View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Queensland eSpace

Kemper KE, Daetwyler HD, **Visscher PM** & Goddard ME (2012) Comparing linkage and association analyses in sheep points to a better way of doing GWAS. **Genetics Research** 94: 191-203.
The final publication is available at http://dx.doi.org/10.1017/S0016672312000365

1 ***RESEARCH PAPER***

2 Comparing linkage and association analyses in sheep (*Ovis aries*) points to a
3 better way of doing GWAS

4

5 Kathryn E. Kemper[1*], Hans D. Daetwyler[2,3], Peter M. Visscher[4,5], Michael E.
6 Goddard[1,3]

7

8 [1]Department of Agriculture and Food systems, University of Melbourne,
9 Parkville, Victoria 3010, Australia
10 [2]Victorian Department of Primary Industries, AgriBiosciences Centre, LaTrobe
11 Research and Development Park, Bundoora, Victoria 3083, Australia
12 [3]Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW,
13 2351, Australia
14 [3]Victorian Department of Primary Industries, AgriBiosciences Centre, LaTrobe
15 Research and Development Park, Bundoora, Victoria 3083, Australia
16 [4]University of Queensland Diamantina Institute, University of Queensland,
17 Princess Alexandra Hospital, Brisbane, Queensland 4102, Australia
18 [5]The Queensland Brain Institute, The University of Queensland, Brisbane,
19 Queensland 4072, Australia

20

21 *corresponding author: Kathryn Kemper
22 Email: kathryn.kemper@dpi.vic.giv.au

23

24 *Running heading:*
25 Comparing linkage and association mapping

26 ## Summary

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

45

46 ## Keywords

47 mutation effect size; linkage disequilibrium; rare variants; common variants

48

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. The approaches both use genetics markers to discovery loci but differ in their experimental design. Linkage analysis relies on within family segregation of alleles while 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 (Daetwyler *et al.*, 2008; McKenzie *et al.*, 2001). Rarely has the equivalence between the estimated effects of loci from the two methods been explored. When comparisons of several linkage studies are made, result 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-negatives). This paper compares linkage and association analysis genome-wide and shows that the results are in agreement provided the differences between the methods are taken into consideration.

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

(Darvasi *et al.*, 1993). In contrast, GWAS rely on linkage disequilibrium (LD) between QTL and markers to detect polymorphisms. Since LD extends for only a short distance (i.e. < 80kb 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 QTL are overestimated (e.g. Beavis, 1998; Goddard *et al.*, 2009; Goring *et al.*, 2001; Sun *et al.*, 2011; Xiao & Boehnke, 2011; Xu, 2003b; Zöllner & Pritchard, 2007). 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 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. $E(b \mid \hat{b}) = \hat{b}$ (Goddard *et al.*, 2009). This is not the conventional definition of unbiased but it leads to desirable properties. For

98 instance, if the most significant effects are re-estimated in an independent
99 dataset, then, on average, their effects will not change.
100
101 The second problem with both GWAS and linkage analyses as usually
102 practiced, is that the effect of one position is estimated ignoring all other
103 positions. In a GWAS, for example, each SNP is tested independently for an
104 association with the trait. Consequently many nearby SNP may have
105 significant effects because they are all in LD with the same QTL. Alternatively,
106 significant SNP may be near several possible causal polymorphisms (e.g.
107 Barrett *et al.*, 2008). This can cause confusion about the number, location and
108 effect size of QTL that have been detected. One approach to partially
109 overcome this problem in a GWAS is to fit all positions simultaneously as
110 random effects (Meuwissen *et al.*, 2001), so that the effect of a single SNP is
111 estimated conditional on the effect of all other positions.
112
113 Multiple QTL also cause confusion for results from linkage analyses. The
114 simplest interpretation of a significant peak in the likelihood of a QTL is that
115 there is a single QTL near the peak. However, if more than one QTL
116 contributes to the linkage signal (Haley & Knott, 1992; Martínez & Curnow,
117 1992), this can lead to the wrong conclusion being drawn and possibly a futile
118 attempt to fine map the single QTL (i.e. a so called 'ghost' QTL). The effect
119 estimated in a linkage analysis is actually the combined effect of all the QTL
120 on the chromosome after accounting for recombination between QTL and the
121 position being tested. By design, there is strong linkage between adjacent
122 positions in a linkage analysis and, if there are many QTL, it is impossible to

