., G. Moser, R. E. Crump, M. S. Khatkar, K. R. Zenger *et al.* 2008 Predicting Genetic Merit for Mastitis and Fertility in Dairy Cattle using Genome Wide Selection and High Density SNP Screens. *Animal Genomics for Animal Health* **132**: 219-223.

Reich, D. E., and E. S. Lander, 2001 On the allelic spectrum of human disease. *Trends in Genetics* **17**: 502-510.

Shepherd R.K. & Woolliams J.A. Genomic selection using a fast EM algorithm 1. Understanding the methodolgy. Association for the Advancement of Animal Breeding and Genetics 30th Anniversary Conference. Proc.Assoc.Advmt.Anim.Breed. 2009.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen, 2009 Reducing dimensionality for prediction of genome-wide breeding values. *Genet. Sel. Evol.* **41**.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. *J. Anim. Sci.* **86**: 2447-2454.

The International Schizophrenia Consortium, 2009 Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**: 748-752.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc.Ser.B-Method.***58**: 267-288.

VanRaden P.M., daSilva M. & Sullivan P. National and International Genomic Evaluation in Dairy Cattle. USDA, http://aipl.arsusda.gov/publish/presentations/ADSA09/ADSA09_pvr.ppt . 2009a.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.* 2009b Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **92**: 16-24.

Villanueva, B., R. Pong-Wong, J. Fernandez, and M. A. Toro, 2005 Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* **83**: 1747-1752.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu *et al.* 2006 Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *Plos Genetics* **2**: 316-325.

Weedon, M. N., H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans *et al.* 2008 Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics* **40**: 575-583.

Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661-678.

Wray, N. R., M. E. Goddard, and P. M. Visscher, 2007 Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**: 1520-1528.

Yi, N. J., and S. H. Xu, 2008 Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045-1055.

Young, C. W., and A. J. Seykora, 1996 Estimates of inbreeding and relationship among registered Holstein females in the United States. *J. Dairy Sci.* **79**: 502-505.

Summary

This thesis investigates the use of genomic marker data in the genetic evaluation of populations. Incorporating molecular data into breeding value estimation can improve its accuracy or correlation of true and estimated breeding values. The availability of large numbers of single nucleotide polymorphisms (SNP) has caused a shift in the way that genetic marker data is applied. Before dense SNP chips, the focus was primarily on marking particular quantitative trait loci (QTL), which are segments of the genome found to be associated with a phenotypic trait. These QTL would then be incorporated into breeding schemes through marker-assisted selection. The arrival of dense SNP markers datasets was preceded with the proposal of a method called genome-wide evaluation (GWE). In GWE, all chromosome segments effects are estimated simultaneously in one step and these effects are then summed to breeding value for an individual. This new method reduces the number of distinct steps necessary to use genetic marker data in genetic evaluation and also overcomes some of the challenges faced by QTL detection approaches.

The chapters in this thesis present work in both QTL detection and GWE. However, there is significant focus on GWE after Chapter 2. The chapters on GWE have a strong emphasis on the accuracy of GWE. Deterministic predictions are proposed and tested and the impact of genetic architecture on GWE methods is evaluated. A method is presented which can impute missing genotypes to increase the density of genotypes and, in turn, increase GWE accuracy. Furthermore, ways to quantify the missing genetic variance and challenges to implementing GWE are discussed.

Chapter 2 is a QTL detection study in Holstein cattle using both a linkage analysis variance component method and an association approach in a 10K Bovine SNP chip. The first approach exploits linkage disequilibrium within families and found 102 potential QTL, whereas the second method makes use of linkage disequilibrium across the whole population and detected 144 significant SNP associations.

