Genet. Res., Camb. (2010), **92**, *pp.* 115–125. © Cambridge University Press 2010
doi:10.1017/S0016672310000091

Including copy number variation in association studies to predict genotypic values

M. P. L. CALUS[1]*, D. J. DE KONING[2] AND C. S. HALEY[2,3]
[1] *Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB, Lelystad, The Netherlands*
[2] *Division of Genetics and Genomics, Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin EH25 9PS, UK*
[3] *MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK*

(*Received 4 June 2009 and in revised form 9 February 2010*)

Summary

The objective of this study was to investigate, both empirically and deterministically, the ability to explain genetic variation resulting from a copy number polymorphism (CNP) by including the CNP, either by its genotype or by a continuous derivation thereof, alone or together with a nearby single nucleotide polymorphism (SNP) in the model. This continuous measure of a CNP genotype could be a raw hybridization measurement, or a predicted CNP genotype. Results from simulations showed that the linkage disequilibrium (LD) between an SNP and CNP was lower than LD between two SNPs, due to the higher mutation rate at the CNP loci. The model R^2 values from analysing the simulated data were very similar to the R^2 values predicted with the deterministic formulae. Under the assumption that x copies at a CNP locus lead to the effect of x times the effect of 1 copy, including a continuous measure of a CNP locus in the model together with the genotype of a nearby SNP increased power to explain variation at the CNP locus, even when the continuous measure explained only 15% of the variation at the CNP locus.

1. Introduction

The application of genome-wide association (GWA) studies has become increasingly common, due to the availability of genome-wide dense marker maps at relatively low-cost. GWA studies typically can have two main objectives. The first objective is to derive the position of a gene or a genomic region that has an influence on one or more traits of interest. Common examples are identification of disease loci in human (Thomson, 1995), or identification of quantitative trait loci (QTLs) in livestock (Weller *et al.*, 1990; Andersson & Georges, 2004). The second objective of GWA is to predict the genetic potential or phenotype of an individual for a certain trait. Examples are estimation of breeding values in livestock to enable genomic selection (Meuwissen *et al.*, 2001), or prediction of the (genetic) susceptibility of an individual for a disorder or disease (Wray *et al.*, 2007). Generally, the applied models for both types of GWA studies may be the same, although fine-tuning for one of both objectives may result in subtle differences in the applied models (Calus *et al.*, 2009).

GWA studies are typically performed using markers such as single nucleotide polymorphisms (SNPs) that represent a sample of the variation in the genome. Another source of structural genomic variation is in the form of differences between different individuals in numbers of copies of genomic regions, referred to as copy number variation (CNV) or copy number polymorphisms (CNPs). Recent studies in human genetics have revealed that CNV may underlie an appreciable amount of variation at the trait level (Khaja *et al.*, 2006; Locke *et al.*, 2006). Moreover, it has been shown that CNVs can be associated with disease susceptibility and that disease genes are located in CNV regions (Sebat *et al.*, 2004; Blasko *et al.*, 2007; Kehrer-Sawatzki, 2007; Zhang *et al.*, 2009). GWA studies typically use dense SNP maps to associate genetic variation with genomic regions.

* Corresponding author: Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, P. O. Box 65, 8200 AB, Lelystad, The Netherlands. Tel: 31 320 238265. Fax: 31 320 293591. e-mail: mario.calus@wur.nl

Considering that CNPs may be directly associated with phenotypic variation, an important question is whether this phenotypic variance can also be captured using a dense SNP map, or whether CNPs should be genotyped and included in GWA studies. In the situation where CNPs and SNPs are located in the same regions and therefore physically closely related, the difference in mutation rates and number of alleles between the two types of loci may still result in relatively low linkage disequilibrium (LD) between both types of loci. Genotyping of CNP loci may be straightforward for loci with only two alleles, each representing a different number of segregating copies, but may be difficult for loci with more than two segregating alleles (Locke *et al*., 2006). A proposed solution to this problem is to use raw (continuous) hybridization intensities at those CNP loci, rather than derived (discrete) genotypes to provide an estimate of the number of copies (over both gametes) in an individual (Locke *et al*., 2006). In addition to measuring the CNP genotypes for all individuals, they may be predicted for some individuals. The predicted number of copies (over both gametes) could also be used as a continuous measure of the CNP locus.

The objective of this study was to investigate, both empirically and deterministically, the ability to explain genetic variation resulting from a CNP by including the CNP, either by its genotype or by a continuous derivation thereof, alone or together with a nearby SNP in the model.

2. Material and methods

Formulae are derived to predict the captured variance at CNP loci when CNP genotypes are not measured with 100% accuracy. Predictions from those formulae are compared to R^2 values from simulated data. To derive reasonable distributions of LD between SNP and CNP loci, data were simulated with two segregating CNP and several segregating SNP loci. The derived distributions were used to gain insight into the LD between SNP and CNP loci, and to inform additional simulations to investigate the possibility to associate genetic variance caused by a CNP locus with allelic variation of a linked SNP. Additionally, the benefit of including CNP phenotypes, i.e. a continuous measure or prediction for CNP genotypes, or CNP genotypes in the model was investigated.

(i) *Simulations to estimate associations between segregating SNP and CNP loci*

An important characteristic that we want to derive is the association between CNP and SNP loci. A measure of the association between a biallelic (SNP) locus and a multiallelic locus (in this case a CNP), was presented by Zhao *et al*. (2005) as

$$r^2 = \frac{\sum_{i=1}^{k} p(A_i)[p(Q|A_i) - p(Q)]^2}{p(Q)[1 - p(Q)]},$$

where A_i is one of k alleles at the multiallelic (CNP) locus and Q is one of the two alleles at the biallelic (SNP) locus. Considering this formula, the association between an SNP and a CNP locus depends on the number of alleles at the CNP locus, the allele frequencies at the SNP and CNP loci, and the frequencies of haplotypes consisting of CNP and SNP alleles. These parameters, in turn, depend on the population history and the mutation rates at both loci. Therefore, simulation of a large number of replicated datasets will give insight into the distribution of r^2 values between SNP and CNP loci, for the simulated population history and mutation rates.

