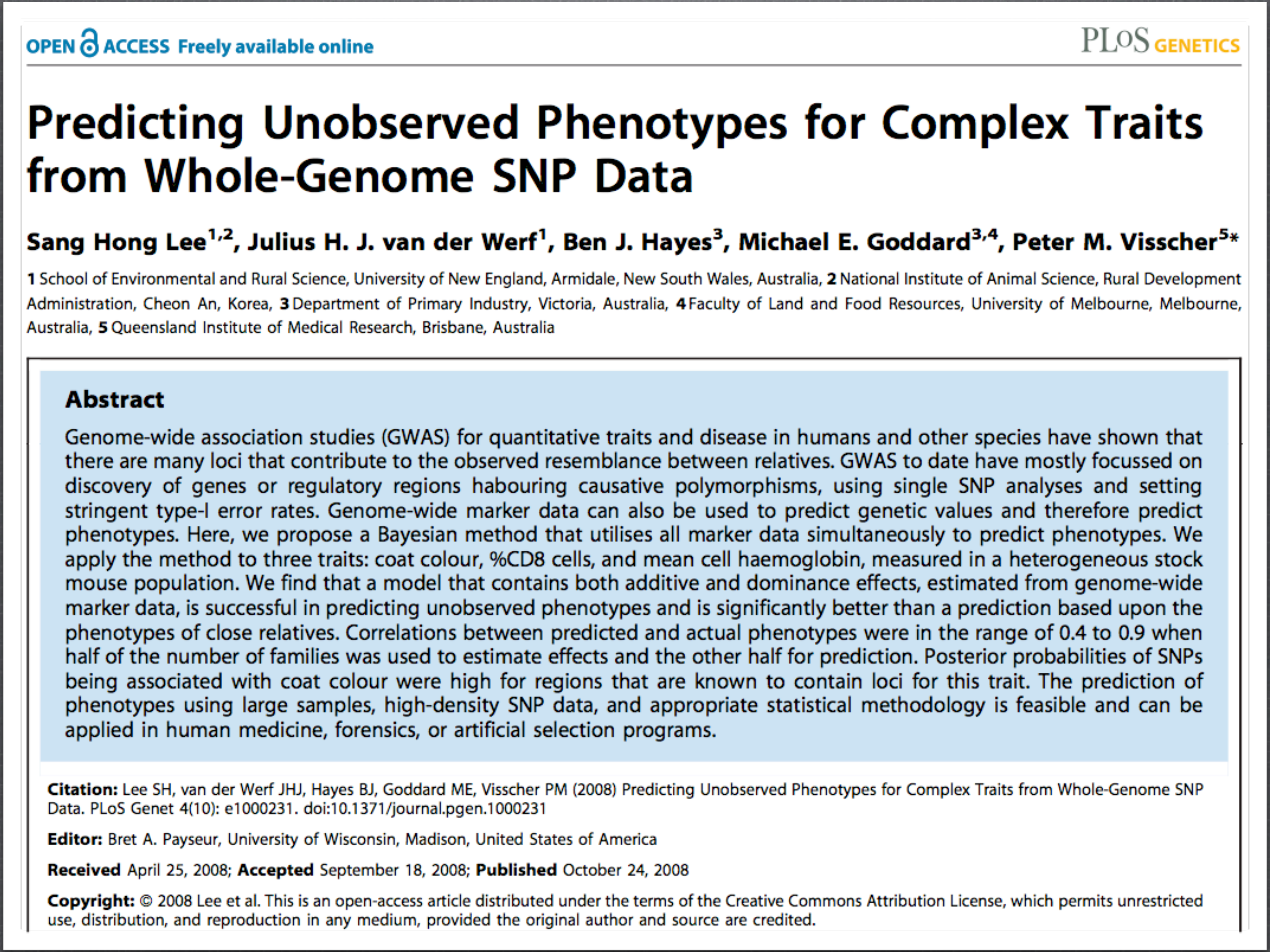


OPEN ACCESS Freely available online

PLoS GENETICS

Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data

Sang Hong Lee[1,2], Julius H. J. van der Werf[1], Ben J. Hayes[3], Michael E. Goddard[3,4], Peter M. Visscher[5]*

1 School of Environmental and Rural Science, University of New England, Armidale, New South Wales, Australia, 2 National Institute of Animal Science, Rural Development Administration, Cheon An, Korea, 3 Department of Primary Industry, Victoria, Australia, 4 Faculty of Land and Food Resources, University of Melbourne, Melbourne, Australia, 5 Queensland Institute of Medical Research, Brisbane, Australia

Abstract

Genome-wide association studies (GWAS) for quantitative traits and disease in humans and other species have shown that there are many loci that contribute to the observed resemblance between relatives. GWAS to date have mostly focussed on discovery of genes or regulatory regions habouring causative polymorphisms, using single SNP analyses and setting stringent type-I error rates. Genome-wide marker data can also be used to predict genetic values and therefore predict phenotypes. Here, we propose a Bayesian method that utilises all marker data simultaneously to predict phenotypes. We apply the method to three traits: coat colour, %CD8 cells, and mean cell haemoglobin, measured in a heterogeneous stock mouse population. We find that a model that contains both additive and dominance effects, estimated from genome-wide marker data, is successful in predicting unobserved phenotypes and is significantly better than a prediction based upon the phenotypes of close relatives. Correlations between predicted and actual phenotypes were in the range of 0.4 to 0.9 when half of the number of families was used to estimate effects and the other half for prediction. Posterior probabilities of SNPs being associated with coat colour were high for regions that are known to contain loci for this trait. The prediction of phenotypes using large samples, high-density SNP data, and appropriate statistical methodology is feasible and can be applied in human medicine, forensics, or artificial selection programs.

