Genet. Res., Camb. (2012), **94**, *pp.* 191–203.
doi:10.1017/S0016672312000365

Comparing linkage and association analyses in sheep points to a better way of doing GWAS

KATHRYN E. KEMPER[1]*, HANS D. DAETWYLER[2,3], PETER M. VISSCHER[4,5] AND MICHAEL E. GODDARD[1,2]

[1] *Department of Agriculture and Food, University of Melbourne, Parkville, Victoria 3010, Australia*
[2] *Victorian Department of Primary Industries, AgriBiosciences Centre, LaTrobe Research and Development Park, Bundoora, Victoria 3083, Australia*
[3] *Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW, 2351, Australia*
[4] *University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland 4102, Australia*
[5] *The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia*

(*Received 13 March 2012; revised 5 June 2012; accepted 7 June 2012*)

Summary

Genome wide association studies (GWAS) have largely succeeded family-based linkage studies in livestock and human populations as the preferred method to map loci for complex or quantitative traits. However, the type of results produced by the two analyses contrast sharply due to differences in linkage disequilibrium (LD) imposed by the design of studies. In this paper, we demonstrate that association and linkage studies are in agreement provided that (i) the effects from both studies are estimated appropriately as random effects, (ii) all markers are fitted simultaneously and (iii) appropriate adjustments are made for the differences in LD between the study designs. We demonstrate with real data that linkage results can be predicted by the sum of association effects. Our association study captured most of the linkage information because we could predict the linkage results with moderate accuracy. We suggest that the ability of common single nucleotide polymorphism (SNP) to capture the genetic variance in a population will depend on the effective population size of the study organism. The results provide further evidence for many loci of small effect underlying complex traits. The analysis suggests a more informed method for GWAS is to fit statistical models where all SNPs are analysed simultaneously and as random effects.

1. Introduction

Genome wide association studies (GWAS) and family-based linkage studies have both been widely used to map genes causing variation in complex or quantitative traits. The two approaches have a similar aim, and so it is surprising that the results from the two methods have been subjected to little systematic comparison, particularly with regard to the size of estimated effects. Both the approaches use genetic markers to discover loci but differ in their experimental design. Linkage analysis relies on segregation of alleles within the family, whereas association analysis simply correlates markers with phenotypes across a population. Some studies compare the methods, but primarily aim to identify influential loci and sometimes only a selected portion of the genome is investigated (McKenzie *et al*., 2001; Daetwyler *et al*., 2008). The equivalence between the estimated effects of loci from the two methods has rarely been explored. When comparisons of several linkage studies are made, results are inconsistent (Altmüller *et al*., 2001); implying either false-positive results, systematic differences, such as different alleles segregating in different families, or lack of statistical power (false-negative results). This paper compares linkage and GWAS and shows that the results are in agreement, provided the differences between the methods are taken into consideration.

A key difference between linkage and association mapping is in the precision with which they map the location of quantitative trait loci (QTLs). A linkage analysis uses recombination events only within the recorded pedigree and so the confidence interval for

* Corresponding author: Kathryn Kemper, Department of Agriculture and Food Systems, University of Melbourne, Parkville, Victoria 3010, Australia. E-mail: kathryn.kemper@dpi.vic.giv.au

the position of the QTL is typically large (Darvasi *et al*., 1993). In contrast, GWAS rely on linkage disequilibrium (LD) between QTLs and markers to detect polymorphisms. As LD extends only for a short distance (i.e. <80 kb in humans Clark *et al*., 2003), the confidence interval for the position of the QTL is generally smaller for a GWAS than for a linkage analysis. Thus, although some GWAS find a QTL in the same region as linkage studies, linkage studies have found QTL on most chromosomes for extensively studied traits and regions identified with linkage tend to extend for long distances (Altmüller *et al*., 2001).

Both GWAS and linkage studies suffer from two deficiencies when carried out using standard procedures. First, the estimated size of effect for significant QTLs are overestimated (e.g. Beavis, 1998; Goring *et al*., 2001; Xu, 2003*b*; Zöllner & Pritchard, 2007; Goddard *et al*., 2009; Sun *et al*., 2011; Xiao & Boehnke, 2011). This arises because a single dataset is used for both discovery and parameter estimation, causing a correlation between the test statistic and the estimated effect size of alleles (Goring *et al*., 2001). Verification of locus effects in an independent population can avoid this bias, provided that the validation results are not conditioned on statistical tests (Goring *et al*., 2001). Alternatively, Goddard *et al*. (2009) argue that this bias can be overcome by fitting the effect of a single nucleotide polymorphism (SNP) or chromosome position as a random effect. If the mean of the posterior distribution of effect size for the estimate is $\hat{b}$, then the expectation of the true effect (b) has the desirable property of being the mean of the estimates, i.e. $E(b|\hat{b})=\hat{b}$ (Goddard *et al*., 2009). This is not the conventional definition of unbiased, but it leads to desirable properties. For instance, if the most significant effects are re-estimated in an independent dataset, then, on average, their effects will not change.

The second problem with both GWAS and linkage analyses as usually practiced is that the effect of one position is estimated ignoring all other positions. In a GWAS, for example, each SNP is tested independently for an association with the trait. Consequently many nearby SNPs may have significant effects because they are all in LD with the same QTL. Alternatively, significant SNP may be near several possible causal polymorphisms (e.g. Barrett *et al*., 2008). This can cause confusion about the number, location and effect size of QTLs that have been detected. One approach to partially overcome this problem in a GWAS is to fit all positions simultaneously as random effects (Meuwissen *et al*., 2001), so that the effect of an SNP is estimated conditional on the effect of all other positions.

Multiple QTLs also cause confusion for results from linkage analyses. The simplest interpretation of a significant peak in the likelihood of a linkage analysis is that there is a single QTL near the peak. However, if more than one QTL contributes to the linkage signal (Haley & Knott, 1992; Martínez & Curnow, 1992), this can lead to a wrong conclusion being drawn and possibly a futile attempt to fine map a single locus (i.e. a so-called 'ghost' QTL). The effect estimated in a linkage analysis is actually the combined effect of all QTLs on the chromosome after accounting for recombination between QTLs and the position being tested. By design, there is strong linkage between adjacent positions in a linkage analysis and, if there are many QTLs, it is impossible to distinguish between adjacent loci because of inadequate recombination. If the effect of all QTLs detected in a GWAS could be combined along a chromosome, allowing for recombination between the position being tested and all other positions, then this effect should be the same as that estimated by a linkage analysis. Yang *et al*. (2010) indicate that common SNP markers capture approximately half of the genetic variance for human height. This could cause a discrepancy between linkage analysis and GWAS as imperfect LD would affect the association analysis but not linkage results. Studies with domesticated species indicate that markers generally capture a higher proportion of the genetic variance (Daetwyler, 2009; Boyko *et al*., 2010; Aitman *et al*., 2011; Haile-Mariam *et al*., 2012), suggesting that this discrepancy should be minimized using a livestock population.

This study tests the hypothesis that effects estimated from a GWAS and from a linkage analysis agree, provided both are estimated appropriately as random effects and that SNPs are fitted simultaneously in both analysis. To test the hypothesis, we needed to conduct a linkage analysis and a GWAS in the same population. We used a sheep population with large half-sib families because this design maximizes power for the linkage analysis and, with appropriate methods, the impact of family structure in the GWAS can be minimized (MacLeod *et al*., 2010). Our approach first demonstrates the consequence of treating the marker effects as random and of fitting all markers simultaneously. Then we show how the effects observed in the linkage analysis can be predicted by combining the effects estimated from the GWAS and allowing for recombination along a chromosome.