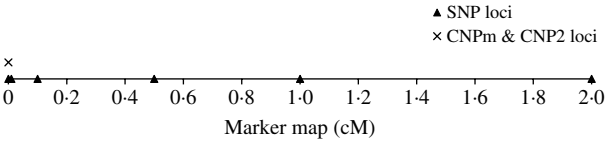

Fig. 1. The simulated marker map, with one CNP locus with only 2 alleles (CNP2), one CNP locus with two or more alleles (CNPm), and SNP loci at respectively 0·0, 0·1, 0·5, 1·0 and 2·0 cM distance with 10 SNPs at each locus.

To derive the distribution of LD between an SNP and a CNP, d cM apart, we simulated a population with 500 individuals that were randomly mated for 4000 generations. Twenty CNP loci were simulated on the same position and 70 SNP loci; 10 SNPs at each of seven positions located at 0·0, 0·0, 0·01, 0·1, 0·5, 1·0 and 2·0 cM distance of the CNP loci (Fig. 1). Two sets of SNPs were simulated at a distance of 0·0 to allow estimation of SNP–SNP LD at 0·0 cM distance. All CNP and SNP loci had alleles 1 and 2 segregating in the first generation, where both alleles were drawn per individual and locus with equal chance. Segregating loci in the first generation combined with 4000 generations of random mating ensure reaching a mutation-drift balance. The applied mutation rate for the SNP loci was 2×10^{-8} per haploid locus per generation (e.g. Drake *et al*., 1998; Kumar & Subramanian, 2002). Initially, SNP alleles were coded as 1 or 2. A mutation on an SNP locus changed an allele 1 (2) to become 2 (1). The applied mutation rate for the CNP loci was 10^{-4} per haploid locus per generation based on several reported estimates (Tusieluna & White, 1995; Nuzhdin, 1999; Shaffer & Lupski, 2000; van Ommen, 2005; Repping *et al*., 2006). For 10 CNP loci, it was assumed that a mutation event caused the number of copies to decrease or increase by one copy with equal probability. Whenever a mutation occurred at a locus with 0 copies, the only possible outcome was 1 copy. These

https://doi.org/10.1017/S0016672310000091 Published online by Cambridge University Press

CNPs are from here on referred to as CNPm loci. For the other 10 CNPs, as well as for the SNP loci, a mutation of allele 1 always resulted in allele 2, and a mutation of allele 2 always resulted in allele 1. This latter type of CNP therefore represented an SNP with the assumed mutation rate of a CNP. These CNPs are from here on referred to as CNP2. Recombination, based on Haldane's mapping function, was considered among all loci.

In total, 100 000 replicates were simulated. From those replicates, distributions for allele frequencies at SNP and CNP loci, r^2 values between SNP and CNP loci, and the number of alleles at CNP loci, were derived for each of the six considered distances, using the simulated genotypes from the last generation (i.e. using 500 individuals).

(ii) *Simulation of phenotypes with a CNP and an SNP*

The first set of simulations yielded a large number of haplotypes combining alleles from an SNP locus and a CNP locus. The obtained distributions of the frequencies of those haplotypes, in which a CNPm locus was included, were used to simulate six sets of data with one segregating SNP and one segregating CNP locus, again with different distances (0·0, 0·01, 0·1, 0·5, 1·0 and 2·0 cM) between those loci. Those simulated datasets in turn were used to evaluate the accuracy and bias of models including different combinations of SNP and CNP information (as explained in the next section). Each simulated dataset contained 500 individuals. Each individual received haplotypes, i.e. combinations of SNP and CNP alleles, with probabilities equal to the haplotype frequencies from the previous simulations. This ensured that this simulation represented the original simulated population. Each CNP allele received an effect on the phenotype of the individuals such that the CNP locus explained 10% of the total phenotypic variance. The remaining 90% of the phenotypic variation was explained by a residual effect, drawn for each individual from $N\sim(0,\ 0{\cdot}90)$. For biallelic CNP, the allele substitution effect was calculated using $\sigma^2_{\text{cnp}} = 2p(1-p)a^2$ (Falconer & Mackay, 1996), where σ^2_{cnp} is the (simulated) variance explained by the CNP locus (kept at 0·10 in this case to ensure consistency across replicates), p is the allele frequency of one of both alleles at the locus and a is the allele substitution effect. a was calculated per replicate such that the variance was constant across replicates. Assuming that the effect of z copies was zx, and that on a biallelic CNP locus one segregating allele consisted of b and the other of c copies, a^2 in the above formula was replaced by $(c-b)^2x^2$.

For CNP with more than two alleles, the variance was written in terms of the number of copies nc_j that individual j carried at the CNP locus, in a population of n individuals, and again assuming that the effect of z copies was zx:

$$\sigma^2_{\text{cnp}} = \frac{\sum_{j=1}^{n}\text{nc}_j^2x^2 - \left(\sum_{j=1}^{n}\text{nc}_j x\right)^2/n}{n-1}.$$

Solving for x yields:

$$x = \sqrt{\frac{(n-1)\sigma^2_{\text{cnp}}}{\sum_{j=1}^{n}\text{nc}_j^2 - \left(\sum_{j=1}^{n}\text{nc}_j\right)^2}}.$$

CNP phenotypes, mimicking raw hybridization levels or predicted CNP genotypes, were simulated using the following model:

$$\text{CNPphen} = \text{CNPgen} + e,$$

where CNPgen is the sum of the CNP alleles, reflecting the total true number of copies at this locus and e is drawn from a distribution $N(0,\ \sigma^2_{\text{cnp}}/h^2_{\text{CNPp,CNPg}} - \sigma^2_{\text{cnp}})$. The heritability of the CNP phenotype, $h^2_{\text{CNPp,CNPg}}$, was varied from 0·05 to 0·95, and represents the squared correlation between the simulated CNP phenotypes and CNP genotypes. Consequently, a high (low) value of $h^2_{\text{CNPp,CNPg}}$ means that the CNP phenotype predicts the CNP genotype with high (low) accuracy.