In Chapter 3 deterministic formulae for the prediction of accuracy are derived for continuous and dichotomous traits in population and case control studies using a GWE least

squares approach. The core formula proposed for the accuracy of GWE in a continuous trait is $r_{g\hat{g}} = \sqrt{N_P h^2/(N_P h^2 + N_G)}$, where $r_{g\hat{g}}$ is the correlation between true and estimated breeding values, N_P is the number of phenotypes, h^2 is the trait heritability, and N_G is the number of loci. The predictions are extensively tested using stochastic simulations and found to be appropriately responsive to the factors affecting accuracy. The formulae in this chapter represent the foundation for theoretical work in later chapters.

In Chapter 4, a hypothesis prompted by Chapter 3 is investigated regarding the impact of population genomic structure and trait genetic architecture on GWE methods. Genomic best linear unbiased prediction (GBLUP) and a Bayesian variable selection method (BayesB) are compared at three different effective population sizes and a wide range of number of QTL affecting the trait. GBLUP had a constant accuracy regardless of the number of QTL, confirming that its accuracy depends crucially on the number of independent chromosome segments. BayesB had higher accuracy than GBLUP when the number of QTL was low, but its accuracy decreased as the number of QTL increased. Eventually, BayesB accuracy reached a lower plateau which was just below GBLUP accuracy, suggesting that BayesB also depends on the number of independent chromosome segments at high numbers of QTL. Furthermore, deterministic equations of Chapter 3 are extended to predict the accuracy of these two methods.

Chapter 5 compares deterministic predictions from Chapter 3, and those suggested by other authors, to accuracies achieved in real Holstein and Jersey cattle populations both in the USA and Australia. Deterministic predictions match real accuracies generally well, though there is a need to extend the equations to account for the proportion of the genetic variance captured by a SNP chip.

In Chapter 6, theoretical concepts established from studies on inbreeding from traditional methods and the use of molecular markers, are used to extrapolate what inbreeding rates are expected with selection based on GWE. It concludes that genomic selection will result in lower rates of inbreeding per generation for the same rate of genetic gain when compared to selection based on traditional best linear unbiased prediction.

Chapter 7 describes a method to phase and impute missing genotypes. The approach is tested in datasets containing varying number of generations and with three different proportions of loci missing. Performance is very good when more than two generations of individuals are available. The main application of this long-range phasing approach will be to impute missing SNP in individuals genotyped at lower density with information from relatives genotyped at higher density. This will increase the sample sizes available for GWE, which, in turn, will increase GWE accuracy.

Chapter 8 is the General Discussion which raises four main topics. First, a method is presented to estimate the proportion of the total genetic variation tagged by current SNP chips. This proportion is estimated for the 50K Illumina Bovine chip using US Holstein data. Furthermore, work in Chapter 4 is revisited and empirical estimates of the number of independent chromosome segments, which account for the proportion of the genetic variance tagged by markers, are presented. Secondly, the performance of GWE methods in different genetic population and trait architectures is discussed. The third topic discussed is the impact that sequence data is likely to have on GWE. Finally issues related to implementation are considered.

Samenvatting

Dit proefschrift onderzoekt het gebruik van moleculaire merkers voor genetische evaluatie van populaties. Moleculaire merkers kunnen worden gebruikt om de nauwkeurigheid van geschatte fokwaardes te verhogen. De manier waarop merkers worden gebruikt is veranderd door het beschikbaar komen van grote hoeveelheden zogenaamde SNP-merkers. In het verleden was men gericht op het opsporen van een beperkt aantal zogenaamde QTL, delen van het genoom die direct in verband staan met een kenmerk. Het doel was om deze QTL te benutten in fokprogramma's met behulp van merker-ondersteunde selectie. Met het beschikbaar komen van grote hoeveelheden SNP-merkers kan gebruik worden gemaakt van een methode die gericht is op het gehele genoom, en bekend staat als "genome-wide evaluation" (GWE). Bij GWE worden effecten voor alle chromosoomsegmenten in één keer geschat, en deze worden opgeteld tot een totale fokwaarde voor het individu. Deze methode beperkt het aantal stappen dat nodig is voor toepassing van moleculaire merkers in genetische evaluatie, en lost ook een aantal problemen op die bij QTL-detectie optreden.