Citation: Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. PLoS Genet 4(10): e1000231. doi:10.1371/journal.pgen.1000231

Editor: Bret A. Payseur, University of Wisconsin, Madison, United States of America

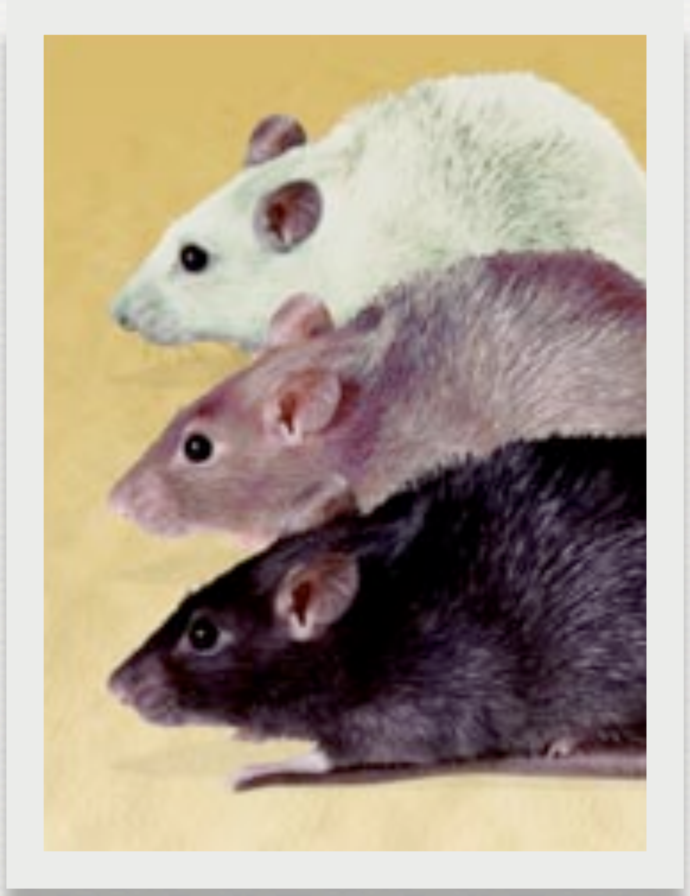
Received April 25, 2008; **Accepted** September 18, 2008; **Published** October 24, 2008

INTRODUCTION

- Genome wide association studies
  - Multi staged, focused on gene discovery, set stringent error rates, one SNP at a time
  - Alternative - prediction of phenotypes using all genome-wide SNP information simultaneously
  - Relationship between genome-wide marker data and phenotypes needs to be modeled