(iii) *Analyses to predict the effect of the CNP locus*

To assess the ability to predict the effect of the CNP locus with different sources of information in the model, we considered the following five models:

$$y_i = \mu + \beta \times \text{snp}_i + e_i, \quad (1)$$

$$y_i = \mu + \delta \times \text{cnpg}_i + e_i, \quad (2)$$

$$y_i = \mu + \beta \times \text{snp}_i + \delta \times \text{cnpg}_i + e_i, \quad (3)$$

$$y_i = \mu + \gamma \times \text{cnpp}_i + e_i, \quad (4)$$

$$y_i = \mu + \beta \times \text{snp}_i + \gamma \times \text{cnpp}_i + e_i, \quad (5)$$

where y_i is a phenotypic record of individual i, μ is an average phenotypic effect, β is the regression coefficient on the genotype snp_i at the SNP locus (0 for homozygotes 11, 1 for heterozygotes and 2 for homozygotes 22), δ is the regression coefficient on the number of copies cnpg_i at the CNP locus, γ is the regression coefficient on CNP phenotype cnpp_i and e_i is a random residual. All analyses were performed using ASReml (Gilmour *et al.*, 2006).

(iv) *Model comparison*

The different proposed models were compared for their ability to estimate the effect of the CNP

https://doi.org/10.1017/S0016672310000091 Published online by Cambridge University Press

genotype. To assess the accuracy of the predicted genotype effects, the mean-squared correlation between the predicted genotype effect and the simulated genotype effect was calculated for each of the five models. To assess the bias of the predicted genotype effects, the mean-squared error of the prediction (MSEP) of the genotype was calculated for each of the five models. The simulated (true) genotypic effect was per individual calculated as the sum of the simulated effects of its alleles at the CNP locus. Estimates per individual were derived as the sum of estimates of its SNP genotype (model 1), CNP genotype (model 2), SNP and CNP genotypes (model 3), CNP phenotype (model 4) or CNP phenotype and SNP genotype (model 5).

To gain more insight into the predictive ability of CNP phenotypes compared to SNP genotypes, the r^2 values between SNP and CNP genotypes were compared to r^2 values between SNP genotypes and CNP phenotypes. The r^2 values between SNP genotypes and CNP phenotypes were calculated as the squared correlation coefficient between the recoded SNP genotypes (0, 1 or 2) and the CNP phenotypes.

(v) *Theory: deterministic derivation of model* R^2

To allow direct prediction of the model R^2 for each of models 1–5, deterministic formulae were derived. Multiple coefficients of determination, i.e. R^2 values, between CNPg and each of the four (combinations of) explanatory variables were derived as follows. For model 1,

$$R^2(\text{CNPg, SNP}) = r^2(\text{CNPg, SNP}), \tag{6}$$

where r^2(CNPg, SNP) is calculated by the formula presented by Zhao *et al.* (2005).
For model 2,

$$R^2(\text{CNPg, CNPg}) = 1{\cdot}0. \tag{7}$$

For model 4, it was assumed that a CNP phenotype was measured with a certain heritability $h^2_{\text{CNPp,CNPg}}$ (here denoted as h^2). Since r(CNPg, CNPph) is equal to h,

$$R^2(\text{CNPg, CNPph}) = h^2. \tag{8}$$

For models 3 and 5, the following general formula is used, which calculates the multiple coefficients of determination for n loci that are used to predict the variation that is associated with a locus (Bastiaansen *et al.*):

$$R^2 = c'K^{-1}c,$$

where c is an $n \times 1$ vector that contains values of r (i.e. the correlation) between each of the loci included in the analysis and the predicted locus and K is an $n \times n$ square matrix with values of r between each pair of predicting loci on the off-diagonal elements and values of 1 on the diagonal. Thus, for model 3,

$$c' = [r(\text{SNPg, CNPg}) \quad 1]$$

and

$$K = \begin{bmatrix} 1 & r(\text{SNPg, CNPg}) \\ r(\text{SNPg, CNPg}) & 1 \end{bmatrix}$$

yielding that

$$R^2 = 1{\cdot}0. \tag{9}$$

For model 5,

$$c' = [r(\text{SNPg, CNPg}) \quad r(\text{CNPph, CNPg})] = [r(\text{SNPg, CNPg}) \quad h]$$

and

$$K = \begin{bmatrix} 1 & r(\text{SNPg, CNPph}) \\ r(\text{SNPg, CNPph}) & 1 \end{bmatrix} = \begin{bmatrix} 1 & r(\text{SNPg, CNPg}) \times h \\ r(\text{SNPg, CNPg}) \times h & 1 \end{bmatrix}$$

yielding, after rearranging, that

$$R^2 = \frac{(1-2h^2) \times r^2(\text{SNPg, CNPg}) + h^2}{1 - h^2 r^2(\text{SNPg, CNPg})}. \tag{10}$$

3. Results

(i) *Allele frequencies of SNP* versus *CNP*

Of all segregating loci in the first set of simulations, one CNPm, one CNP2 and one SNP locus at each of the seven distances were randomly selected and used in the analysis. At some of the positions none of the SNP loci were segregating after 4000 generations, leading to a total of 40 184 replicates with segregating SNP (out of 100 000) that were retained for analysis. The CNPm loci mainly had 2, 3 or 4 and only rarely 5 segregating alleles (Table 1). For all CNPm loci, the alleles consisting of 1 and 2 copies were segregating with the highest frequency.