123 distinguish between adjacent loci because of inadequate recombination. If the
124 effect of all QTL detected in a GWAS could be combined along a
125 chromosome, allowing for recombination between the position being tested
126 and all other positions, then this effect should be the same as that estimated
127 by a linkage analysis. Yang *et al.* (2010) indicates that common single
128 nucleotide polymorphism (SNP) markers capture approximately ½ of the
129 genetic variance for humans height. This could cause a discrepancy between
130 linkage analysis and GWAS as imperfect LD would affect association but not
131 linkage results. Studies with domesticated species indicate that markers
132 generally capture a higher proportion of the genetic variance (Aitman *et al.*,
133 2011; Boyko *et al.*, 2010; Daetwyler, 2009; Haile-Mariam *et al.*, 2012)
134 suggesting that this discrepancy should be minimised using a livestock
135 population.
136
137 This study tests the hypothesis that effects estimated from a GWAS and from
138 a linkage analysis agree provided both are estimated appropriately as random
139 effects and that SNP are fitted simultaneously in both analysis. To test the
140 hypothesis we needed to conduct a linkage analysis and a GWAS in the same
141 population. We used a sheep population with large half-sib families because
142 this design maximises power for the linkage analysis and, with appropriate
143 methods, the impact of family structure in GWAS can be minimised (MacLeod
144 *et al.*, 2010). Our approach first demonstrates the consequence of treating the
145 marker effects as random and of fitting all SNP simultaneously. Then we show
146 how the effects observed in the linkage analysis can be predicted by

147 combining the estimated effects from the GWAS and allowing for
148 recombination along a chromosome.
149

150 2. Materials and Methods

151 *Data.* Genotypes and phenotypes were obtained for 1971 merino sheep from
152 12 half-sib families from the SheepGenomics project (White *et al.*, 2012). The
153 average family size was 164 animals (range: 68 to 349). Genotypes consisted
154 of 48,640 SNP from the Illumina Ovine SNP50 BeadChip which were quality
155 checked and missing genotypes imputed (see Kemper *et al.*, 2011). The trait
156 analysed was eye muscle depth (mm) corrected for body weight, measured by
157 ultrasound scanning at approximately 10 months of age. This trait was chosen
158 because many records were available and the trait has an approximate
159 normal distribution. Heritability estimates for eye muscle depth range between
160 0.22 (±0.04) and 0.33 (±0.03) (Huisman & Brown, 2009; Mortimer *et al.*, 2010;
161 Safari *et al.*, 2005). Full details of the data collection and procedures can be
162 found in White *et al.* (2012). Genotypes for the 48,640 SNP were available for
163 9 sires while the genotypes for the remaining 3 sires were imputed using a
164 rules based approach from the progeny genotypes and ChromoPhase
165 (Daetwyler *et al.*, 2011). Calculations of LD between pairs of markers (r^2) were
166 made using the correlation of genotypes.
167

168 *Assigning inheritance of the paternal alleles.* Alleles for sires and their
169 progeny were phased into paternal and maternal haplotypes using
170 ChromoPhase (Daetwyler *et al.*, 2011). Although the sire genotypes were
171 phased there is no information on which haplotype is paternal or maternal and

172 so they are referred to below as the first and second chromosome of a sire
173 where the designation of first and second is arbitrary. The paternal alleles of
174 each offspring were assigned to either the first or second chromosome of their
175 sire based on runs of successive alleles that matched one of the two
176 chromosomes of their sire. The algorithm allowed up to one mismatch per
177 section to account for genotyping and map errors. Unassigned SNP were
178 treated as missing data. Further details of the algorithm are provided in Part A
179 of the supplementary materials.