2. Materials and methods

(i) *Data*

Genotypes and phenotypes were obtained for 1971 merino sheep from 12 half-sib families from the SheepGenomics project (White *et al*., 2012). The average family size was 164 animals (range: 68–349). Genotypes consisted of 48 640 SNPs from the Illumina Ovine SNP50 BeadChip, which were quality

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

checked and missing genotypes imputed (see Kemper *et al*., 2011). The trait analysed was eye muscle depth (mm) corrected for body weight, measured by ultrasound scanning at approximately 10 months of age. This trait was chosen because many records were available and the trait has an approximate normal distribution. Heritability estimates for eye muscle depth range between 0·22 (±0·04) and 0·33 (±0·03) (Safari *et al*., 2005; Huisman & Brown, 2009; Mortimer *et al*., 2010). Full details of the data collection and procedures can be found in White *et al*. (2012). Genotypes for the 48 640 SNP were available for nine sires, while the genotypes for the remaining three sires were imputed using a rules-based approach from the progeny genotypes and ChromoPhase (Daetwyler *et al*., 2011). Calculations of LD between pairs of markers (r^2) were made using the correlation of genotypes.

(ii) *Assigning inheritance of the paternal alleles*

Alleles for sires and their progeny were phased into paternal and maternal haplotypes using ChromoPhase (Daetwyler *et al*., 2011). Although the sire genotypes were phased, there is no information on which haplotype is paternal or maternal, and so they are referred to below as the first and second chromosome of a sire, where the designation of first and second is arbitrary. The paternal alleles of each offspring were assigned to either the first or second chromosome of their sire based on runs of successive alleles that matched one of the two chromosomes of their sire. The algorithm allowed up to one mismatch per section to account for genotyping and map errors. Unassigned SNPs were treated as missing data. Further details of the algorithm are provided in Part A of the supplementary materials (available at http://journals.cambridge.org/grh).

(iii) *Within-family linkage analysis – fixed effect model*

A fixed effects model was fitted sequentially for all SNP positions. The model was

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zv} + \mathbf{W}\alpha + \mathbf{e}, \quad (1)$$

where **y** is a vector of phenotypes, **X** is a design matrix assigning progeny to fixed effects (including covariates), **b** is a vector of fixed effect solutions, **Z** is a design matrix allocating phenotypes to sires, **v** is a vector of sire solutions, **W** is an incidence matrix assigning progeny to groups according to the allele inherited from their sire, **α** is a vector of effects contrasting each sire's first and second chromosome and **e** is a vector of residuals distributed $N(0, \mathbf{I}\sigma^2_e)$. Fixed effects in **b** were year of birth (2 levels), a regression coefficient for age (in days, mean age 304 days), birth and rearing type (three levels), sex nested within year (four levels) and four regression coefficients for the first four principal components from the genomic relationship matrix (Yang *et al*., 2010). Principal components were fitted as covariates to account for population structure within the maternal haplotypes as maternal pedigree was unknown (Patterson *et al*., 2006). Thus, the estimate of the effect of the sire's allele ($\hat{\alpha}$) is

$$\hat{\alpha} = (W'W)^{-1}W'(y - X\hat{b} - Z\hat{v}), \quad (2)$$

where $\hat{b}$ and $\hat{v}$ are the estimates for the fixed effects and sire solutions. The false discovery rate was calculated as $(1-s)p/[s(1-p)]$ (Storey, 2002; Bolormaa *et al*., 2011), where s and p are the realized and expected proportion of significant SNP.

(iv) *Within-family linkage analysis – random effect model*

The model is similar to the fixed effect analysis (i.e. (1)) except that α is treated as a vector of random effects distributed $\alpha \sim N(0, \mathbf{I}\sigma^2_{\text{sire.snp}})$, where **I** is an identity matrix and $\sigma^2_{\text{sire.snp}}$ is the sire segregation variance. That is, $\sigma^2_{\text{sire.snp}}$ is the variance in the trait attributed to the segregation of alleles within sire families, estimated across all families. To estimate this variance, we took the average variance component estimated using restricted maximum likelihood over all positions with ASReml (Gilmour *et al*., 2006). To avoid an upward bias, imposed by the default settings in ASReml, both positive and negative estimates of $\sigma^2_{\text{sire.snp}}$ were permitted. This variance component was then fixed and used to calculate the allele effect at each position for each sire. The solutions vector, from Henderson's mixed model equations (Henderson, 1950; Mrode, 2005), was

$$\hat{\alpha} = (W'W + \lambda I)^{-1}W'(y - X\hat{b} - Z\hat{v}), \quad (3)$$

where terms are as described in (1), $\lambda = \sigma^2_{\text{error}}/\sigma^2_{\text{sire.snp}}$ and σ^2_{error} is the residual variance. This was computed with ASReml for all positions. An alternative cross-validation method to estimate the sire segregation variance, with respect to the error variance, and therefore the degree of overestimation in the fixed effect model is given in Part B of the supplementary materials (available at http://journals.cambridge.org/grh).

(v) *Association analysis – fixed effect model*

A regression of phenotype on allele dosage was made at each SNP position. That is, the SNP marker effect was fitted as fixed following a conventional linkage analysis. The model was

$$y = \mathbf{X}b + \mathbf{Z}v' + \mathbf{T}\gamma + \mathbf{e}, \quad (4)$$

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

where **X**, **Z** and **e** were as defined in (1), $\mathbf{v}'$ is a vector of random sire effects [distributed $N(0, \mathbf{I}\sigma^2_{\text{sire}})$], **T** is a vector assigning progeny to their SNP genotype (i.e. 0, 1 or 2 copies of an SNP allele) and $\boldsymbol{\gamma}$ is the SNP allele effect (a scalar). The solution for $\hat{\gamma}$ was estimated using ASReml (Gilmour *et al*., 2006), where the sire variance (σ^2_{sire}) was estimated at each position.

(vi) *Association analysis – simultaneous effect of all SNPs with random SNP effects*

Simultaneous estimates of all SNP effects were obtained using the Bayesian approach (BayesA) of Meuwissen *et al*. (2001). The model is

$$\mathbf{y}' = \mathbf{T}\gamma + \mathbf{Z}\boldsymbol{v}' + \mathbf{e} \tag{5}$$

where **T, Z,** $\mathbf{v}'$ and **e** are as defined above (4), $\mathbf{y}'$ is a vector of phenotypes corrected for fixed effects (i.e. $y' = y - X\hat{b}$, as described in (1)) and $\boldsymbol{\gamma}$ is a vector of marker effects assumed to be $N(0, \mathbf{I}\sigma^2_{\gamma i})$, where $\sigma^2_{\gamma i}$ is the variance for the *i*th SNP. This method assumes that allele effects ($\boldsymbol{\gamma}$) come from a *t*-distribution with 4·012 df following Meuwissen *et al*. (2001). This model, in contrast to (4), directly accounts for the LD between nearby markers, the overestimation bias in the marker effects and, by extrapolation of Kang *et al*. (2010) and Yang *et al*. (2011), spurious results due to population stratification. Fitting all SNPs simultaneously indirectly accounts for population stratification because SNP effects are estimated conditional on all other SNPs, whereby eliminating spurious associations (e.g. associations caused by SNP in LD with QTL on different chromosomes). SNP allele effects were estimated as the posterior mean of 10 replicates of a Gibbs chain with 30 000 iterations, with 5000 iterations discarded in each replicate as burn-in.