Minor allele frequencies indicated that the allele frequencies at SNP, CNP2 and CNPm loci with 2 alleles were similar (Table 2). The U-shaped distribution of the allele frequencies confirmed the similarity between CNP2 (Fig. 2) and SNP (not shown), albeit that the frequency of rare alleles was lower for SNPs. Note that grouping 0 and 1 copies and 2 and more copies for CNPm loci with 2 alleles yields a similar distribution as the CNP2 loci (Fig. 2). With 3 or more copies, the frequency at the CNPm loci of the higher numbers of copies increased (Table 1). This

https://doi.org/10.1017/S0016672310000091 Published online by Cambridge University Press

Table 1. *Distribution of number of segregating alleles at simulated CNPm loci*

Total no. of alleles CNPm			2	3	4	5
Percentage of replicates			67·73	29·76	2·44	0·07
	Frequency of numbers of copies at CNP loci					
No. of segregating alleles at CNP locus	0	1	2	3	4	5
2	9·54	42·31	40·66	6·98	0·49	0·02
3	5·94	41·02	41·81	9·99	1·16	0·08
4	7·47	28·89	39·57	19·56	4·02	0·48
5	10·46	25·86	17·61	28·35	15·93	1·68

Table 2. *Average minor allele frequencies (MAFs) across segregating loci, in ascending order*

Locus type	MAF(1)	MAF(2)	MAF(3)	MAF(4)	MAF(5)
SNP	0·245	0·755			
CNP2	0·132	0·868			
CNPm, 2 alleles	0·111	0·889			
CNPm, 3 alleles	0·026	0·161	0·813		
CNPm, 4 alleles	0·011	0·064	0·251	0·675	
CNPm, 5 alleles	0·004	0·023	0·090	0·259	0·623

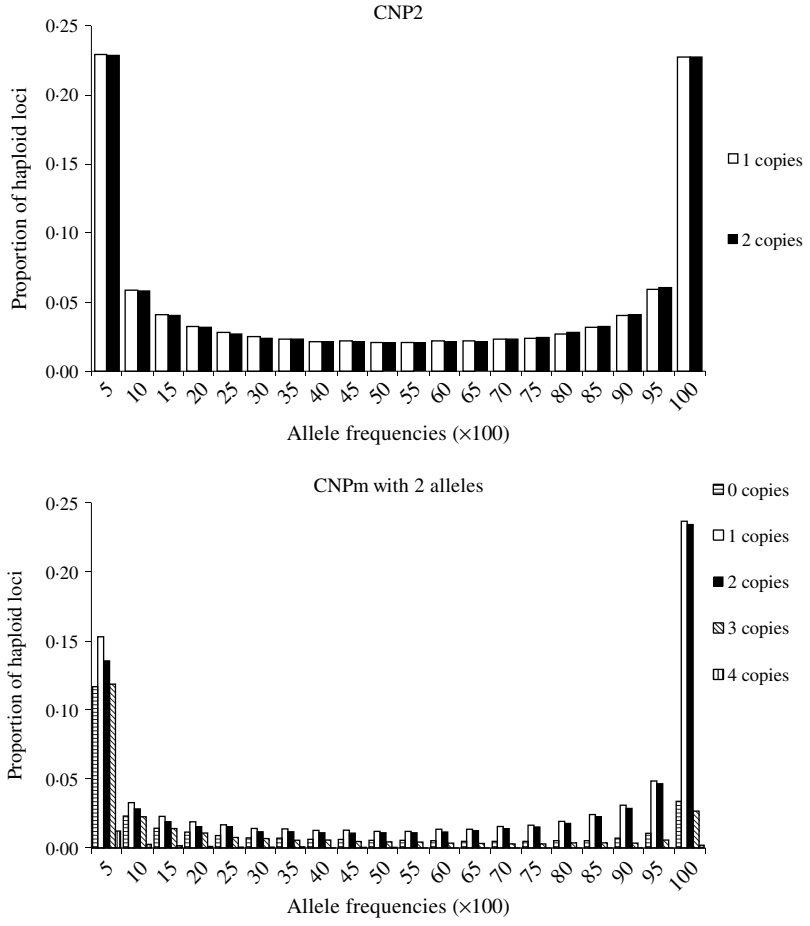

Fig. 2. Average frequencies of alleles across all CNP2 loci and CNPm loci with 2 alleles in generation 4000.

https://doi.org/10.1017/S0016672310000091 Published online by Cambridge University Press

Table 3. *Estimated* r^2 *values between an SNP and an SNP, CNP2 or CNPm loci, and the predicted* r^2 *between two bi-allelic loci according to Sved (1971), located at different distances*

r^2	Distance to SNP (cM)					
Method	0	0·01	0·1	0·5	1·0	2·0
Sved	1·000	0·833	0·333	0·091	0·048	0·024
SNP	1·000	0·885	0·336	0·060	0·026	0·011
CNP2	0·451	0·374	0·129	0·039	0·019	0·010
CNPm(2)	0·504	0·394	0·131	0·035	0·018	0·010
CNPm(3)	0·703	0·582	0·202	0·068	0·033	0·019
CNPm(4)	0·912	0·810	0·327	0·095	0·050	0·022
CNPm(5)	0·981	0·878	0·335	0·054	0·079	0·031

resulted in a distribution of allele frequencies that further deviated from the distribution of allele frequencies at an SNP locus (results not shown).

(ii) *Average LD between SNP and CNP loci*

The first set of simulations was also used to calculate LD between different loci at different distances. The average LD between two SNP loci, an SNP and a CNP2 locus and an SNP and a CNPm locus was calculated at all six distances, as well as the expected LD between two SNP loci based on the formula by Sved (1971) (Table 3). The LD between two SNPs was generally close to the expectation. Increased mutation rates, that is CNP2 and CNPm(2) compared to SNP loci, led to lower LD with the nearby SNP and larger deviation of the LD from its expected value. An increase in the number of alleles at the CNPm locus resulted in higher LD with the nearby SNP at all distances.