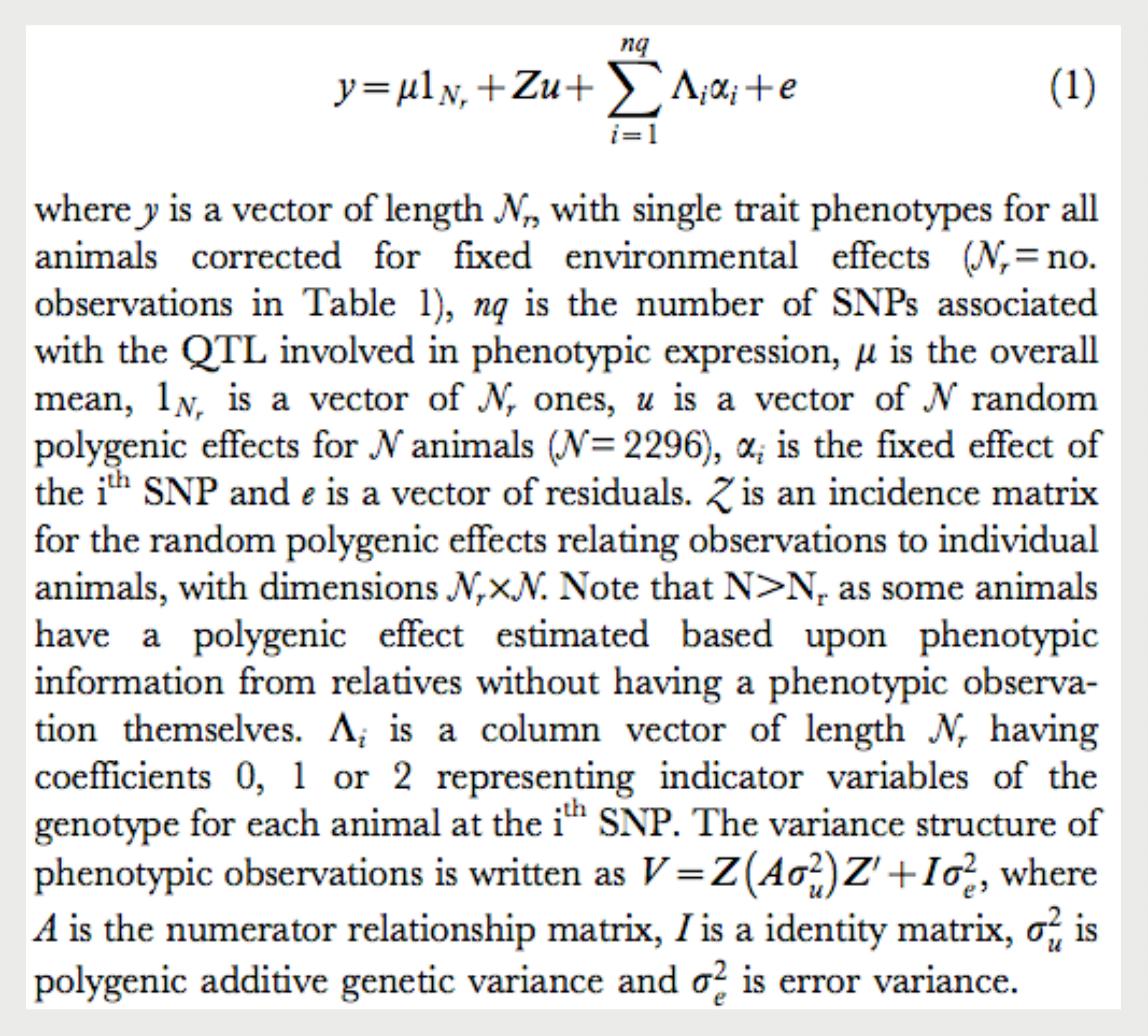
METHODS

- Publicly available data on heterogenous stock mice
  - 2296 animals from 85 families
  - 11730 SNPs
- Linear mixed models
  - Additive genetic model
  - Additive and dominance genetic model