Dit proefschrift presenteert resultaten van zowel QTL-detectie als GWE. Vanaf Hoofdstuk 3 richt het proefschrift zich op GWE, met een nadruk op nauwkeurigheid. Deterministische voorspellingen van nauwkeurigheid worden gepresenteerd en getest, en de invloed van de genetische structuur op nauwkeurigheid wordt onderzocht. Een methode wordt gepresenteerd voor het berekenen van missende genotypes, met als doel merkerdichtheid en nauwkeurigheid van GWE te verhogen. Daarnaast worden praktische toepassing van GWE en manieren om ontbrekende genetische variatie te kwantificeren bediscussieerd.

Hoofdstuk 2 gaat over QTL-detectie in Holstein melkvee met een 10K SNP-chip, en vergelijkt een linkage-analyse variantie-componenten methode met een associatie studie. Met de eerste methode, die linkage-disequilibrium binnen families benut, zijn 102 potentiële QTL gevonden. Met de tweede methode, die linkage-disequilibrium in de gehele populatie benut, zijn 144 significante SNPs gevonden.

Hoofdstuk 3 presenteert formules voor de nauwkeurigheid van GWE voor zowel continue als 0/1-kenmerken, in populaties en case-control studies, gebruik makend van de kleinste-kwadraten methode. De nauwkeurigheid van GWE voor een continu kenmerk wordt gegeven door $r_{g\hat{g}} = \sqrt{N_P h^2/(N_P h^2 + N_G)}$, waarin $r_{g\hat{g}}$ de correlatie tussen echte en geschatte fokwaarde voorstelt, N_P het aantal fenotypes, h^2 de erfelijkheidsgraad, en N_G het aantal loci. De voorspellingen zijn uitgebreid getest met behulp van simulatie, en voorspellen het verloop van nauwkeurigheid goed. Deze formules vormen de basis voor theoretisch werk in de volgende hoofdstukken.

Hoofdstuk 4 onderzoek een hypothese over de invloed van populatie en genoomstructuur op de nauwkeurigheid van GWE-methoden, die voortkomt uit Hoofdstuk 3. Genomische "best linear unbiased prediction" (GBLUP) en een Bayesiaanse variabele selectiemethode worden vergeleken voor verschillende waardes van effectieve populatieomvang en aantallen QTL. Resultaten laten zien dat de nauwkeurigheid van GBLUP niet wordt beïnvloed door het aantal QTL. Dit bevestigt dat nauwkeurigheid afhangt van het aantal onafhankelijke chromosoom segmenten, en niet van het aantal QTL. BayesB was nauwkeuriger dan GBLUP bij weinig QTL, en vertoonde een afnemende nauwkeurigheid bij een toenemend aantal QTL. Bij veel QTL bereikte BayesB een iets lagere nauwkeurigheid dan GBLUP. Dit suggereert dat, bij veel QTL, de nauwkeurigheid van BayesB ook wordt bepaald door het aantal onafhankelijke chromosoom segmenten. In Hoofdstuk 4 zijn de formules voor nauwkeurigheid uit Hoofdstuk 3 uitgebreid naar GBLUP en BayesB.

Hoofdstuk 5 vergelijkt voorspellingsformules uit Hoofdstuk 3 en formules voorgesteld door anderen met gerealiseerde nauwkeurigheden in Holstein en Jersey melkvee uit de VS en Australië. Resultaten tonen een goede overeenkomst tussen voorspelde en gerealiseerde nauwkeurigheden, maar ook de noodzaak tot uitbreiding van de formules om rekening te houden met de fractie genetische variatie die door SNPs wordt verklaard.