(iii) *Deterministic* R^2 *values and estimated MSEP of different models*

The second set of simulations, based on the haplotype frequencies of the first set, was used to compare deterministically predicted versus obtained model R^2 values. Model R^2 values obtained from analysing the simulated data using models 1, 2, 3, 4 and 5 were similar to those calculated using, respectively, formulae (6), (7), (8), (9) and (10) (Table 4). The small differences are such that generally the R^2 values based on the analysis are smaller than the predicted R^2 values. Only at a distance of 0·0 cM, the model including the SNP and CNP phenotypes always yielded higher R^2 values than the predicted values (Table 4). The R^2 values show that including an SNP in the model, in addition to a CNP phenotype, increases model R^2, when the heritability of the CNP phenotype $<0{\cdot}5$, and the distance between the CNP and SNP is short ($\sim <0{\cdot}5$ cM). The predicted R^2 values for models including only CNP phenotypes (using formula (8)) or CNP phenotypes and SNP genotypes (using formula (10)), were plotted as a function of h^2 of the CNP phenotypes for different levels of LD between CNP and SNP loci (Fig. 3). This figure also shows that the gain in R^2 due to including the SNP locus was substantial, depending on the r^2 between the SNP and CNP loci.

The MSEP across models 1–5 was clearly largest (i.e. the bias was greatest) when only the SNP genotype or the CNP phenotype with very low h^2 (with or without the SNP genotype) was included in the model (Table 5). Lowest MSEP was found when only the CNP genotype or the CNP phenotype with very high h^2 was included in the model. Including the SNP genotype in addition to the CNP genotype or phenotype in the model hardly changed the MSEP.

4. Discussion

The objective of this study was to investigate, both empirically and deterministically, the ability to explain genetic variation resulting from a CNP by including the CNP, either by its genotype or by a continuous derivation thereof, alone or together with a nearby SNP in the model. The model R^2 values from analysing the simulated data were very similar to the values predicted with the deterministic formulae. The results indicated that using CNP phenotypes in the model next to a nearby SNP can increase the power of the model substantially, when CNP genotypes cannot easily be derived. It should be noted that the heritability of the CNP phenotype can be interpreted as the reliability of measuring or predicting the CNP genotype. This means that the presented formulae also apply for situations where CNP genotypes are predicted for groups of individuals with a given reliability, conditional on known CNP genotypes in other related individuals.

(i) r^2 *(LD) between different loci*

In this study, we chose to evaluate LD between a CNP and an SNP locus, because SNPs are nowadays widely used as genetic markers in many species. In our analyses, we limited ourselves to including only one SNP in the model, while for instance in cattle nowadays $\sim$50 000 SNPs are used and in humans $\sim$1 000 000 SNPs are used. At a genome of 30 Morgan in length, this implies an average marker spacing of 1 SNP per 0·06 cM. On the cattle 50 k SNP chip, for Holstein the r^2 between adjacent loci is between 0·15 and 0·20, for an average distance of $\sim$0·06 cM (De Roos *et al.*, 2008; Khatkar *et al.*, 2008). In our simulation, after interpolation, we here find an r^2 value of $\sim$0·44 between two SNP loci,

https://doi.org/10.1017/S0016672310000091 Published online by Cambridge University Press

Table 4. *Realized and predicted model* R^2 *values for different models averaged across 1000 replicates*

			Distance CNP–SNP (cM)													
			0		0·01		0·1		0·5		1		2		−SNP[b]	
Model	Frm[a]	h^2 CNPph	an[a]	frm[a]	an	frm	an	frm	an	frm	an	frm	an	frm	an	frm
SNP	6		0·74	057	0·45	0·46	0·15	0·16	0·04	0·05	0·03	0·02	0·01	0·01		
CNP + SNP	9 (7)[c]		0·98	1	0·93	1	0·92	1	0·93	1	0·92	1	0·94	1	0·95	1
CNPph	10 (8)	0·95	0·95	0·95	0·89	0·95	0·87	0·95	0·89	0·95	0·87	0·95	0·89	0·95	0·90	0·95
+ SNP		0·85	0·91	0·88	0·83	0·87	0·79	0·85	0·79	0·85	0·78	0·85	0·80	0·85	0·81	0·85
		0·75	0·88	0·81	0·78	0·79	0·70	0·76	0·70	0·75	0·69	0·75	0·70	0·75	0·71	0·75
		0·65	0·86	0·76	0·73	0·73	0·62	0·67	0·61	0·66	0·60	0·65	0·61	0·65	0·62	0·65
		0·55	0·83	0·72	0·68	0·67	0·54	0·58	0·52	0·56	0·50	0·55	0·51	0·55	0·52	0·55
		0·45	0·81	0·68	0·63	0·63	0·46	0·50	0·43	0·46	0·41	0·46	0·42	0·45	0·43	0·45
		0·35	0·79	0·65	0·59	0·58	0·39	0·42	0·34	0·37	0·32	0·36	0·32	0·36	0·33	0·35
		0·25	0·77	0·63	0·54	0·54	0·32	0·34	0·25	0·28	0·23	0·26	0·23	0·26	0·24	0·25
		0·15	0·75	0·60	0·50	0·51	0·24	0·27	0·16	0·18	0·14	0·17	0·14	0·16	0·14	0·15
		0·05	0·73	0·58	0·45	0·48	0·17	0·19	0·08	0·09	0·06	0·07	0·05	0·06	0·05	0·05

[a] Predicted using the formulae (frm) or analysis of simulated data (an).
[b] The same model, but without the SNP. Values are averaged across all 6000 replicates (1000 per distance) for the values based on analysis of simulated data.
[c] The formula in brackets indicates the formula used for the last column (−SNP).