http://gscan.well.ox.ac.uk/

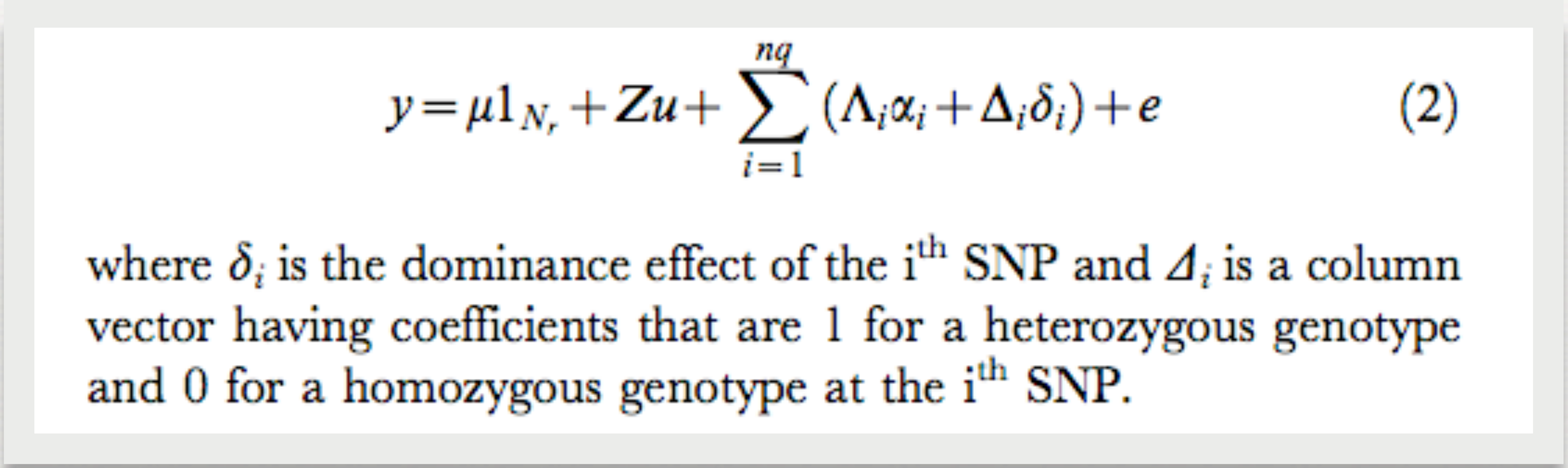
ADDITIVE GENETIC MODEL



$$y=\mu 1_{N_r}+Zu+\sum_{i=1}^{nq}\Lambda_i\alpha_i+e \tag{1}$$

where y is a vector of length N_r, with single trait phenotypes for all animals corrected for fixed environmental effects (N_r = no. observations in Table 1), nq is the number of SNPs associated with the QTL involved in phenotypic expression, μ is the overall mean, 1_{N_r} is a vector of N_r ones, u is a vector of N random polygenic effects for N animals (N = 2296), α_i is the fixed effect of the i^{th} SNP and e is a vector of residuals. Z is an incidence matrix for the random polygenic effects relating observations to individual animals, with dimensions $N_r \times N$. Note that $N>N_r$ as some animals have a polygenic effect estimated based upon phenotypic information from relatives without having a phenotypic observation themselves. Λ_i is a column vector of length N_r having coefficients 0, 1 or 2 representing indicator variables of the genotype for each animal at the i^{th} SNP. The variance structure of phenotypic observations is written as $V=Z\left(A\sigma_u^2\right)Z'+I\sigma_e^2$, where A is the numerator relationship matrix, I is a identity matrix, σ_u^2 is polygenic additive genetic variance and σ_e^2 is error variance.

ADDITIVE AND DOMINANCE GENETIC MODEL



$$y=\mu 1_{N_r}+Zu+\sum_{i=1}^{nq}(\Lambda_i\alpha_i+\Delta_i\delta_i)+e \tag{2}$$

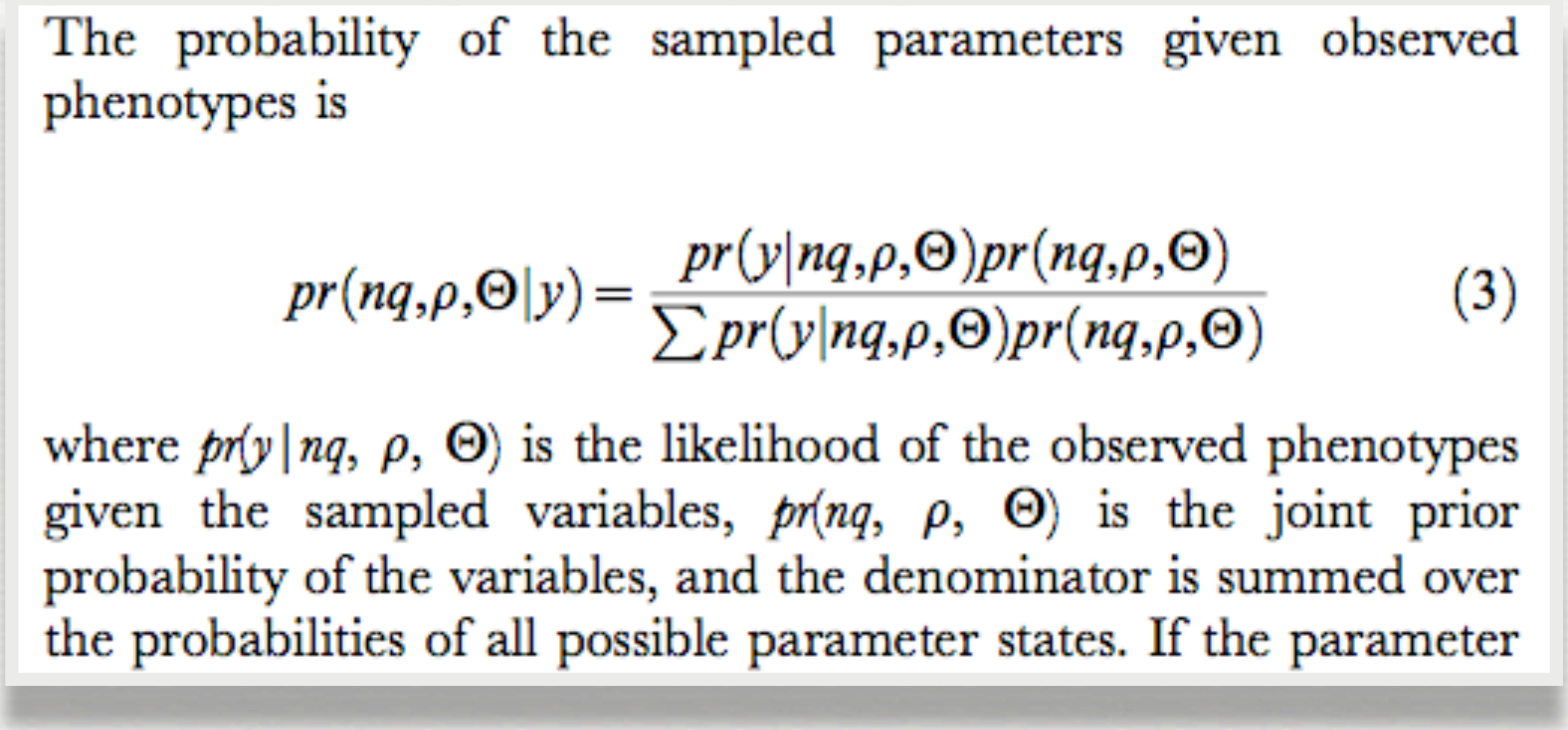
where δ_i is the dominance effect of the ith SNP and $\mathit{\Delta}_i$ is a column vector having coefficients that are 1 for a heterozygous genotype and 0 for a homozygous genotype at the ith SNP.