180

181 *Within-family linkage analysis – fixed effect model.* A fixed effects model was
182 fitted sequentially for all SNP positions. The model was:

183 $$\mathbf{y} = \mathbf{Xb} + \mathbf{Zv} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{e} \quad [1]$$

184 where **y** is a vector of phenotypes, **X** is a design matrix assigning progeny to
185 fixed effects (including covariates), **b** is a vector of fixed effect solutions, **Z** is a
186 design matrix allocating phenotypes to sires, **v** is a vector of sire solutions, **W**
187 is an incidence matrix assigning progeny to groups according to the allele
188 inherited from their sire, $\boldsymbol{\alpha}$ is a vector of effects contrasting each sire's first and
189 second chromosome and **e** is a vector of residuals distributed $N(0,\mathbf{I}\sigma^2_e)$. Fixed
190 effects in **b** were year of birth (2 levels), a regression coefficient for age (in
191 days, mean age 304 days), birth and rearing type (3 levels), sex nested within
192 year (4 levels) and 4 regression coefficients for the first 4 principal
193 components from the genomic relationship matrix (Yang *et al.*, 2010).
194 Principal components were fitted as covariates to account for population
195 structure within the maternal haplotypes as maternal pedigree was unknown

196 (Patterson *et al.*, 2006). Thus the estimate of the effect of the sire's allele ($\hat{\alpha}$)
197 is:

198 $$\hat{\alpha} = (W'W)^{-1}W'(y - X\hat{b} - Z\hat{v}) \quad [2]$$

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

203

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

216 $$\hat{\alpha} = (W'W + \lambda I)^{-1}W'(y - X\hat{b} - Z\hat{v}) \quad [3]$$

217 where terms are as described in [1], $\lambda = \sigma^2_{error} / \sigma^2_{sire.snp}$ and σ^2_{error} is the residual
218 variance. This was computed with ASReml for all positions. An alternative
219 cross-validation method to estimate the sire segregation variance, with

220 respect to the error variance, and therefore the degree of overestimation in
221 the fixed effect model is given in Part B of the supplementary materials.
222
223 *Genome wide association analysis – fixed effect model.* A regression of
224 phenotype on allele dosage was made at each SNP position. That is, the SNP
225 marker effect was fitted as fixed following a conventional linkage analysis. The
226 model was

227 $$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{v}' + \mathbf{T}\gamma + \mathbf{e} \quad [4]$$

228 where **X, Z** and **e** were as defined for [1], **v'** is a vector of random sire effects
229 [distributed N(0, $\mathbf{I}\sigma^2_{sire}$)], **T** is a vector assigning progeny to their SNP
230 genotype (i.e. 0, 1 or 2 copies of a SNP allele) and γ is the SNP allele effect (a
231 scalar). The solution for $\hat{\gamma}$ was estimated using ASReml (Gilmour *et al.*, 2006)
232 where the sire variance (σ^2_{sire}) was estimated at each position.
233
234 *Genome wide association analysis – simultaneous effect of all SNP with*
235 *random SNP effects.* Simultaneous estimates of all SNP effects were obtained
236 using the Bayesian approach (BayesA) of Meuwissen *et al.* (2001). The model
237 is

238 $$\mathbf{y}' = \mathbf{T}\gamma + \mathbf{Z}\mathbf{v}' + \mathbf{e} \quad [5]$$

239 where **T, Z, v'** and **e** are as defined above [4], **y'** is a vector of phenotypes
240 corrected for fixed effects (i.e. $y' = y - X\hat{b}$, as described in [1]) and γ is a vector
241 of marker effects assumed to be N(0, $\mathbf{I}\sigma^2_{\gamma i}$), where $\sigma^2_{\gamma i}$ is the variance for the
242 i^{th} SNP. This method assumes that allele effects (γ) come from a t-distribution
243 with 4.012 df following Meuwissen *et al.* (2001). This model, in contrast to [4],
244 directly accounts for the LD between nearby markers, the overestimation bias