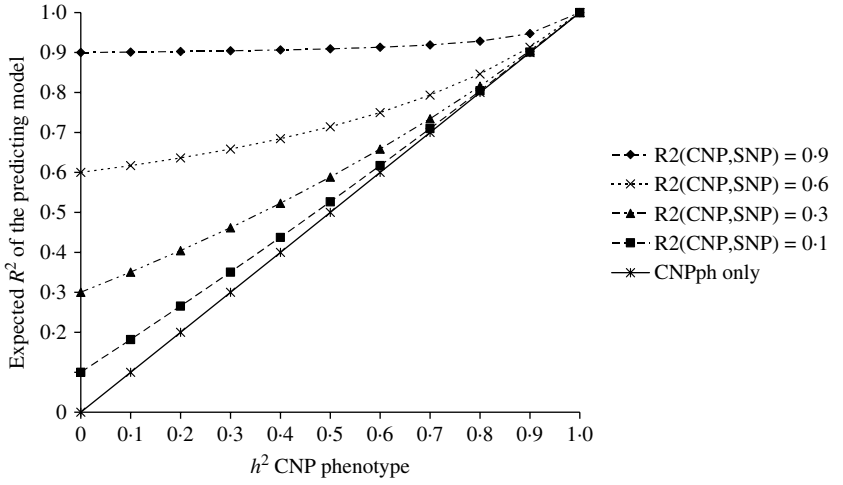

Fig. 3. Deterministic R^2 values obtained for models including CNP phenotypes and SNP genotypes assuming different r^2 values between CNP and SNP loci, as a function of h^2 of the CNP phenotypes.

indicating that our achieved LD within the ranges of considered distances is much higher than LD in the 50 k cattle SNP chip. Applications for cattle data may, however, consider multiple SNPs simultaneously. This would lead to explaining higher proportions of variation at the CNP locus then the expected value of 0·20 based on average LD between the adjacent SNPs, since the SNP with the highest LD with the CNP locus (not necessarily the closest SNP) would predict most phenotypic variance and therefore be favoured in an association study. Note that the presented formulae can easily be extended to include multiple SNPs by including more than one SNP in the vector c and matrix K.

In our simulated data an average r^2 of 0·20, that is the average expected value for the 50 k cattle SNP chip, is expected at a distance of 0·23 cM after interpolation. At this distance, after interpolation the r^2 between an SNP and a CNP locus was at least ~0·2. Using formula (10) indicates that by including the CNP phenotype in this scenario, the model R^2 could be increased from 0·2 to over 0·55, when the heritability of the CNP phenotype is at least 0·5.

Based on the 1 000 000 SNPs currently available on commercial human genotyping products, the expected distance to an unobserved CNP is expected to be on average a maximum of $1{\cdot}5 \times 10^{-3}$ Mb, here assumed to be equal to $1{\cdot}5 \times 10^{-3}$ cM. At such a distance, the

https://doi.org/10.1017/S0016672310000091 Published online by Cambridge University Press

Table 5. *MSEP for different models, averaged across 1000 replicates*

Model	h^2 CNPph	Distance CNP–SNP (cM)						
		0	0·01	0·1	0·5	1	2	−SNP[a]
SNP		0·029	0·053	0·082	0·094	0·094	0·098	
CNP + SNP		0·005	0·005	0·005	0·005	0·005	0·005	0·003
CNPph	0·95	0·008	0·008	0·010	0·010	0·010	0·010	0·009
+ SNP	0·85	0·011	0·014	0·018	0·019	0·019	0·020	0·017
	0·75	0·014	0·020	0·027	0·028	0·028	0·029	0·024
	0·65	0·017	0·025	0·034	0·038	0·038	0·039	0·032
	0·55	0·019	0·030	0·043	0·046	0·047	0·048	0·039
	0·45	0·022	0·035	0·050	0·056	0·056	0·057	0·046
	0·35	0·024	0·039	0·058	0·064	0·065	0·067	0·053
	0·25	0·026	0·044	0·065	0·073	0·074	0·076	0·060
	0·15	0·027	0·048	0·073	0·082	0·083	0·085	0·066
	0·05	0·030	0·053	0·080	0·091	0·092	0·095	0·073

[a] The result of this model, with the SNP excluded, averaged across all 6000 replicates.

r^2 between SNPs in the human genome is ∼0·25 (P. Navarro, personal communication). In our simulated data, this level of LD between two SNP loci was found at a distance of ∼0·17 cM after interpolation. Using formula (10) indicates that by including the CNP phenotype in this scenario, the model R^2 could be increased from 0·17 to over 0·53, when the heritability of the CNP phenotype is at least 0·5.

The differences between the loci considered here are that CNPm and CNP2 loci have a much higher mutation rate than an SNP locus, while a CNPm locus may have more than two alleles segregating. The results showed that the average LD of a CNPm locus with a nearby SNP always increased with increasing number of segregating alleles at the CNPm locus. The results also showed that loci with a higher mutation rate have lower LD with a linked SNP locus. Some studies have reported that microsatellites and short tandem repeats explain more variation at a nearby locus than SNPs do (Ohashi & Tokunaga, 2003; Varilo *et al*., 2003; Mueller, 2004; Payseur & Cutter, 2006). Both microsatellites and short tandem repeats are comparable to CNP, in the sense that their number of segregating alleles may be larger than two, and that their mutation rate is similar to that of CNP loci. Hinds *et al*. (2006) reported similar LD between deletion loci and SNPs compared to pairwise SNP LD. Payseur *et al*. (2008) reported that LD between short-tandem-repeat polymorphisms and SNP loci was lower than pairwise SNP LD, in agreement with our results. Summarized, an increased mutation rate leads to lower LD, when the number of segregating alleles is left unchanged. An increased mutation rate can, however, indirectly lead to increased LD, if it increases the number of segregating alleles. A mutation at a segregating locus may lead to breakdown of LD in the short run. When the mutation stays in the population for a longer time, genetic drift will re-establish LD. When the mutation leads to a new allele that was not segregating yet, our results show that the LD with a nearby SNP eventually ends up being on average higher than with fewer alleles segregating at the locus.