- Dominance effects due to SNPs are added

ESTIMATION OF EFFECTS AND MODEL SELECTION

- Reversible jump MCMC to simultaneously consider the whole genome

- Unknown phenotypes are predicted based on parameter estimates and averaged over all MCMC rounds.



The probability of the sampled parameters given observed phenotypes is

$$pr(nq,\rho,\Theta|y) = \frac{pr(y|nq,\rho,\Theta)pr(nq,\rho,\Theta)}{\sum pr(y|nq,\rho,\Theta)pr(nq,\rho,\Theta)} \qquad (3)$$

where $pr(y\,|\,nq,\ \rho,\ \Theta)$ is the likelihood of the observed phenotypes given the sampled variables, $pr(nq,\ \rho,\ \Theta)$ is the joint prior probability of the variables, and the denominator is summed over the probabilities of all possible parameter states. If the parameter

ESTIMATION OF EFFECTS AND MODEL SELECTION

- Fixed polygenic heritability used in models
  - 0.72, 0.99, 0.55 for coat colour, CD8%, MCH
  - Best linear unbiased prediction
  - SNP modeling
  - saved computer time (10,000 MCMC iterations)
- Single SNP analyses

PREDICTING UNOBSERVED PHENOTYPES

- Best linear unbiased prediction of polygenic values
  - pedigree and phenotype values only, or additional genomic information (model A), or both (model AD)
  - half of phenotypic data for estimation, remaining for prediction and validation

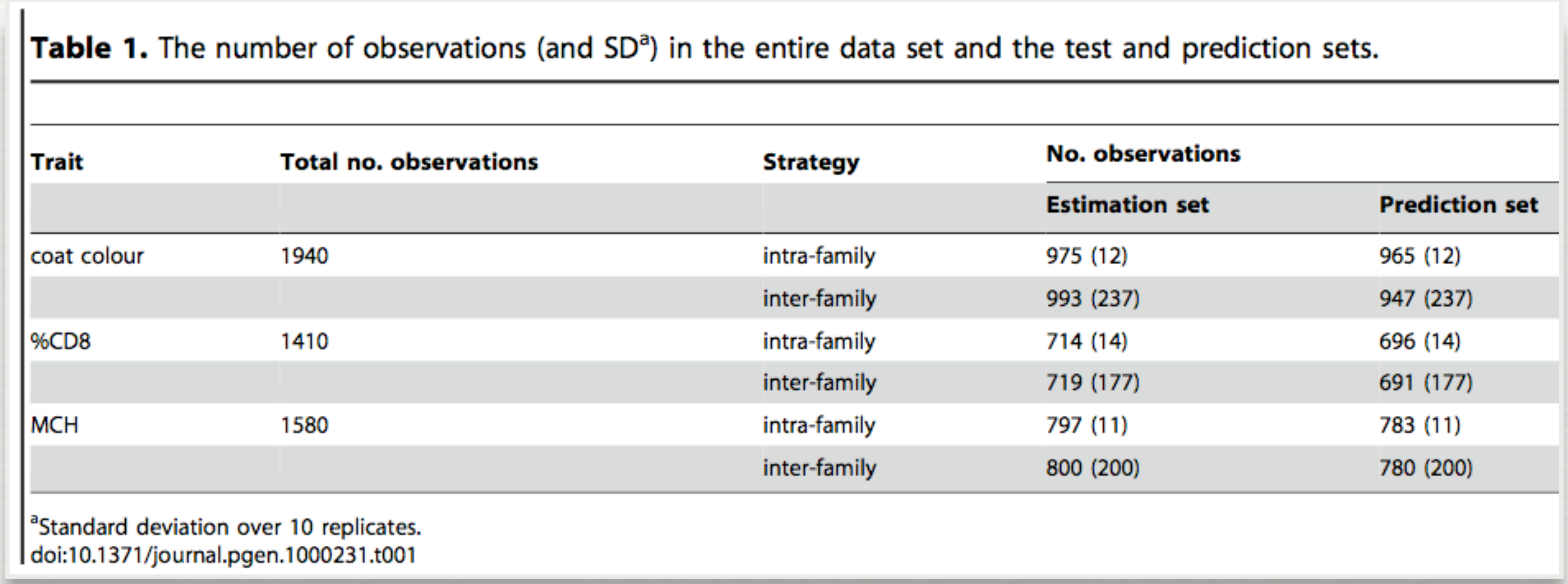


Table 1. The number of observations (and SD[a]) in the entire data set and the test and prediction sets.