Hoofdstuk 6 bespreekt de verwachte inteelttoename bij selectie op GWE-fokwaarden, gebruikmakend van theoretische concepten over inteelttoename in geselecteerde populaties. De conclusie is dat, bij eenzelfde genetische vooruitgang, selectie op GWE-fokwaarden zal resulteren in lagere inteelttoenames per generatie dan selectie op klassieke “best linear unbiased prediction” van fokwaarden.

Hoofdstuk 7 beschrijft een methode om missende genotypes te berekenen en allelen toe te wijzen aan haplotypes. De methode is getest voor verschillende aantallen generaties en proporties missende loci, en werkt zeer goed als tenminste drie generaties beschikbaar zijn. De belangrijkste toepassing van deze methode ligt in het berekenen van ontbrekende SNPs in individuen die zijn gegenotypeerd voor een beperkt aantal merkers, gebruik makend van uitgebreide genotypes gemeten aan hun verwanten. Dit verhoogt de steekproefomvang van GWE en daarmee ook de nauwkeurigheid.

Hoofdstuk 8 bevat een algemene discussie gericht op vier onderwerpen. Als eerste wordt er een methode gepresenteerd voor het schatten van de fractie genetische variantie die door huidige SNP-chips wordt verklaard. Deze fractie wordt geschat voor de 50K Illumina Bovine chip, toegepast op Holstein gegevens uit de VS. Daarnaast wordt werk uit Hoofdstuk 4 herzien, en worden empirische schattingen voor het aantal onafhankelijke chromosoom segmenten gegeven, rekening houdend met de fractie variantie verklaard door SNP-chips. Ten tweede wordt de nauwkeurigheid van GWE-methoden bediscussieerd, in relatie tot de structuur van de populatie en het genoom. Ten derde wordt de verwachte impact van kennis van de DNA-volgorde op GWE besproken. Als laatste wordt toepassing van GWE in de praktijk bediscussieerd.

Curriculum Vitae

Hans Dieter Daetwyler was born in Moosleerau, Switzerland on May the 8th 1973 as the sixth child of Hans and Nelly Daetwyler-Neeser. Hans attended primary school in Moosleerau between 1980 and 1985 and high school in Reitnau, Switzerland from 1985 to 1989. He then started a farmer apprenticeship while working on an organic farm in Noflen, Switzerland. In 1991, Hans immigrated to Canada with his family where they bought a dairy farm in Atwood, Ontario. From 1991 to 1992 he attended Listowel District Secondary School, Listowel, Ontario to learn English. In 1994, he graduated from Ridgetown College, Ridgetown, Ontario with an Honours Diploma in Animal Production after which he returned home to co-manage the dairy farm with his parents and his brother for 6 years. In 2000 he started a B.A. in Political Science at the University of Toronto but a year later he switched to the B.Sc. (Agr.) program at the University of Guelph and majored in Animal Science. After graduating in 2004, he started a M.Sc. by research in Animal Breeding and Genetics within the Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph. The M.Sc. thesis consisted of a genome scan to detect quantitative trait loci in dairy cattle using a variance component approach and an association method. He married Michael Nunn in 2005. In October 2006, Hans successfully defended his M.Sc. thesis and later that month he started a Ph.D. in Animal Breeding and Genetics at the Roslin Institute, Roslin, UK as a Marie Curie Fellow while being matriculated at Wageningen University. Hans' Ph.D. research was mostly completed at the Roslin Institute, but he also completed longer term research stays at the Animal Breeding and Genomics Centre, Wageningen University working with Dr. Piter Bijma and at the Department of Primary Industries Victoria, Bundoora, Australia doing research with Dr. Ben Hayes and Professor Mike Goddard.

List of Publications

Peer Reviewed Articles

Daetwyler, H.D., Villanueva, B., and J.A. Woolliams. 2008. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3(10): e3395.

Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M., and J. A. B. Robinson. 2008. A Genome Scan to Detect Quantitative Trait Loci for Economically Important Traits in Holstein Cattle Using Two Methods and a Dense Single Nucleotide Polymorphism Map. J. Dairy Sci. 91:3225-3236.