245 in the marker effects and, by extrapolation of Kang *et al.* (2010) and Yang *et*
246 *al.* (2011), spurious results due to population stratification. Fitting all SNP
247 simultaneously indirectly accounts for population stratification because SNP
248 effects are estimated conditional on all other SNP, whereby eliminating
249 spurious associations (e.g. associations caused by SNP in LD with QTL on
250 different chromosomes). SNP allele effects were estimated as the posterior
251 mean of 10 replicates of a Gibbs chain with 30,000 iterations, with 5,000
252 iterations discarded in each replicate as burn-in.
253
254 *Predicting linkage results from the association analysis.* The estimates of SNP
255 effects from [5] were used to predict the linkage effects at each position. The
256 predicted effect at position *j* for sire *k* ($\eta_{j,k}$) was calculated as:

257 $$\eta_{j,k} = \sum_{i=1}^{M} \hat{\gamma}_i p_{i,j} x_{i,k,1} - \sum_{i=1}^{M} \hat{\gamma}_i p_{i,j} x_{i,k,2} \quad [6]$$

258 where $\hat{\gamma}_i$ is the estimate of the SNP allele effect at positions *i*, $p_{i,j}$ is the
259 probability of co-inheritance of positions *i* and *j*, $x_{i,k,1}$ and $x_{i,k,2}$ are sire k's allele
260 at position *i* (i.e. 0 or 1) for the first (*k=1)* and second *(k=2)* chromosomes and
261 *M* is the total number of SNP positions on the chromosome. Thus [6] is the
262 difference between the sum of allele effects for the first and second
263 chromosome at each position, where the sum of allele effects on each
264 chromosome accounts for the probability of recombination events along the
265 chromosome. The probability of co-inheritance of positions *i* and *j* was
266 calculated as $p_{i,j} = 1 - 2c_{i,j}$, where $c_{i,j}$ was the recombination fraction from
267 Haldane's mapping function (1919), i.e. $c_{i,j} = 0.5 [1-exp(-2m)]$ where *m* is the
268 distance (in Morgans) between *i* and *j* and assuming 1 cM = 1Mbp (Botstein *et*

269 *al.*, 1980 citing; Renwick, 1969). The regression coefficient of the observed
270 effect on the predicted linkage effect will be one if (1) the association analysis
271 captures all of the genetic information in the linkage analysis, (2) SNP allele
272 effects are additive and (3) Haldane's mapping function is an accurate
273 predictor of recombination.
274
275 *Predicting linkage results from the association analysis with independent data.*
276 The data from the association analysis used to predict the linkage effects in
277 [5] are not independent of the data used in the linkage analysis. This is
278 because the segregating alleles from the linkage analysis in the 12 sires also
279 contribute to the association analysis. To achieve complete independence
280 between the association and linkage analyses it was necessary to exclude, in
281 turn, the offspring of each sire from the association analysis. That is, model [5]
282 was run 12 times. SNP marker effects were then used to predict the linkage
283 results using [6] for the sire excluded from the association analysis. For
284 comparison, an analysis which predicts the between sire differences from
285 markers effects estimated from data including all sires and excluding the sire
286 to be predicted (i.e. independent data) is described in Part C of the
287 supplementary materials.
288

289 3. Results

290 *Tracking the paternal alleles.* Paternal alleles were assigned to either the 1st
291 or 2nd chromosome of the sire at 92.1% of positions (range per sire: 81.5 to
292 95.8%), excluding uninformative positions (Supplementary Figure S1). There
293 was an average of 7.2% unassigned progeny per SNP per sire.