Trait	Total no. observations	Strategy	No. observations	
			Estimation set	Prediction set
coat colour	1940	intra-family	975 (12)	965 (12)
		inter-family	993 (237)	947 (237)
%CD8	1410	intra-family	714 (14)	696 (14)
		inter-family	719 (177)	691 (177)
MCH	1580	intra-family	797 (11)	783 (11)
		inter-family	800 (200)	780 (200)

[a] Standard deviation over 10 replicates.

doi:10.1371/journal.pgen.1000231.t001

RESULTS

- Use of genomic information substantially increases the accuracy of predicting unobserved phenotypes

- prediction accuracies are generally better for AD model

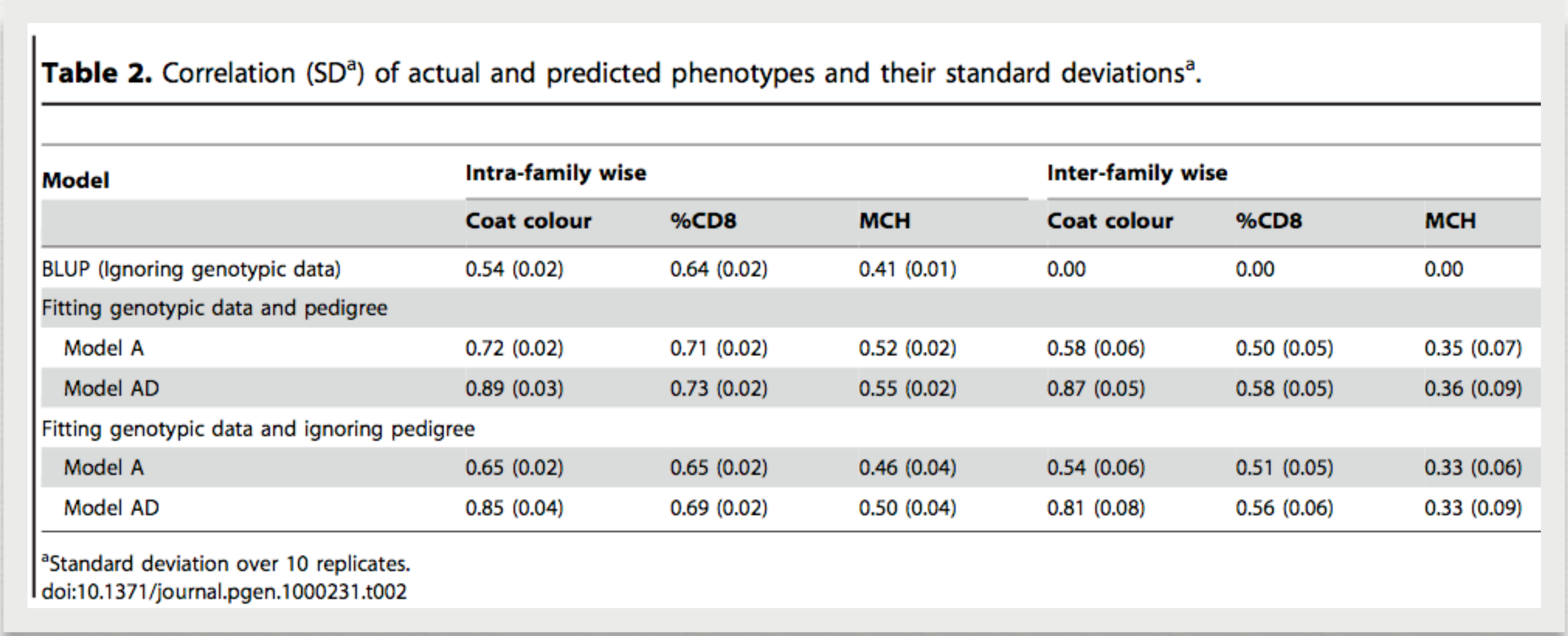


Table 2. Correlation (SD[a]) of actual and predicted phenotypes and their standard deviations[a].

Model	Intra-family wise			Inter-family wise		
	Coat colour	**%CD8**	**MCH**	**Coat colour**	**%CD8**	**MCH**
BLUP (Ignoring genotypic data)	0.54 (0.02)	0.64 (0.02)	0.41 (0.01)	0.00	0.00	0.00
Fitting genotypic data and pedigree						
Model A	0.72 (0.02)	0.71 (0.02)	0.52 (0.02)	0.58 (0.06)	0.50 (0.05)	0.35 (0.07)
Model AD	0.89 (0.03)	0.73 (0.02)	0.55 (0.02)	0.87 (0.05)	0.58 (0.05)	0.36 (0.09)
Fitting genotypic data and ignoring pedigree						
Model A	0.65 (0.02)	0.65 (0.02)	0.46 (0.04)	0.54 (0.06)	0.51 (0.05)	0.33 (0.06)
Model AD	0.85 (0.04)	0.69 (0.02)	0.50 (0.04)	0.81 (0.08)	0.56 (0.06)	0.33 (0.09)

[a] Standard deviation over 10 replicates.

doi:10.1371/journal.pgen.1000231.t002

RESULTS

- Prediction of phenotypes from genetic data
  - not accurate for traits with low heritability
  - large proportion of genetic variation detected