Daetwyler, H.D., Villanueva, B., Bijma, P., and J.A. Woolliams. 2007. Inbreeding in Genome-Wide Selection. J. Anim. Breed. Genet. 124: 369-376.

Conference Proceedings

Hayes, B.J., Daetwyler, H.D., Bowman, P.J., Moser, G., Tier, B., Crump, R., Khatkar, M., Raadsma, H., and M.E. Goddard. 2009. Accuracy of genomic selection: comparing theory and results. Association for the Advancement of Animal Breeding and Genetics 30th Anniversary Conference, September 28 – October 1, 2009. Proc. Assoc. Advmt. Anim. Breed. Genet. 17: 352 -355.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., and J.A. Woolliams. 2009. A comparison of genomic BLUP and BayesB at various population and trait genetic architectures. QTL-MAS Workshop 13 2009, Wageningen, Netherlands, April 20-21, 2009.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., and J.A. Woolliams. 2009. Genome-wide evaluation: genomic BLUP or BayesB? 60th Annual Meeting of the European Association for Animal Production, Barcelona, Spain. August 24-27, 2009.

Daetwyler, H.D., Villanueva, B., and J.A. Woolliams. 2008. Accuracy of predicting breeding values and genetic risk of disease using a genome-wide approach. Kennedy Conference and Canadian Society of Animal Science, Guelph, Ontario, Canada. August 11-14, 2008. Can. J. Anim. Sci. 89: 124-125.

Daetwyler, H.D., Villanueva, B., and J.A. Woolliams. 2008. Accuracy of predicting breeding values using a genome-wide approach. QTLMAS Workshop 12, Uppsala, Sweden. May 15-16, 2008.

Daetwyler, H.D., Villanueva, B., and J.A. Woolliams. 2008. Accuracy of predicting genetic risk of disease for population and case control designs using genome-wide markers. Easter Bush Research Consortium Launch Conference, Edinburgh, UK. April 7-8, 2008.

Karrow, N.A., Mallard, B.A., Schenkel, F.S., Sharma, B.S., Pant, S.D., and H.D. Daetwyler. 2007. Integrative immunogenomics and health of the dairy cow: SNPs and chips and latte to go. 13th International Conference on Production Diseases in Farm Animals, Leipzig, Germany. July 29- August 4, 2007.

Sargolzaei, M., Schenkel, F.S., and H.D. Daetwyler. 2007. First screening of QTL using a segment mapping approach. Proceedings of the Joint Annual Meeting of the American Dairy Science Association, American Poultry Association, Asociacion Mexicana de Produccion Animal, and American Society of Animal Science. San Antonio, TX, USA. July 8 – 12, 2007. J. Dairy Sci. 90 (Suppl. 1): 669.

Daetwyler, H.D., Schenkel, F.S., Sargolzaei, M., and J.A.B. Robinson. 2007. Genome Scan to Detect Quantitative Trait Loci for Economically Important Traits in Holstein Cattle Using a Dense SNP Map. Proceedings of the British Society of Animal Science Annual Conference. April 2 to 4, 2007. Southport, UK. Summary 6.

Daetwyler, H.D., Schenkel, F.S., and J.A.B. Robinson. 2006. Relationship of Multilocus Homozygosity and Inbreeding in Canadian Holstein Sires. Proceedings of the Joint Colloquium of the Canadian Society of Animal Science, Canadian Society of Agronomy and Canadian Society of Horticultural Science. August 1 to 5, 2006. Halifax, Nova Scotia, Canada. Can. J. Anim. Sci. 86: 578-579.