294

295 *Linkage analysis and GWAS using conventional methods.* Using the
296 conventional fixed effect linkage analysis ([1]), 3109 positions were identified
297 as significant on 15 of 26 chromosomes at a false discovery rate of 14.8% (P
298 < 0.01, Figure 1). When significant SNP were tested using the genome-wide
299 association analysis ([4]), there are 132 SNP identified as significant with a
300 false-discovery rate of 22.8% (P < 0.01), SNP details in Supplementary Table
301 S1). The false-discovery rate suggests many true discoveries, although the
302 closer inspection below creates some confusion for QTL underlying our trait.
303
304 Doubts over the results from the conventional analysis arise because some
305 chromosomes suggest discrete QTL while for other chromosomes the results
306 are inconsistent. For example, consider chromosomes 3 and 6 (Figure 2).
307 Chromosome 3 presents seemingly reliable answers where the 43 positions
308 significant in both analyses appear to cluster near two likely QTL, one at
309 (approx) 30Mbp and another at 50Mbp. The effect of the SNP with the highest
310 significance from the association analysis at about 50 Mbp is -0.39 (± 0.08)
311 mm and the estimated (absolute) effect ranges from 0.01 (± 0.27) to 0.71 (±
312 0.38) mm for the linkage analysis. However, chromosome 6 shows a strong
313 linkage signal from 80Mbp onwards and 21 SNP significant from both the
314 linkage and association analysis over a wide region. It is not clear which, or if
315 all, these SNP are associated with the linkage peak. The linkage analysis
316 suggests possibly 3 QTL while the SNP also significant in the association
317 analysis suggests maybe 4 or more QTL. Also contradictory are the several
318 significant SNP at about 40Mbp which do not have any corresponding linkage

319 signal. It is difficult to ascertain using the two approaches in this form, which
320 analysis is more reliable, which effects are due to experimental noise, how
321 many QTL exist and what is the best estimate of the position of each QTL.
322 <Figure 1; Figure 2>
323
324 *Linkage analysis – impact of the random effects model.* The mean maximum
325 likelihood estimate for $\sigma^2_{sire.snp}$ from all positions was 0.013, and thus the
326 average proportion of phenotypic variance explained by the paternally
327 inherited allele was 0.0037 (i.e. $\sigma^2_{sire.snp}/\ \sigma^2_{phen}$ = 0.013 / 3.15). Although the
328 likelihood failed to converge at 5407 (11.1% of all) positions; a subsequent
329 restricted (positive definite) maximum likelihood analysis at these positions
330 showed an almost zero variance attributed to $\sigma^2_{sire.snp}$. This method
331 overestimates the average proportion of phenotypic variance explained by
332 markers because the sum for all markers is much greater than the genetic
333 variance of the trait (i.e. if the genetic variance is 0.3 σ^2_{phen}; 0.0037 σ^2_{phen} per
334 SNP *x* 48,640 SNP > 0.3). The overestimation occurs because of the strong
335 LD between makers in the linkage analysis.
336
337 Comparison of the fixed and random effects models for SNP alleles from the
338 linkage analysis (i.e. models [2] and [3]) shows broad agreement for most
339 sires at most positions (Figure 3A). The regression indicates that the random
340 effects analysis explains 91% of the variation in the fixed effect analysis but
341 that the fixed effect model is estimating the size of the allele effect to be about
342 10 times greater than the random effect model. Adjacent allele effects for a
343 sire are correlated in Figure 3A (i.e. adjacent SNP positions have correlated

344 effects and form lines in the plot). This correlation between positions is
345 maintained by the random model but the estimated size of effect is reduced.
346 Notably there are several SNP positions with large effects estimated by the
347 fixed model (> ± 2 mm) for which the random model estimates an effect near
348 zero. This more severe regression by the random effects model suggests that
349 there was little support for the large effect estimated by the fixed model.
350 These positions are probably regions where poor tracking of the paternal
351 allele occurred and, consequently, there were few progeny who were
352 recorded to inherit each of the sire's alleles.
353 <Figure 3>
354
355 *Association study – impact of the random effects model.* The regression of the
356 association allele effects from the fixed and random models (i.e. [4] and [5])
357 show that the fixed model estimates the effect of alleles almost 100 times
358 larger than the random model (Figure 3*b*). The regression of the fixed effect
359 solutions on the random effects solutions also explains a lower amount of
360 variation compared to the linkage analysis (R^2 = 0.58). The differences
361 between the models and the lower proportion of variance explained by the
362 random effect model is partially due to over-estimation of the effects when
363 they are fitted one at a time as fixed effects and partially because the BayesA
364 analysis may spread the effect of each QTL over several adjacent SNP. For
365 example, Figure 4 compares the fixed and BayesA analysis over a 20Mbp
366 region on chromosome 6 where there appears to be a strong QTL signal at
367 around 42Mbp. The random effects analysis maps this effect in a location
368 slightly further along the chromosome (41.5 Mbp) compared to the fixed effect