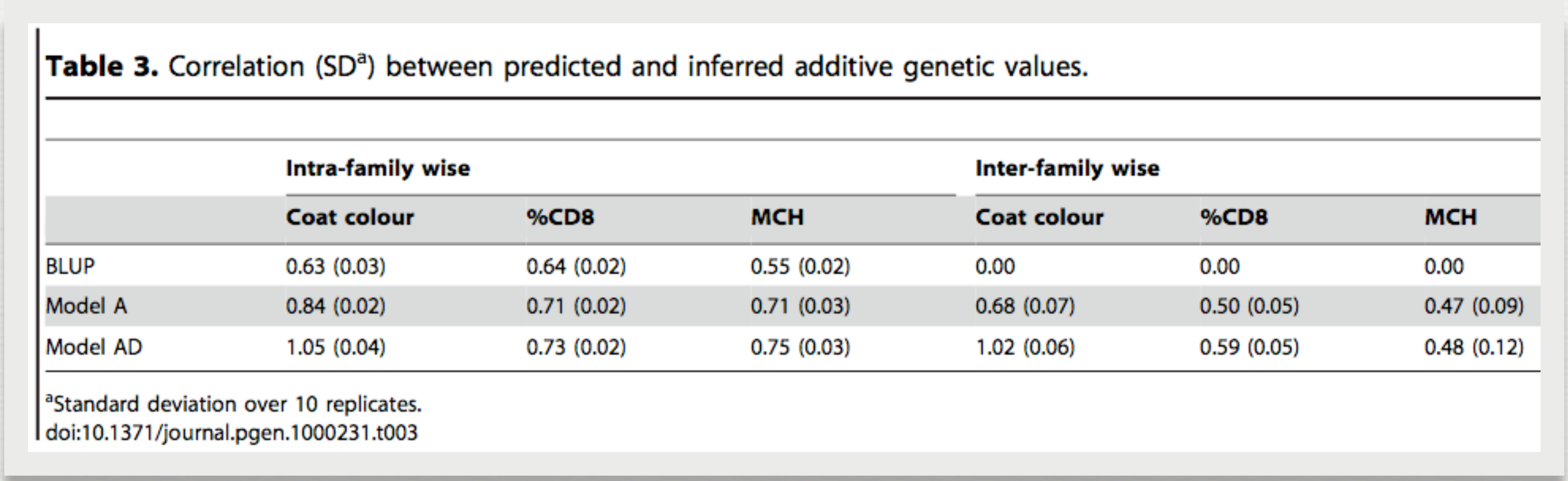


Table 3. Correlation (SD[a]) between predicted and inferred additive genetic values.

	Intra-family wise			Inter-family wise		
	Coat colour	**%CD8**	**MCH**	**Coat colour**	**%CD8**	**MCH**
BLUP	0.63 (0.03)	0.64 (0.02)	0.55 (0.02)	0.00	0.00	0.00
Model A	0.84 (0.02)	0.71 (0.02)	0.71 (0.03)	0.68 (0.07)	0.50 (0.05)	0.47 (0.09)
Model AD	1.05 (0.04)	0.73 (0.02)	0.75 (0.03)	1.02 (0.06)	0.59 (0.05)	0.48 (0.12)

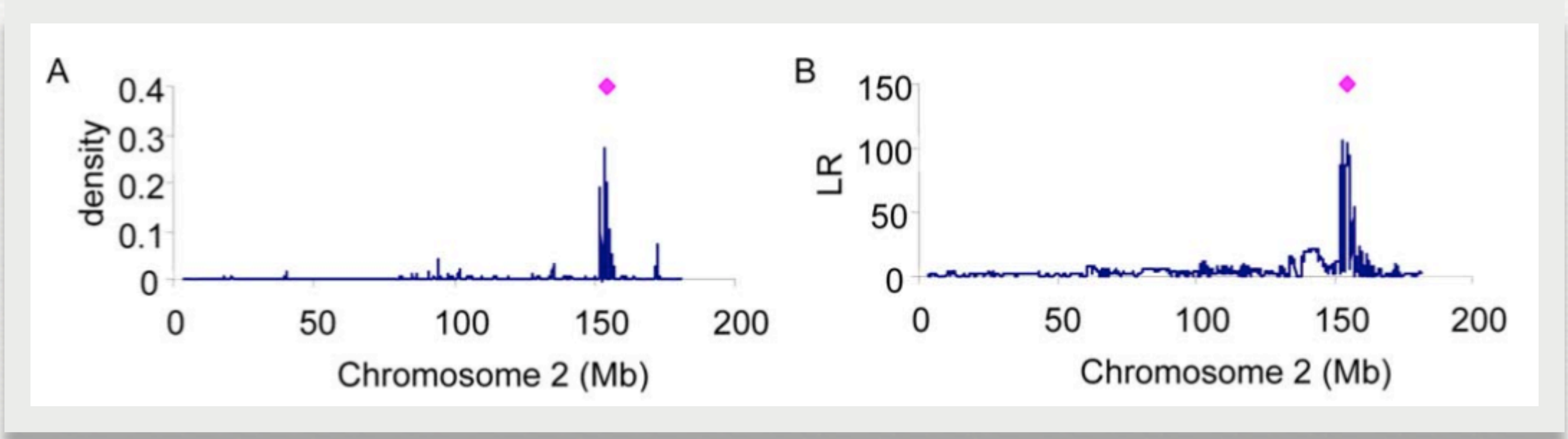
[a] Standard deviation over 10 replicates.