Kolbehdari, D., Daetwyler, H.D., Robinson, J.A.B., and F. S. Schenkel. 2006. Mapping of quantitative trait loci for economically important traits in Canadian Holstein bulls. Proceedings of the Joint Colloquium of the Canadian Society of Animal Science, Canadian Society of Agronomy and Canadian Society of Horticultural Science. August 1 to 5, 2006. Halifax, Nova Scotia, Canada. Can. J. Anim. Sci. 86: 580-580.

Acknowledgements

Collaboration and the exchange of ideas is the life blood of science. It is no surprise then that the work in this thesis is a result of collaborative research within and across research groups. In my relatively short academic career, I have been fortunate to work at several research centres, and in each I have learned a great deal from colleagues, both senior and junior. Here I would like to express my gratitude to those who have contributed to my professional and personal life in the last few years.

John and Beatriz, both of you are scientists of high integrity and skill and I have learned a lot about theory, application and good scientific practice from you. It has been a pleasure to work with you and I am grateful for all the advice and help received. Your enthusiasm for genetics is infectious and I seem to have ‘caught the bug’. Piter, thank you for all the help navigating through the Wageningen Ph.D. and a special thank you for your frank comments on my writings. Johan, I thank you for your guidance through my Ph.D. and also for comments on this thesis.

I very much enjoyed working at the Roslin Institute. I think the main reason for this is the collegial work environment where help on any particular subject was just a door knock away. I shall also miss the humorous lunch time conversations. I especially would like to thank Ricardo for his help with programming and for the many discussions. There are too many names to mention individually, so, collectively, I would like to thank all the staff and students in the North Wing Extension and the whole Institute for their support. Furthermore, I would like to thank the Genesis-Faraday team for organising the SABRETRAIN Ph.D. studentships. Your efforts were greatly appreciated.

Preceding my time at the Roslin Institute I did research for my M.Sc. at the University of Guelph. I would like to thank Flavio, Mehdi, Andy and everyone in the Centre for Genetic Improvement of Livestock for all their help during that time. During my Ph.D. I spent a couple of months at the Wageningen Animal Breeding and Genomics Centre. I thank everyone for all the stimulating discussions and any help received. Thank you Piter for your help on getting my simulation program started. Andreia, your help in getting used to Wageningen was greatly appreciated. A big thank you to Ada and Monique for their excellent organisational support. In addition, I spent three months at the Department of

Primary Industries Victoria. I would like to thank Ben and Mike for their guidance in scientific matters. Ben, Kath, Jennie, Anna, Theo, and Graziano, thank you for inviting me to dinners and social outings.

My dear parents Hans and Nelly, you have worked hard all your life and while doing so you have created a loving and supportive home and a strong base from which to start my own life. I am forever grateful and I could not have asked for better.

Last, but certainly not least, I would like to thank my husband Mike for all of his love, support, patience and understanding. I am reminded daily of how lucky I am to have met you and I cannot wait to tackle the next chapter of our journey together.

Training and Supervision Plan

The Graduate School
WAGENINGEN INSTITUTE of ANIMAL SCIENCES

Name	Hans D. Daetwyler
Groups	The Roslin Institute and Animal Breeding and Genomics Centre
Daily Supervisors	John Woolliams and Beatriz Villanueva
Supervisors	John Woolliams, Beatriz Villanueva, Piter Bijma, Johan van Arendonk
Period	October 2006 to October 2009