369 analysis (40.8 Mbp), but it also shows the spread of QTL effects for SNP in
370 modest LD (r^2 > 0.5) with this SNP in the region. Further, from the random
371 effects model, it is clearer that there are possibility 3 QTL at 30.7, 45.0 and
372 50.6 Mbp for markers which are not in strong LD with the SNP at 41.5 Mbp. A
373 further SNP at 42.1 Mbp may be associated with the same QTL tracked by the
374 SNP at 41.5 Mbp or this association could indicate another nearby QTL.
375 Similar to the linkage analysis, many SNP alleles estimated with large effects
376 (> ± 1mm) from the fixed model were regressed to almost zero using the
377 simultaneous method (Figure 3*b*). This occurs because of unreliable
378 estimates of effects from the fixed effect model. For example, of the 23
379 markers with large effects (> ± 1mm) from fixed effect model and very small
380 effects (< 0.001 mm) in the random model, 20 (87%) were not significant (P >
381 0.05). The remaining 3 markers may represent spurious results from the
382 standard GWAS, presumably caused by LD with other SNP.
383 <Figure 4>
384

385 *Predicting the linkage results from the association study.* Despite the
386 correction for bias in the linkage and association analyses the magnitude of
387 the association effects are still in the order of 100 times smaller than those
388 estimated from the linkage analysis (Figure 3). A prediction of the linkage
389 results from the association analysis needs to account for the stronger LD
390 between adjacent positions in the linkage analysis. Using the linkage results
391 from random model (i.e. [3]), the prediction was the contrast between sire
392 chromosomes for the sum of the association effects accounting for
393 recombination (i.e. models [5] and [6]). For individual sires, the expectation of

394 the linkage effects shows good agreement with the linkage results (Figure 5,
395 Supplementary Figure S2). To compare the effects across all sires and at all
396 positions we plotted the estimate from the linkage analysis against that
397 predicted from the association study (Figure 6*a*). The regression is almost
398 one (slope: $0.975 \pm 1.2 \times 10^{-3}$, intercept: $3.7 \times 10^{-3} \pm 6.9 \times 10^{-5}$) and accounts for
399 about half of the variation in the linkage results ($R^2 = 0.523$). Considering the
400 sampling errors in both estimates, this suggests that the association analysis
401 is capturing the majority of the within-family information. There were no data
402 points which showed a notable deviation from the regression slope
403 (Supplementary Figure S3).
404 <Figure 5; Figure 6>
405

406 *Predicting the linkage results with independent data.* There was a high
407 correlation between the SNP effects estimated with all animals and those
408 estimated excluding progeny from each sire using the random effects model
409 (average $R^2 = 0.91$, range: 0.85 to 0.93). However these analyses predicted
410 the linkage effects for the excluded sire very inaccurately (Figure 6*b*, R^2 =
411 0.002). This contrasts sharply to results when the sire to be predicted is
412 included in the analysis (Figure 6*a*). Thus the sire whose linkage analysis is to
413 be predicted must be included in the association analysis to achieve good
414 agreement between the two approaches. Predictive ability with independent
415 data is slightly improved when predicting differences between sires (R^2 =
416 0.04, appendix 3).
417