(vii) *Predicting linkage results from the association analysis*

The estimates of SNP effects from (5) were used to predict the linkage effects at each position. The predicted effect at position *j* for sire *k* ($\eta_{j,k}$) was calculated as

$$\eta_{j,k} = \sum_{i=1}^{M} \hat{\gamma}_i p_{i,j} x_{i,k,1} - \sum_{i=1}^{M} \hat{\gamma}_i p_{i,j} x_{i,k,2}, \tag{6}$$

where $\hat{\gamma}_i$ is the estimate of the SNP allele effect at positions *i*, $p_{i,j}$ is the probability of co-inheritance of positions *i* and *j*, $x_{i,k,1}$ and $x_{i,k,2}$ are sire *k*'s allele at position *i* (i.e. 0 or 1) for the first and second chromosomes and *M* is the total number of SNP positions on the chromosomes. Thus, (6) is the difference between the sum of allele effects for the first and second chromosome at each position, where the sum of allele effects on each chromosome accounts for the probability of recombination events along the chromosome. The probability of co-inheritance of positions *i* and *j* was calculated as $p_{i,j} = 1 - 2c_{i,j}$, where $c_{i,j}$ was the recombination fraction from Haldane's mapping function (Haldane, 1919), i.e. $c_{i,j} = 0{\cdot}5[1 - \exp(-2m)]$ where *m* is the distance (in Morgans) between *i* and *j* and assuming cM = 1 Mbp (Botstein *et al*., 1980 citing Renwick, 1969). The regression coefficient of the observed effect on the predicted linkage effect will be one if (1) the association analysis captures all of the genetic information in the linkage analysis, (2) SNP allele effects are additive and (3) Haldane's mapping function is an accurate predictor of recombination.

(viii) *Predicting linkage results from the association analysis with independent data*

The data from the association analysis used to predict the linkage effects in (5) are not independent of the data used in the linkage analysis. This is because the segregating alleles from the linkage analysis in the 12 sires also contribute to the association analysis. To achieve complete independence between the association and linkage analyses, it was necessary to exclude, in turn, the offspring of each sire from the association analysis. That is, model (5) was run 12 times. SNP marker effects were then used to predict the linkage results using (6) for the sire excluded from the association analysis. For comparison, an analysis that predicts the between sire differences using marker effects estimated with data from all sires and excluding the sire to be predicted (i.e. independent data) is described in Part C of the supplementary materials (available at http://journals.cambridge.org/grh).

3. Results

(i) *Tracking the paternal alleles*

Paternal alleles were assigned to either the 1st or 2nd chromosome of the sire at 92·1 % of positions (range per sire: 81·5–95·8 %), excluding uninformative positions (Supplementary Fig. S1, available at http://journals.cambridge.org/grh). There was an average of 7·2 % unassigned progeny per SNP per sire.

(ii) *Linkage analysis and GWAS using conventional methods*

Using the conventional fixed effect linkage analysis (1), 3109 positions were identified as significant on 15 of 26 chromosomes at a false discovery rate of 14·8 % ($P < 0{\cdot}01$, Fig. 1). When significant SNPs were tested using the genome-wide association analysis (4), there are 132 SNPs identified as significant with a false-discovery rate of 22·8 % ($P < 0{\cdot}01$, SNP details in Supplementary Table S1, available at http://journals.cambridge.org/grh). The false-discovery rate suggests

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

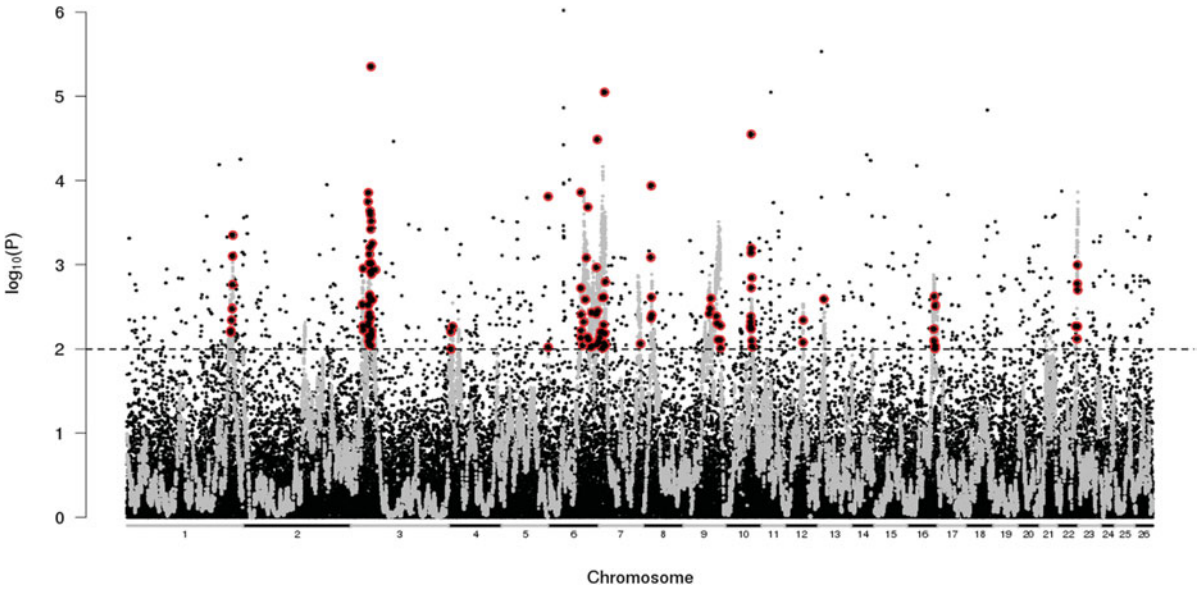

Fig. 1. Comparison of the test statistics across the genome for linkage (grey) and the association (black) analyses. Markers significant in both analyses are highlighted in red ($P<0{\cdot}01$).

many true discoveries, although the closer inspection below creates some confusion for QTLs underlying our trait.

Doubts over the results from the conventional analysis arise because some chromosomes suggest discrete QTL, while for other chromosomes the results are inconsistent. For example, consider chromosomes 3 and 6 (Fig. 2). Chromosome 3 presents seemingly reliable answers where the 43 positions significant in both analyses appear to cluster near two likely QTLs, one at (approx) 30 Mbp and another at 50 Mbp. The effect of the SNP with the highest significance from the association analysis at about 50 Mbp is $-0{\cdot}39$ ($\pm 0{\cdot}08$) mm and the estimated (absolute) effect ranges from 0·01 ($\pm 0{\cdot}27$) to 0·71 ($\pm 0{\cdot}38$) mm for the linkage analysis. However, chromosome 6 shows a strong linkage signal from 80 Mbp onwards and 21 SNP significant from both the linkage and association analysis over a wide region. It is not clear which or if all these SNPs are associated with the linkage peak. The linkage analysis suggests possibly three QTLs, while the SNP also significant in the association analysis suggests maybe four or more QTLs. Also contradictory are the several significant SNPs at about 40 Mbp, which do not have any corresponding linkage signal. It is difficult to ascertain using the two approaches in this form, which analysis is more reliable, which effects are due to experimental noise, how many QTLs exist and what is the best estimate of the position of each QTL.