doi:10.1371/journal.pgen.1000231.t003

RESULTS

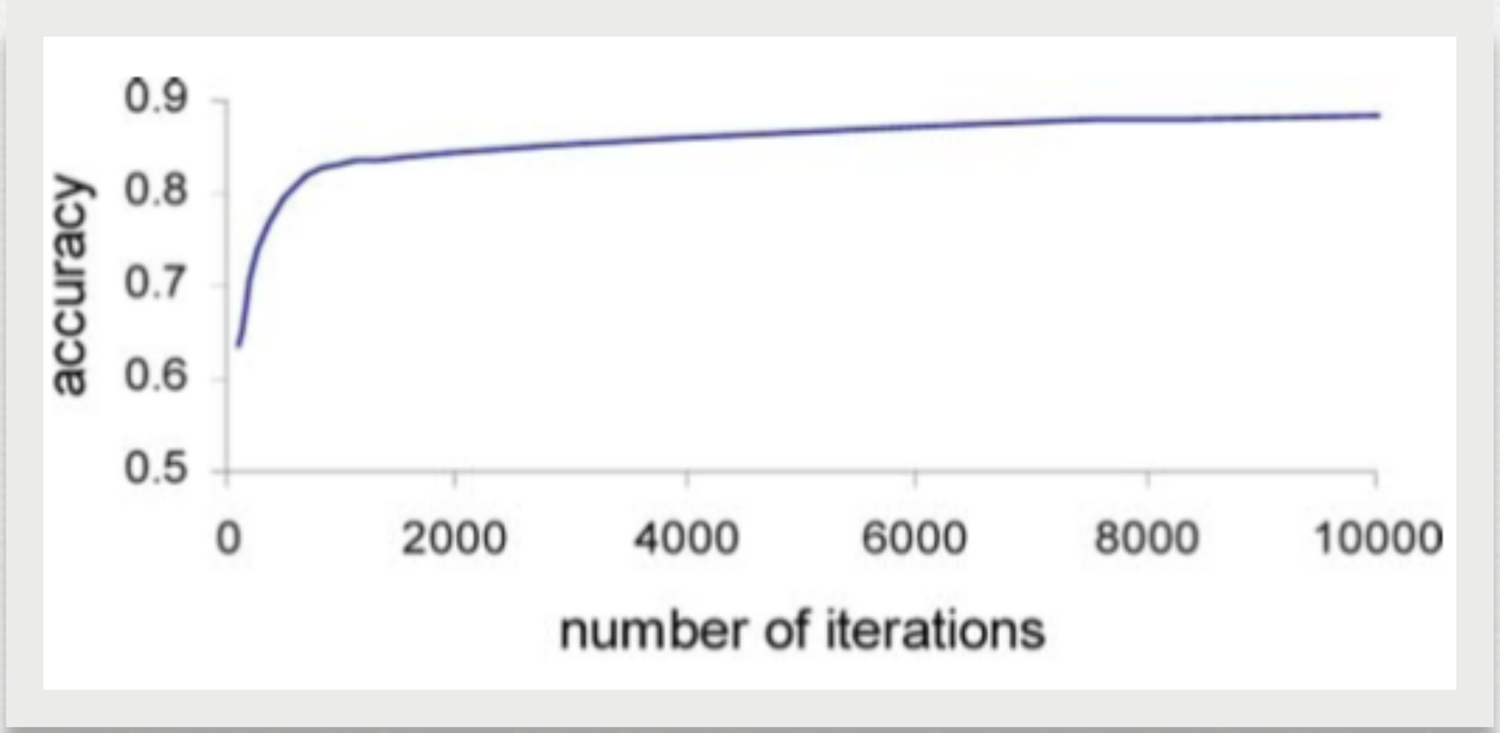
- Accuracy of prediction is higher when considering whole genome information instead of using one chromosome at a time

- Whole-genome approach and RJMCMC provides posterior density of each SNP being associated with the phenotype, so the positions of trait loci can be estimated



RESULTS

- Converge of parameter estimates
    - Accuracy rapidly increases in early rounds, stable after 10,000 iterations



DISCUSSION

- Study proposed a method to simultaneously analyse whole genome SNP data for association with phenotypes, applied this method to three traits measured in a heterogeneous mouse stock and successfully predicted unobserved phenotypes.
- The prediction of unobserved phenotypes for complex traits from genome-wide marker data is feasible and can be accurate
- Applications of the method are plentiful

THANKS