418 4. Discussion

419 This study suggests two reasons why there is often little agreement between
420 linkage analysis and GWAS on the same complex trait. First, when the effects
421 are estimated as fixed effects in statistical models, the most significant effects
422 are often grossly overestimated. This is evident in our study for both the
423 linkage and association analysis. Overestimation of fixed effects has been
424 highlighted previously by several authors (e.g. Beavis, 1998) and contributes
425 to the often smaller than expected or perhaps non significant results for loci
426 when replication is attempted. Naturally this problem also occurs if one
427 attempts to verify the results of a linkage analysis with a GWAS or vice versa.
428 Our GWAS predicted the linkage results provided both are estimated as
429 random effects, SNP are fitted simultaneously in the GWAS, and GWAS
430 effects on a chromosome are combined to account for LD in the linkage
431 analysis. The regression of the observed linkage effect on the effect predicted
432 from the GWAS is close to 1.0 indicating an approximate agreement in size.
433 The proportion of the variance in the linkage results explained by our
434 prediction is high (R^2 = 0.52) considering that both sets of effects are
435 estimated with error.
436
437 Second, multiple linked QTL can be the underlying cause of significant linkage
438 results. In contrast to the simulation studies with multiple QTL tracked by
439 microsatellite markers (e.g. Haley & Knott, 1992), our results in real data
440 suggest that likelihood peaks can be caused by the sum of many QTL along a
441 chromosome. We do not suggest that all linkage peaks are detecting multiple
442 small QTL because some studies have been successful in identifying
443 important loci (e.g. Charlier *et al.*, 1995; Coppieters *et al.*, 1998; Gusella *et al.*,

444 1983; Tsui *et al.*, 1985). However, successful linkage studies involve
445 polymorphisms of large effect and these loci probably overwhelm any
446 interference in the signal caused by multiple linked loci. The effect of the
447 linked loci could be to increase or decrease the apparent effect size of the
448 major loci, depending on the phase of the interacting loci. Here we
449 demonstrate with real data that the additive effect of multiple loci in strong LD
450 can cause apparent linkage signals. This conclusion is consistent with
451 simulation and theoretical studies (e.g. Dekkers & Dentine, 1991; Visscher &
452 Haley, 1996) and is also supported by mice studies when single QTL
453 fractionate into multiple smaller loci with fine mapping (Flint *et al.*, 2005).
454
455 The influence of nearby linked loci cannot be excluded when using
456 association rather than linkage analysis. Even in a conventional GWAS
457 analysis, fitting one SNP at a time, SNP with significant effects may be
458 influenced by multiple nearby QTL, some in phase and some out of phase
459 with the tested SNP. However, LD in GWAS probably has less influence than
460 in linkage because LD usually extends for shorter distances, i.e. < 1 Mbp in
461 Merino sheep (Kemper *et al.*, 2011). Hence a large number of significant SNP
462 most likely indicates a large number of QTL. This conclusion is made clearer
463 by fitting all SNP simultaneously. Then SNP which have no marginal effect
464 after fitting all other SNP, including SNP in strong LD with the causal
465 polymorphisms, will show no association with the trait. Figure 4 shows a
466 typical result where there are several positions along the chromosome
467 associated with the trait of interest.
468

469 The high degree of agreement (R^2 = 0.52, regression coefficient ~ 1.0)
470 between our observed and predicted linkage results is surprising. This
471 consistency suggests that the association analysis is tracking the majority of
472 the linkage information and that imperfect LD (between causal mutations and
473 SNP) is not a strong influence on the results from our association analysis.
474 This is because the linkage analysis has strong LD within families and
475 imperfect LD is not limiting as it can be in GWAS. Incomplete LD between
476 common SNP and causative mutations has been hypothesised to be
477 responsible for ~ 50% of the genetic variation in human populations which is
478 not explained by common SNP (Yang *et al.*, 2010). Here, we suggest that the
479 importance of incomplete LD between SNP and causative mutations is
480 influenced strongly by genetic diversity. Our observation is supported by other
481 studies with domestic species where common SNP capture a high proportion
482 of the genetic variance (e.g. Boyko *et al.*, 2010; Daetwyler, 2009; Haile-
483 Mariam *et al.*, 2012). Thus, as the population's diversity, or effective
484 population size (N_e), increases the ability of common SNP to capture the
485 genetic variance reduces. Incomplete LD may occur when causative SNP are
486 at a lower frequency than the genotyped SNP (Yang *et al.*, 2010), suggesting
487 an increased importance for these mutations in, for example, human
488 compared to livestock populations.
489
490 Extensive QTL mapping experiments in many species suggests that alleles
491 with a large effect on quantitative traits are uncommon (e.g. Darvasi &
492 Pisanté-Shalom, 2002). The results of the association analysis reported here
493 suggest many QTL for our trait but we found no evidence of large effect QTL