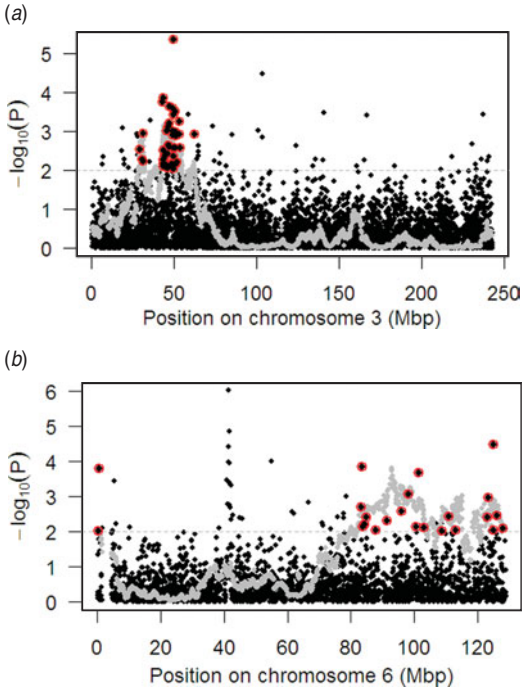

Fig. 2. Comparison of test statistics for chromosomes 3 (*a*) and 6 (*b*) using the linkage (grey) and association (black) analyses. Markers significant in both analyses are highlighted in red ($P<0{\cdot}01$).

(iii) *Linkage analysis – impact of the random effects model*

The mean maximum likelihood estimate for $\sigma^2_{\text{sire.snp}}$ from all positions was 0·013, and thus the average proportion of phenotypic variance explained by the paternally inherited allele was 0·0037 (i.e. $\sigma^2_{\text{sire.snp}}/\sigma^2_{\text{phen}} = 0{\cdot}013/3{\cdot}15$). Although the likelihood failed to converge at 5407 (11·1% of all) positions; a subsequent restricted (positive definite) maximum likelihood analysis at these positions showed an almost zero variance attributed to $\sigma^2_{\text{sire.snp}}$. This method

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

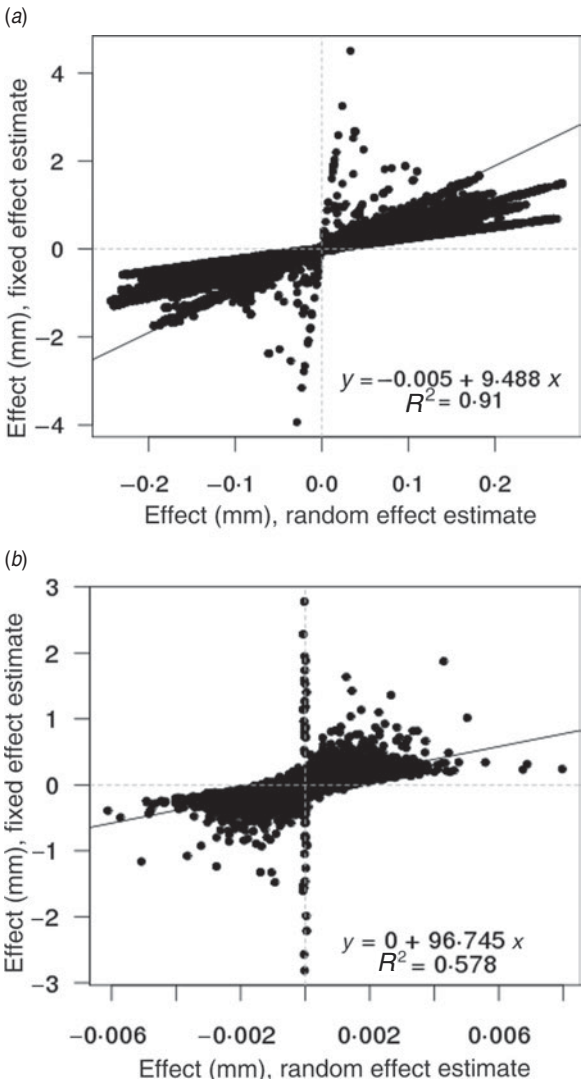

Fig. 3. Effect of fitting SNP alleles as fixed (*y*-axis) or random (*x*-axis) using linkage (*a*) or association (*b*) analysis. Allele effects using linkage are estimated for every sire at all positions (*a*) or across all animals at all positions using association (*b*). Each point represents a single estimate of an allele effect.

overestimates the average proportion of phenotypic variance explained by markers because the sum for all markers is much greater than the genetic variance of the trait (i.e. the expected genetic variance is approximately 0·3 σ^2_{phen} but 0·0037 σ^2_{phen}/SNP 48 640 SNP$>>$0·3 σ^2_{phen}). The overestimation occurs because of the strong LD between makers in the linkage analysis.

Comparison of the fixed and random effects models for SNP alleles from the linkage analysis (i.e. models (2) and (3)) shows broad agreement for most sires at most positions (Fig. 3 *a*). The regression indicates that the random effects analysis explains 91 % of the variation in the fixed effect analysis but that the fixed effect model is estimating the size of the allele effect to be about ten times greater than the random effect model. Adjacent allele effects for a sire are correlated in Fig. 3 *a* (i.e. adjacent SNP positions have correlated effects and form lines in the plot) and this correlation between positions is maintained by the random model. Notably there are several SNP positions with large effects estimated by the fixed model ($>\pm 2$ mm) for which the random model estimates an effect near zero. This severe regression by the random effects model suggests that there was little support for the large effect estimated by the fixed model. These positions are probably regions where poor tracking of the paternal allele occurred, and consequently, there were few progeny who were recorded to inherit each of the sire's alleles.

(iv) *Association study – impact of the random effects model*

The regression of the association allele effects from the fixed and random models (i.e. (4) and (5)) show that the fixed model estimates the effect of alleles almost 100 times larger than the random model (Fig. 3 *b*). Similar to the linkage analysis, many SNP alleles estimated with large effects ($>\pm 1$ mm) from the fixed model were regressed to almost zero using the simultaneous method (Fig. 3 *b*). This occurs because of unreliable estimates of effects from the fixed effect model. For example, of the 23 markers with large effects ($>\pm 1$ mm) from fixed effect model and very small effects ($<$0·001 mm) in the random model, 20 (87 %) were not significant ($P>0{\cdot}05$). The remaining three markers may represent spurious results from the standard GWAS, presumably caused by LD with other SNP.