(ii) *Using CNP phenotypes instead of CNP genotypes*

In the simulated datasets, a range of heritabilities of the CNP phenotypes was considered. This applies to CNP loci whose clusters representing the different genotypes are not sufficiently distinct to allow derivation of discrete genotypes. Locke *et al*. (2006) calculated the heritability of CNP loci in two human subpopulations, for 17 loci per subpopulation, and reported an average heritability of 0·86, while the lowest value was only 0·15. This indicates that the whole range of considered heritabilities in our study may actually be present in real data, albeit that most CNP phenotypes are likely to have a heritability relatively close to 1·0.

The maximum average model R^2 value from a model including only an SNP was 0·74. Based on our results, this means that adding CNP phenotypes increases the model R^2 across all distances when the heritability of the CNP phenotypes is >0·05 (Table 4). To illustrate the relation between CNP genotypes and phenotypes when the heritability of CNP phenotypes is low (0·25 in this example), we simulated 500 individuals with one CNP locus with allele frequencies for 0, 1, 2, 3 and 4 copies as in Table 1 for CNPm(3) loci. Visual inspection of the results indicates that in this case a distinction in discrete CNP genotypes is not possible (Fig. 4). Therefore, in such situations including the raw hybridization or predicted CNP genotype in the model provides an opportunity to investigate whether the CNP locus is associated with a trait or disease of interest.

https://doi.org/10.1017/S0016672310000091 Published online by Cambridge University Press

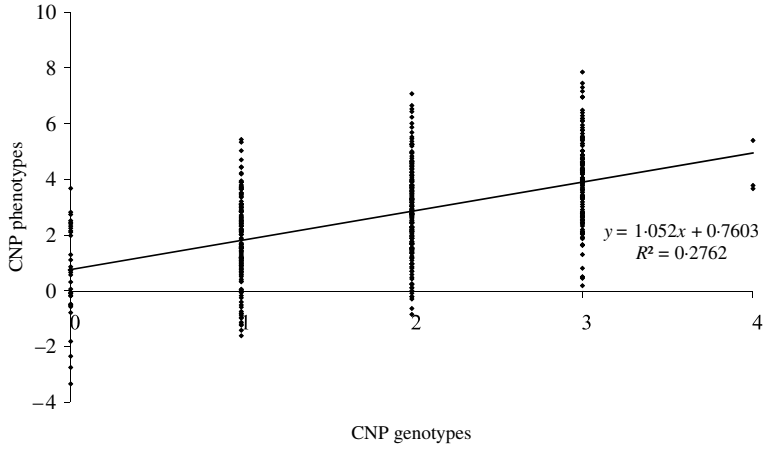

Fig. 4. Simulated CNP phenotypes, assuming a heritability of 0·25, plotted against the CNP genotypes for 500 individuals with a CNP locus with allele frequencies for 0, 1, 2, 3 and 4 copies as in Table 1 for CNPm(3) loci.

(iii) *The effect of a CNP locus on the phenotype*

In the simulations, it was assumed that the effect of a CNP locus on the phenotype was linearly related with the number of copies at the locus. For practical situations where CNP genotypes cannot be derived, this assumption of linearity allows one to define a general model to test for associations between a CNP locus and a phenotype. Whenever an association is found, the nature of the association can be further investigated by comparing the fit of additional models with non-linear regressions on the CNP phenotype.

Although several studies associate CNP loci with appreciable genetic variation and the expression of genes (e.g. Orozco *et al*., 2009; Zhang *et al*., 2009), still too little is known to make an estimate of the distribution of effects of CNP loci on the phenotype. The results in Tables 3 and 4 apply to a situation where the CNP genotype explains 10% of the phenotypic variance. Causal effects of most CNP loci are likely to be (much) lower. Consider that the R^2 of predicting an effect of a causal locus, by a linked SNP marker, is equal to the product of the r^2 (LD) between the marker and the causal locus (here: r^2(SNPg,CNPg)) and the squared accuracy of the predicted marker effect (Goddard, 2009). In the derived prediction formulae for the model R^2 values it is assumed that the accuracy of the predicted marker effect is 1·0. For a model including only one locus, the squared accuracy of its estimated effect can be calculated as follows (Daetwyler *et al*., 2008; Goddard, 2009):

$$r^2_{\text{locus}} = \frac{\lambda h^2}{\lambda h^2 + 1},$$

where $\lambda = n_p/n_G$, n_p is the number of phenotypes (500), n_g is the number of effective loci (considering that 1 locus is included in the model) and h^2 is the heritability, in this case the variance explained by the CNP locus divided by the phenotypic variance (which reduces to σ^2_{cnp}). Following this equation, r^2_{locus} is 0·98 for our simulations. Note that the value of 0·98 is somewhat higher still than the obtained model R^2 value for the model including only the CNP genotype (0·95; Table 4). r^2_{locus} would reduce to 0·96 and 0·83, when σ^2_{cnp} explains, respectively, 5 or 1% of the total phenotypic variance. This means that when changing σ^2_{cnp} to 1% of the phenotypic variance, the obtained model R^2 values for the model only including the CNP or SNP genotype or the CNP phenotype, are expected to be 0·83/0·98 = 0·85 times the obtained values reported in Table 4. The above formula for r^2_{locus} can also be used to derive the impact of using different numbers of phenotypes in the predictions. Thus, the presented formulae can easily be extended to predict the model R^2 values for different scenarios.

5. Conclusion

The simulations showed that an increased mutation rate leads to lower LD, whereas an increased number of segregating alleles at a locus leads to increased LD.