The Basic Package	**Year**	**ECTS**
WIAS Introduction Course	2007	1.5
WIAS course ‘Philosophy of Science and Ethics’	2008	1.5
Subtotal		**3**
Scientific Exposure		
International Conferences		
QTL MAS Workshop, Uppsala, Sweden	2008	0.6
Kennedy Conference, Guelph, Canada	2008	1.1
Sabre/Sabretrain Meeting Conference, Foulum, Denmark	2008	0.6
QTL MAS Workshop, Wageningen, The Netherlands	2009	0.6
European Association of Animal Production, Barcelona, Spain	2009	1.2
Seminars and workshops		
The Roslin Institute and University of Edinburgh Seminar Program	2006/09	1.5
Roslin Student Seminars and Student Day	2007/09	1.0
WIAS Science Day, Wageningen	2008	0.3
Aquagenome Workshop, Stirling, UK	2008	0.3
Presentations		
Bristish Society of Animal Science, Southport, UK (Oral)	2007	1.0
Divisional Seminar Genetics and Genomics, Roslin Institute, UK (Oral)	2007	1.0
Kennedy Conference, Guelph, Canada (Oral)	2008	1.0
Sabre/Sabrtrain Conference, Foulum, Denmark (Oral and Poster)	2008	2.0
QTL MAS Workshop 2008, Uppsala, Sweden (Oral)	2008	1.0
WIAS Science Day, Wageningen, The Netherlands (Oral)	2008	1.0
Aquagenome Applied Training Workshop, Stirling, UK (Oral)	2008	1.0
Victoria Department of Primary Industries, Bundoora, Australia (Oral)	2009	1.0
QTL MAS Workshop 2009, Wageningen, The Netherlands (Oral)	2009	1.0
University of Goettingen, Goettingen, Germany (Oral)	2009	1.0
Annual Meeting of EAAP 2009, Barcelona, Spain (Oral)	2009	1.0
Across Edinburgh Complex Trait Group Seminar (Oral)	2009	1.0
Roslin Students Day, Roslin, UK (Poster)	2007	1.0
Easter Bush Research Consortium Launch Conference, Edinburgh, UK (Poster)	2008	1.0
Subtotal		**22.2**

In-Depth Studies		
Disciplinary and interdisciplinary courses		
Gene Detection and Marker Assisted Selection: Putting the Theory into Practice, University of Guelph, Canada	2005	1.5
Computational Techniques in Animal Breeding (Fortran 90)	2007	1.5
Bayesian Statistics, SABRETRAIN and SAC, Edinburgh, UK	2007	1.5
QTL Mapping, MAS and Genomic Selection, WIAS	2008	1.5
Introduction to R for statistical analysis, WIAS	2008	0.6
PhD students' discussion groups		
Quantitative Genetics Journal Club, UofEdinburgh	2006/09	2.7
Quantitative Genetics Book Club Roslin	2006/09	1.0
MSc level courses		
University of Edinburgh Quantitative Genetics Course Modules	2006/07	1.1
Subtotal		**11.4**
Professional Skills Support Courses		
SABRETRAIN Training business models, oral and written communications	2007	0.6
Biotechnology YES, Commercial Awareness Training for Bioscience Researchers	2007	1.5
SABRETRAIN Training, group dynamics, risk and soft system approaches	2008	0.6
Managing People, University of Edinburgh, Staff Development Course	2009	0.3
Subtotal		**3.0**
Research Skills Training		
Preparing own PhD research proposal	2006	6.0
External training period, Dr. Ben Hayes, Melbourne	2009	2.0
Subtotal		**8.0**
Didactic Skills Training		
Teaching Assistant, University of Guelph, Quantitative Genetics	2005	1.0
Teaching Assistant, University of Guelph, Animal Breeding Methods	2005	2.0
Subtotal		**3.0**
Management Skills Training		
Graduate Student Senator, University of Guelph, Ontario, Canada Serving on Board of Graduate Studies and Admissions and Progress Committee	2005/06	2.0
Subtotal		**2.0**
Education and Training Total		**52.6**

Colophon

Chapter 2 of this thesis was in large part conducted at the University of Guelph, Guelph, Canada with funding from L'Alliance Bovitec Inc. and the Ontario Centre for Agriculture Genomics. The remaining Chapters were completed at The Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, UK and at the Animal Breeding and Genomics Centre, Wageningen University, with financial support from the SABRETRAIN Project, which is funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission.

The printing of this thesis was funded by the Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands

Printed by: Wöhrmann Print Service, Zutphen, The Netherlands