The regression of the fixed effect solutions on the random effects solutions also explains a lower amount of variation compared with the linkage analysis (i.e. $R^2=0{\cdot}91$ vs. $R^2=0{\cdot}58$, Fig. 3). The differences between the models and the lower proportion of variance explained by the random effect model is partially due to overestimation of the effects when they are fitted one at a time as fixed effects and partially because the BayesA analysis may spread the effect of each QTL over several adjacent SNP. For example, Fig. 4 compares the fixed and BayesA analysis over a 20 Mbp region on chromosome 6 where there appears to be a strong QTL signal at around 42 Mbp. The random effects analysis maps this effect in a location slightly further along the chromosome (41·5 Mbp) compared with the fixed effect analysis (40·8 Mbp), but it also shows the spread of QTL effects for SNP in modest LD ($r^2>0{\cdot}5$) with this SNP in the region. Further, from the random effects model, it is clearer that there are possibility of three QTLs at 30·7, 45·0 and 50·6 Mbp for markers which are not in strong LD with the SNP at

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

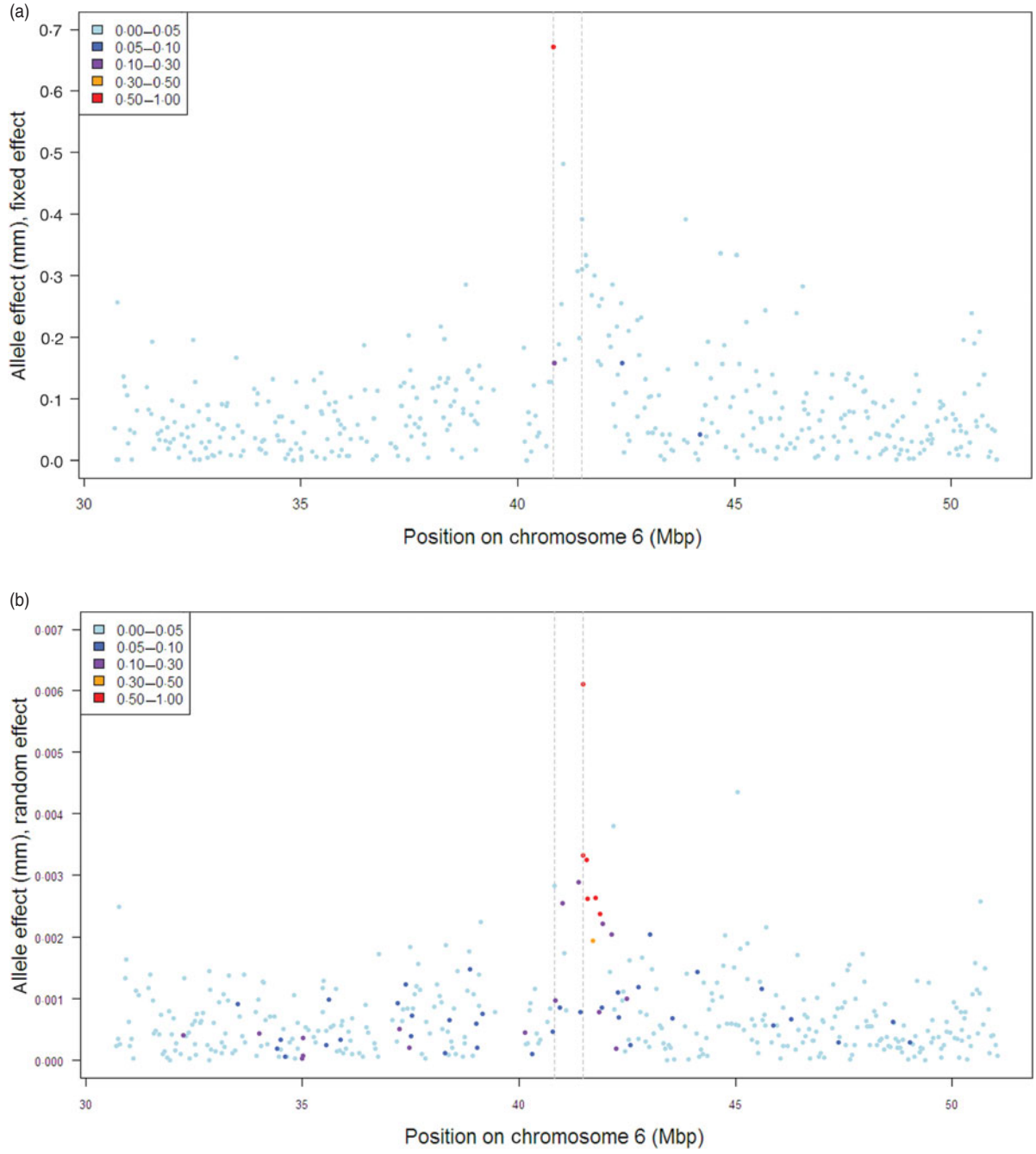

Fig. 4. The absolute effect of SNP alleles when fitted as fixed (*a*) or random (*b*) in the association analysis. Grey lines indicate the positions of the largest effect in (*a*, 40·8 Mbp) or (*b*, 41·5 Mbp) with colours showing the LD (correlation) between these marked SNPs and the surrounding markers.

41·5 Mbp. A further SNP at 42·1 Mbp may be associated with the same QTL tracked by the SNP at 41·5 Mbp or this association could indicate a fourth additional QTL.

(v) *Predicting the linkage results from the association study*

Despite the correction for bias in the linkage and association analyses the magnitude of the association effects are still in the order of 100 times smaller than those estimated from the linkage analysis (Fig. 3). A prediction of the linkage results from the association analysis needs to account for the stronger LD between adjacent positions in the linkage analysis. Using the linkage results from random model (i.e. (3)), the prediction was the contrast between sire chromosomes for the sum of the association effects accounting for recombination (i.e. models (5) and (6)). For individual sires, the expectation of the linkage effects shows

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

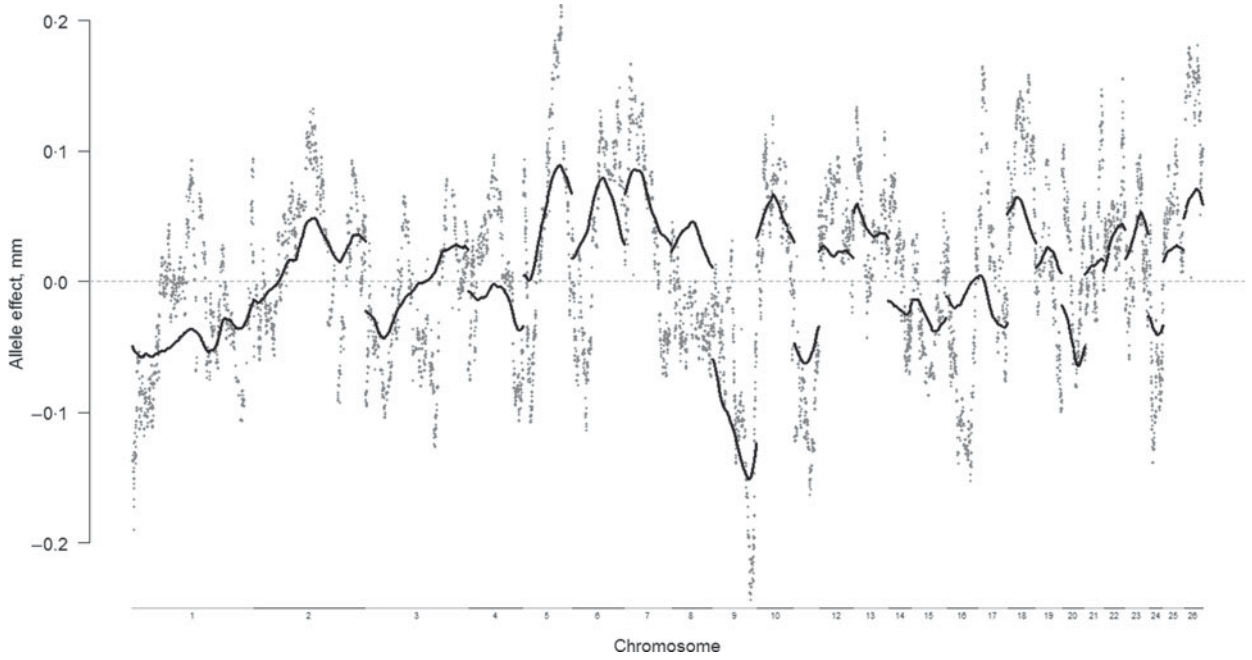

Fig. 5. The size of marker effects (mm) across the genome for a single sire ('W4') when alleles are fitted as random using linkage (grey) or predicted using the sum of association effects accounting for recombination (black).