Under the assumption that x copies at a CNP locus lead to the effect of x times the effect of 1 copy, including the raw hybridizations or predictions of CNP genotypes in the model together with the genotype of a nearby SNP increased power to explain variation at the CNP locus, even when the continuous measure for the CNP explained only 15% of the variation at the CNP locus.

M. P. L. C. thanks John Bastiaansen, Henk Bovenhuis, Mari Smits and Roel Veerkamp for initial discussions on this study, as well as W. G. Hill and two anonymous referees for comments on the performed simulations and earlier versions of the manuscript. The EC-funded Integrated Project SABRE (EC contract number FOOD-CT-2006-01625) is acknowledged for financial support of the stay

of M. P. L. C. at the Roslin Institute. D. J. K and C. S. H. acknowledge support from BBSRC through the ISPG grant to the Roslin Institute and the EC-funded Integrated Project SABRE (EC contract number FOOD-CT-2006-01625) D. J. K. additionally acknowledges support from the EC-funded network of excellence EADGENE (EC contract number FOOD-CT-2004-506416) and C. S. H. additionally acknowledges support from the MRC.

References

Andersson, L. & Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics* **5**, 202–212.

Bastiaansen, J. W. M., Calus, M. P. L., de Roos, A. P. W. & Bovenhuis, H. (2010). Predicting the detectable proportion of QTL variation using linkage disequilibrium between genotyped markers. *Submitted to Genetics Selection Evolution*.

Blasko, B., Szeplaki, G., Varga, L., Ronai, Z., Prohaszka, Z., Sasvari-Szekely, M., Visy, B., Farkas, H. & Fust, G., (2007). Relationship between copy number of genes (C4A, C4B) encoding the fourth component of complement and the clinical course of hereditary angioedema (HAE). *Molecular Immunology* **44**, 2667–2674.

Calus, M. P. L., Meuwissen, T. H. E., Windig, J. J., Knol, E. F., Schrooten, C., Vereijken, A. L. J. & Veerkamp, R. F., (2009). Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genetics Selection Evolution* **41**, 11.

Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, e3395.

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

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics* **148**, 1667–1686.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Essex, UK: Longman Group.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. & Thompson, R. (2006). *ASReml User Guide Release 2.0*. Hemel Hempstead, HP1 1ES, UK: VSN International Ltd.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245–257.

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. Y. & Frazer, K. A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genetics* **38**, 82–85.

Kehrer-Sawatzki, H. (2007). What a difference copy number variation makes. *BioEssays* **29**, 311–313.

Khaja, R., Zhang, J. J., MacDonald, J. R., He, Y. S., Joseph-George, A. M., Wei, J., Rafiq, M. A., Qian, C., Shago, M., Pantano, L., Aburatani, H., Jones, K., Redon, R., Hurles, M., Armengol, L., Estivill, X., Mural, R. J., Lee, C., Scherer, S. W. & Feuk, L. (2006). Genome assembly comparison identifies structural variants in the human genome. *Nature Genetics* **38**, 1413–1418.

Khatkar, M. S., Nicholas, F. W., Collins, A. R., Zenger, K. R., Al Cavanagh, J., Barris, W., Schnabel, R. D., Taylor, J. F. & Raadsma, H. W. (2008). Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* 2008, **9**, 187, doi: 10.1186/1471-2164-9-187.

Kumar, S. & Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 803–808.

Locke, D. P., Sharp, A. J., McCarroll, S. A., McGrath, S. D., Newman, T. L., Cheng, Z., Schwartz, S., Albertson, D. G., Pinkel, D., Altshuler, D. M. & Eichler, E. E. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American Journal of Human Genetics* **79**, 275–290.

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

Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics* **5**, 355–364.

Nuzhdin, S. V. (1999). Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica* **107**, 129–137.

Ohashi, J. & Tokunaga, K. (2003). Power of genome-wide linkage disequilibrium testing by using microsatellite markers. *Journal of Human Genetics* **48**, 487–491.

Orozco, L. D., Cokus, S. J., Ghazalpour, A., Ingram-Drake, L., Wang, S., van Nas, A., Che, N., Araujo, J. A., Pellegrini, M. & Lusis, A. J. (2009). Copy number variation influences gene expression and metabolic traits in mice. *Human Molecular Genetics* **18**, 4118–4129.

Payseur, B. A. & Cutter, A. D. (2006). Integrating patterns of polymorphism at SNPs and STRs. *Trends in Genetics* **22**, 424–429.

Payseur, B. A., Place, M. & Weber, J. L. (2008). Linkage disequilibrium between STRPs and SNPs across the human genome. *American Journal of Human Genetics* **82**, 1039–1050.

Repping, S., van Daalen, S. K. M., Brown, L. G., Korver, C. M., Lange, J., Marszalek, J. D., Pyntikova, T., van der Veen, F., Skaletsky, H., Page, D. C. & Rozen, S. (2006). High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nature Genetics* **38**, 463–467.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M. Y., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A. & Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528.

Shaffer, L. G. & Lupski, J. R. (2000). Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annual Review of Genetics* **34**, 297–329.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125–141.

Thomson, G. (1995). Mapping disease genes - family-based association studies. *American Journal of Human Genetics* **57**, 487–498.

Tusieluna, M. T. & White, P. C. (1995). Gene conversions and unequal crossovers between Cyp21 (Steroid 21-Hydroxylase Gene) and Cyp21p involve different mechanisms. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 10796–10800.

van Ommen, G. J. B. (2005). Frequency of new copy number variation in humans. *Nature Genetics* **37**, 333–334.

Varilo, T., Paunio, T., Parker, A., Perola, M., Meyer, J., Terwilliger, J. D. & Peltonen, L. (2003). The interval of

linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Human Molecular Genetics* **12**, 51–59.

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

Wray, N. R., Goddard, M. E. & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17**, 1520–1528.

Zhang, F., Gu, W. L., Hurles, M. E. & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* **10**, 451–481.

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