good agreement with the linkage results (Fig. 5, Supplementary Fig. S2, available at http://journals.cambridge.org/grh). To compare the effects across all sires and at all positions we plotted the estimate from the linkage analysis against that predicted from the association study (Fig. 6*a*). The regression is almost one (slope: $0{\cdot}975 \pm 1{\cdot}2 \times 10^{-3}$, intercept: $3{\cdot}7 \times 10^{-3} \pm 6{\cdot}9 \times 10^{-5}$) and accounts for about half of the variation in the linkage results ($R^2 = 0{\cdot}523$). Considering the sampling errors in both estimates, this suggests that the association analysis is capturing the majority of within-family information. There were no data points which showed a notable deviation from the regression slope (Supplementary Fig. S3).

(vi) *Predicting the linkage results with independent data*

There was a high correlation between the SNP effects estimated with all animals and those estimated excluding progeny from each sire using the random effects model (average $R^2 = 0{\cdot}91$, range: 0·85–0·93). However, these analyses predicted the linkage effects for the excluded sire very inaccurately (Fig. 6*b*, $R^2 = 0{\cdot}002$). This contrasts sharply to results when the sire to be predicted is included in the analysis (Fig. 6*a*). Thus, the sire whose linkage analysis is to be predicted must be included in the association analysis to achieve good agreement between the two approaches. Predictive ability with independent data is slightly improved when predicting between sires differences ($R^2 = 0{\cdot}04$, Part C of the supplementary material, available at http://journals.cambridge.org/grh).

4. Discussion

This study suggests two reasons why there is often little agreement between linkage analysis and GWAS on the same complex trait. First, when the effects are estimated as fixed effects in statistical models, the most significant effects are often grossly overestimated. This is evident in our study for both the linkage and association analysis. Overestimation of fixed effects has been highlighted previously by several authors (e.g. Beavis, 1998) and contributes to the often smaller than expected or perhaps non-significant results for loci when replication is attempted. Naturally, this problem also occurs if one attempts to verify the results of a linkage analysis with a GWAS or vice versa. Our GWAS predicted the linkage results, provided both are estimated as random effects, SNPs are fitted simultaneously in the GWAS, and GWAS effects on a chromosome are combined to account for LD in the linkage analysis. The regression of the observed linkage effect on the effect predicted from the GWAS is close to 1·0 indicating an approximate agreement in size. The proportion of the variance in the linkage results explained by our prediction is high ($R^2 = 0{\cdot}52$) considering that both sets of effects are estimated with error.

Second, this study shows that multiple linked QTLs can be the underlying cause of significant linkage results. In contrast to the simulation studies with multiple QTLs tracked by microsatellite markers

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

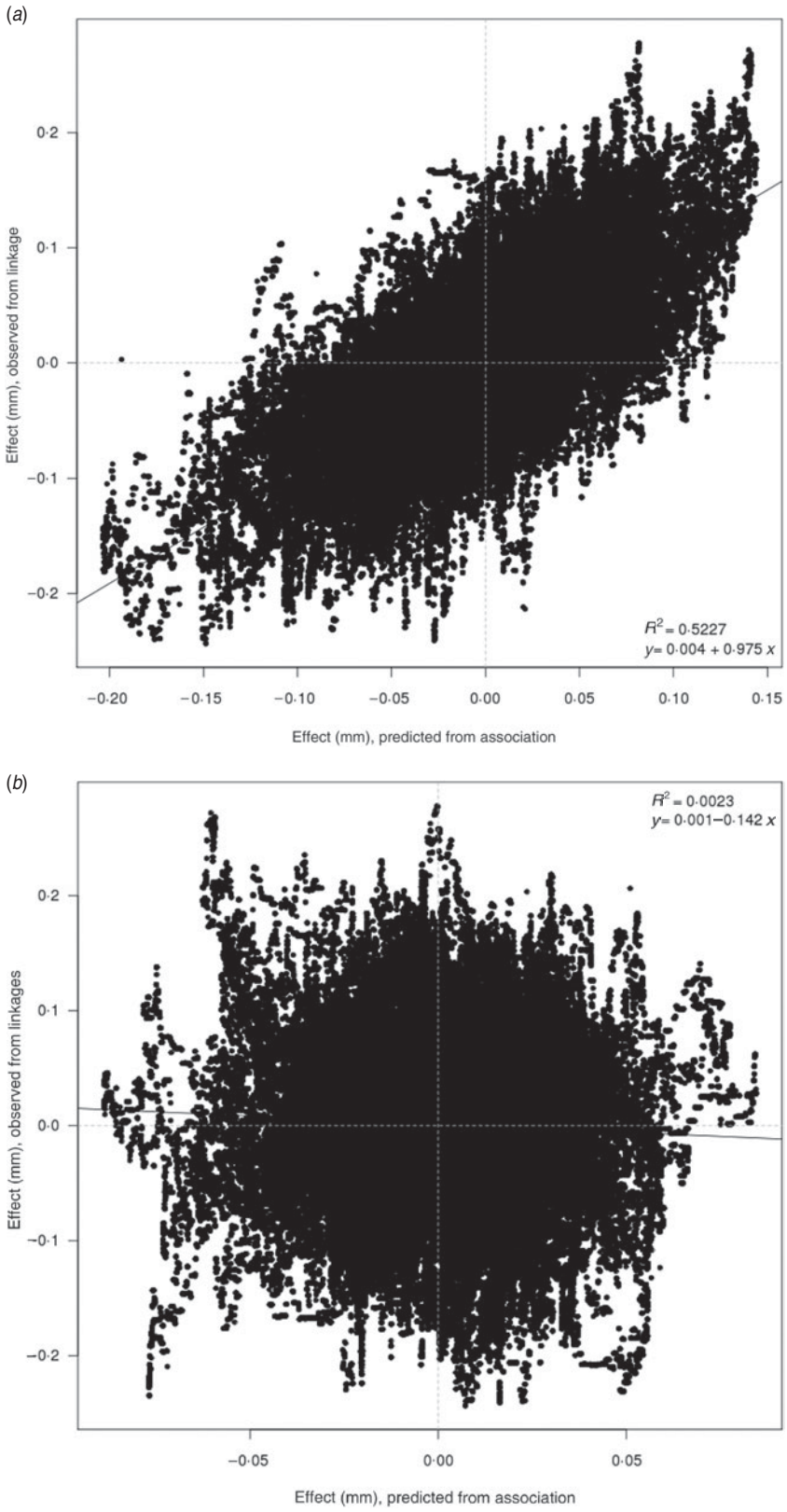

Fig. 6. Marker effects (mm) estimated from linkage when alleles are fitted as random (*y*-axis) or predicted from the sum of the association effects accounting for recombination (*x*-axis). The association analysis either includes all sires (*a*) or excludes the sire to be predicted (*b*).

https://doi.org/10.1017/S0016672312000365 Published online by Cambridge University Press

(e.g. Haley & Knott, 1992), our results in real data suggest that likelihood peaks can be caused by the sum of many QTLs along a chromosome. We do not suggest that all linkage peaks are detecting multiple small QTLs because some studies have been successful in identifying important loci (e.g. Gusella *et al.*, 1983; Tsui *et al.*, 1985; Charlier *et al.*, 1995; Coppieters *et al.*, 1998). However, successful linkage studies involve polymorphisms of large effect and these loci probably overwhelm any interference in the signal caused by multiple linked loci. The effect of the linked loci could be to increase or decrease the apparent effect size of the major loci, depending on the phase of the interacting loci. Here, we demonstrate with real data that the additive effect of multiple loci in strong LD can cause apparent linkage signals. This conclusion is consistent with simulation and theoretical studies (e.g. Dekkers & Dentine, 1991; Visscher & Haley, 1996) and is also supported by mice studies when single QTL fractionate into multiple smaller loci with fine mapping (Flint *et al.*, 2005).

The influence of nearby linked loci cannot be excluded when using association rather than linkage analysis. Even in a conventional GWAS analysis, fitting one SNP at a time, SNP with significant effects may be influenced by multiple nearby QTLs, some in phase and some out of phase with the tested SNP. However, LD in GWAS probably has less influence than in linkage because LD usually extends for shorter distances, i.e. <1 Mbp in Merino sheep (Kemper *et al.*, 2011). Hence, a large number of significant SNPs most likely indicate a large number of QTLs. This conclusion is made clearer by fitting all SNPs simultaneously. Then SNPs that have no marginal effect after fitting all other SNPs, including SNP in strong LD with the causal polymorphisms, will show no association with the trait. Figure 4b shows a typical result where there are several positions along the chromosome associated with the trait of interest.

The high degree of agreement ($R^2 = 0{\cdot}52$, regression coefficient $\sim 1{\cdot}0$) between our observed and predicted linkage results is surprising. This consistency suggests that the association analysis is tracking the majority of the linkage information and that imperfect LD (between causal mutations and SNP) is not a strong influence on the results from our association analysis. This is because the linkage analysis has strong LD within families and imperfect LD is not limiting as it can be in GWAS. Incomplete LD between common SNP and causative mutations has been hypothesized to be responsible for $\sim 50\,\%$ of the genetic variation in human populations which is not explained by common SNP (Yang *et al.*, 2010). Here, we suggest that the importance of incomplete LD between SNP and causative mutations is influenced strongly by genetic diversity. Our observation is supported by other studies with domestic species where common SNP capture a high proportion of the genetic variance (e.g. Daetwyler, 2009; Boyko *et al.*, 2010; Haile-Mariam *et al.*, 2012). Thus, as the population's diversity, or effective population size (N_e), increases the ability of common SNP to capture the genetic variance reduces. Incomplete LD may occur when causative SNPs are at a lower frequency than the genotyped SNPs (Yang *et al.*, 2010), suggesting an increased importance for these mutations in, for example, human compared with livestock populations.

Extensive QTL mapping experiments in many species suggests that alleles with a large effect on quantitative traits are uncommon (e.g. Darvasi & Pisanté-Shalom, 2002). The results of the association analysis reported here suggest many QTLs for our trait but we found no evidence of large effect QTL in our sires. For instance, if most important genes had a variant with large effect, we might expect to see at least one sire with a large estimated effect from the linkage analysis and an inaccurate prediction of this effect from the GWAS. However, we never observed any allele from the linkage analysis which substantially differed from the effect predicted from the association analysis (Fig. 6). We sampled only 12 sires but we analysed each sire at thousands of positions. If most of the genetic variance was due to rare large effect variants then we might expect to observe at least one heterozygous sire in our dataset. The situation of segregating alleles with large effect may occur but it cannot be typical because we predicted our linkage results from an association analysis with moderate accuracy. Further, all of our estimated effects from the association analysis were also very small ($<0{\cdot}008$ mm or $<0{\cdot}008/3{\cdot}15^{1/2} = 0{\cdot}004$ s.d.).

Our results show that most of the linkage information was captured in the prediction from the GWAS results. However, the two approaches are not independent because they use the same data and we also show that when the sire to be predicted is excluded from the association analysis we cannot predict the linkage results. This discrepancy could be explained by high sampling covariance between the effects estimated for SNP in very strong LD with one another. Thus, the combination of SNP alleles has been observed in the data to be predicted accurately. The between sire differences, which are the sum of all SNP effects, were estimated more precisely using independent data. Prediction of between sire differences is equivalent to genomic prediction which, given larger datasets, can reach moderate accuracies in sheep for this trait (Daetwyler *et al.*, 2010). The dependency between SNP when estimating effects of individual markers is not surprising considering that the magnitude of the largest effect was very small (0·004 s.d.) and given the relatively small size of the dataset.

These results suggest that the best analysis is the GWAS in which all SNPs are fitted simultaneously. This method gave us consistent results between linkage and association, and has greater power to discriminate linked QTLs than either the linkage analysis or the standard GWAS fitting one SNP at a time. This is clearly demonstrated in Fig. 4, where numerous GWAS results are consolidated into possibly four QTL signals at 30·7, 41·5, 45·0 and 50·6 Mbp. A potential drawback of this method is that effects may be split between closely linked markers (Xu, 2003*a*). In Fig. 4, this is potentially occurring for several markers in high LD with the largest estimated effect at 41·5 Mbp. These high LD markers may also be capturing multiple different mutations at the locus. However, the effect of this disadvantage should diminish as markers in higher LD with the causal mutations for traits are included in the SNP marker set.

In summary, this study aimed to reconcile some of the differences between linkage and linkage-disequilibrium mapping. We have demonstrated, using real data, the correction for the biases in both linkage and association mapping. We show that multiple linked QTLs can combine to be the primary cause of significant linkage results. In our study, the association analysis captured 52% of the within-family information, which is high considering the sampling error of effects from both analyses. The results support the hypothesis that there are many loci of small effect underlying complex traits. We suggest an improved method for GWAS is to fit statistical models where all SNPs are analysed simultaneously. This method prevents spurious results caused by population structure and accounts for LD surrounding the analysed SNP.

We thank the SheepGenomics and CRC for Sheep Industry Innovation for the provision of the data for this work. This research was supported under Australian Research Council's *Discovery Projects* funding scheme (project DP1093502).

5. Declaration of interest

None.

6. Supplementary material

The online data are available at http://journals.cambridge.org/GRH

References

Aitman, T. J., Boone, C., Churchill, G. A., Hengartner, M. O., Mackay, T. F. C. & Stemple, D. L. (2011). The future of model organisms in human disease research. *Nature Review Genetics* **12**, 575–582.

Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics* **69**, 936–950.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.-P., de Vos, M., Vermeire, S., Louis, E., Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M. & Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40**, 955–962.

Beavis, W. D. (1998). QTL analysis: power, precision and accuracy. *Molecular Dissection of Complex Traits*, pp. 145–161. Boca Raton, FL: CRC Press.

Bolormaa, S., Hayes, B. J., Savin, K., Hawken, R., Barendse, W., Arthur, P. F., Herd, R. M. & Goddard, M. E. (2011). Genome-wide association studies for feedlot and growth traits in cattle. *Journal of Animal Science* **89**, 1684–1697.

Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980). Construction of a genetic linkage map in man using restricted fragment length polymorphisms. *American Journal of Human Genetics* **32**, 314–331.

Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., Lohmueller, K. E., Zhao, K., Brisbin, A., Parker, H. G., von Holdt, B. M., Cargill, M., Auton, A., Reynolds, A., Elkahloun, A. G., Castelhano, M., Mosher, D. S., Sutter, N. B., Johnson, G. S., Novembre, J., Hubisz, M. J., Siepel, A., Wayne, R. K., Bustamante, C. D. & Ostrander, E. A. (2010). A simple genetic architecture underlies morphological variation in dogs. *PLoS Biology* **8**, e1000451.

Charlier, C., Coppieters, W., Farnir, F., Grobet, L., Leroy, P. L., Michaux, C., Mni, M., Schwers, A., vanmanshoven, P., Hanset, R. & Georges, M. (1995). The *mh* gene causing double-muscling in cattle maps to bovine Chromosome 2. *Mammalian Genome* **6**, 788–792.

Clark, A. G., Nielsen, R., Signorovitch, J., Matise, T. C., Glanowski, S., Heil, J., Winn-Deen, E. S., Holden, A. L. & Lai, E. (2003). Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *The American Journal of Human Genetics* **73**, 285–300.

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J. J., Grisart, B., Mackinnon, M. & Georges, M. (1998). A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design. *Genetics* **149**, 1547–1555.

Daetwyler, H. D. (2009). *Genome-Wide Evaluation of Populations*. Wageningen University, Wageningen, NL.

Daetwyler, H. D., Hickey, J., Henshall, J., Dominik, S., Gredler, B., van der Werf, J. H. J. & Hayes, B. J. (2010). Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Animal Production Science* **50**, 1004–1010.

Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M. & Robinson, J. A. B. (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. *Journal of Dairy Science* **91**, 3225–3236.

Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A. & Goddard, M. E. (2011). Imputation of genotypes from sparse to high density using long-range phasing. *Genetics* **189**, 317–327.

Darvasi, A. & Pisanté-Shalom, A. (2002). Complexities in the genetic dissection of quantitative trait loci. *Trends in Genetics* **18**, 489–491.

Darvasi, A., Weinreb, A., Minke, V., Weller, J. I. & Soller, M. (1993). Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134**, 943–951.

Dekkers, J. C. M. & Dentine, M. R. (1991). Quantitative genetic variance associated with chromosomal markers in segregating populations. *TAG Theoretical and Applied Genetics* **81**, 212–220.

Flint, J., Valdar, W., Shifman, S. & Mott, R. (2005). Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Review Genetics* **6**, 271–286.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. & Thompson, R. (2006). *ASReml User Guide 2·0*. Hemel Hempsted, UK: VSN International Ltd.

Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. *Statistical Science* **24**, 514–529.

Goring, H. H. H., Terwilliger, J. D. & Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics* **69**, 1357–1369.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. & Martin, J. B. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238.

Haile-Mariam, M., Nieuwhof, G. J., Beard, K. T., Konstatinov, K. V. & Hayes, B. J. (2012). Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. *Journal of Animal Breeding and Genetics*. doi: 10.1111/j.1439-0388.2012.01001.x.

Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linkaged factors. *Journal of Genetics* **8**, 299–309.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Henderson, C. R. (1950). Estimation of genetic parameters. *Annuals of Mathematical Statistics* **21**, 309.

Huisman, A. E. & Brown, D. J. (2009). Genetic parameters for bodyweight, wool, and disease resistance and reproduction traits in Merino sheep. 3. Genetic relationships between ultrasound scan traits and other traits. *Animal Production Science* **49**, 283–288.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C. & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354.

Kemper, K. E., Emery, D. L., Bishop, S. C., Oddy, H., Hayes, B. J., Dominik, S., Henshall, J. M. & Goddard, M. E. (2011). The distribution of SNP marker effects for faecal worm egg count in sheep, and the feasibility of using these markers to predict genetic merit for resistance to worm infections. *Genetics Research* **93**, 203–219.

MacLeod, I. M., Hayes, B. J., Savin, K. W., Chamberlain, A. J., McPartlan, H. C. & Goddard, M. E. (2010). Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. *Journal of Animal Breeding and Genetics* **127**, 133–42.

Martínez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *TAG Theoretical and Applied Genetics* **85**, 480–488.

McKenzie, C. A., Abecasis, G. R., Keavney, B., Forrester, T., Ratcliffe, P. J., Julier, C., Connell, J. M. C., Bennett, F., McFarlane-Anderson, N., Lathrop, G. M. & Cardon, L. R. (2001). Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Human Molecular Genetics* **10**, 1077–1084.

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.

Mortimer, S. I., van der Werf, J. H. J., Jacob, R. H., Pethick, D. W., Pearce, K. L., Warner, R. D., Geesink, G. H., Hocking Edwards, J. E., Gardner, G. E., Ponnampalam, E. N., Kitessa, S. M., Ball, A. J. & Hopkins, D. L. (2010). Preliminary estimates of genetic parameters for carcass and meat quality traits in Australian sheep. *Animal Production Science* **50**, 1135–1144.

Mrode, R. A. (2005). *Linear Models for the Prediction of Animal Breeding Values*. Wallingford: CABI Publishing.

Patterson, N., Price, A. L. & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2**, e190.

Renwick, J. H. (1969). Progress in mapping human autosomes. *British Medical Bulletin* **25**, 65–73.

Safari, E., Fogarty, N. M. & Gilmore, A. R. (2005). A review of genetic parameters estimates for wool, growth, meat and reproduction traits in sheep. *Livestock Production Science* **92**, 271–289.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) **64**, 479–498.

Sun, L., Dimitromanolakis, A., Faye, L. L., Paterson, A. D., Waggott, D. & Bull, S. B. (2011). BR-squared: a practical solution to the winner's curse in genome-wide scans. *Human Genetics* **129**, 545–552.

Tsui, L., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, J., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., Zsiga, M., Markiewicz, D., Akots, G., Brown, V., Helms, C., Gravius, T., Parker, C., Rediker, K. & Donis-Keller, H. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **230**, 1054–1057.

Visscher, P. & Haley, C. (1996). Detection of putative quantitative trait loci in line crosses under infinitesimal genetic models. *TAG Theoretical and Applied Genetics* **93**, 691–702.

White, J. D., Allingham, P. G., Gorman, C. M., Emery, D., Hynd, P., Owens, J., Bell, A., Siddell, J., Hayes, B., Usmar, J., Goddard, M., Henshall, J., Dominik, S., Brewer, H., van der Werf, J., Nicholar, F. W., Warner, R., Hofmyer, C., Longhurst, T., Swan, P., Forage, R. & Oddy, V. H. (2012). Design and phenotyping proceedures for recording wool, skin, parasite resistance, growth, carcass yield and quality traits of the SheepGENOMICS mapping flock. *Animal Production Science* **52**, 157–171.

Xiao, R. & Boehnke, M. (2011). Quantifying and correcting for the winner's curse in quantitative-trait association studies. *Genetic Epidemiology* **35**, 133–138.

Xu, S. (2003*a*). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

Xu, S. (2003*b*). Theoretical basis of the Beavis effect. *Genetics* **165**, 2259–2268.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., Hill, W. G., Landi, M. T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R. A., Melbye, M., Pugh, E., Cornelis, M. C., Weir, B. S., Goddard, M. E. & Visscher, P. M. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**, 519–525.

Zöllner, S. & Pritchard, J. K. (2007). Overcoming the winner s curse: estimating penetrance parameters from case-control data. *American Journal of Human Genetics* **80**, 605–615.