494 in our sires. For instance, if most important genes had a variant with large
495 effect, we might expect to see at least one sire with a large estimated effect
496 from the linkage analysis and an inaccurate prediction of this effect from the
497 GWAS. However, we never observed any alleles from the linkage analysis
498 which substantially differed from the effect predicted from the association
499 analysis (Figure 5). We sampled only 12 sires but we analysed each sire at
500 thousands of positions. If most of the genetic variance was due to rare large
501 effect variants then we might expect to observe at least one heterozygous sire
502 in our dataset. The situation of segregating alleles with large effect may occur
503 but it cannot be typical because we predicted our linkage results from an
504 association analysis with moderate accuracy. Further, all of our estimated
505 effects from the association analysis were also very small (< 0.008 mm or <
506 $0.008 / 3.15^{1/2} = 0.004$ SD).
507
508 Our results show that most of the linkage information was captured in the
509 prediction from the GWAS results. However the two approaches are not
510 independent because they use the same data and we also show that when
511 the sire to be predicted is excluded from the association analysis we cannot
512 predict the linkage results. This discrepancy could be explained by high
513 sampling covariance between the effects estimated for SNP in very strong LD
514 with one another. Thus the combination of SNP alleles has been observed in
515 the data to be predicted accurately. The between sire differences, which are
516 the sum of all SNP effects, were estimated more precisely using independent
517 data (appendix 3). Prediction of between sire differences is equivalent to
518 genomic prediction which, given larger datasets, can reach moderate

519 accuracies in sheep for this trait (Daetwyler *et al.*, 2010). The dependency
520 between SNP when estimating effects of individual markers is not surprising
521 considering that the magnitude of the largest effect was very small (0.004 SD)
522 and given the relatively small size of the dataset.
523
524 These results suggest that the best analysis is the GWAS in which all SNP
525 are fitted simultaneously. This method gave us consistent results between
526 linkage and association and has greater power to discriminate linked QTL
527 than either the linkage analysis or the standard GWAS fitting one SNP at a
528 time. This is clearly demonstrated in Figure 4 where the numerous GWAS
529 results are consolidated into possibly 4 QTL signals at 41.5, 42.1, 45.0 and
530 50.6 Mbp. A potential drawback of this method is that effects may be split
531 between closely linked markers (Xu, 2003a). In Figure 4, this is potentially
532 occurring for several markers in high LD with the largest estimated effect at
533 41.5 Mbp. These high LD markers may also be capturing multiple mutations at
534 the locus. However the effect of this disadvantage should diminish as markers
535 in higher LD with the causal mutations for traits are included in the SNP
536 marker set.
537
538 In summary this study aimed to reconcile some of the differences between
539 linkage and linkage-disequilibrium mapping. We have demonstrated, using
540 real data, the correction for the biases in both linkage and association
541 mapping. We show that multiple linked QTL can combine to be the primary
542 cause of significant linkage results. In our study, the association analysis
543 captured 52% of the within-family information, which is high considering the

544 sampling error of effects from both analyses. The results support the
545 hypothesis that there are many loci of small effect underlying complex traits.
546 We suggest an improved method for GWAS is to fit statistical models where
547 all SNP are analysed simultaneously. This method prevents spurious results
548 caused by population structure and accounts for LD surrounding the analysed
549 SNP.

550

Acknowledgements

The authors would like to 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).

Declaration of interest

None.

560 # References

561 Aitman, T. J., Boone, C., Churchill, G. A., Hengartner, M. O., Mackay, T. F. C.
562 &Stemple, D. L. (2011). The future of model organisms in human disease
563 research. *Nature Review Genetics* **12,** 575-582.
564 Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. &Wjst, M. (2001). Genomewide
565 scans of complex human diseases: True linkage is hard to find. *The American*
566 *Journal of Human Genetics* **69,** 936-950.
567 Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D.*, et al.*
568 (2008). Genome-wide association defines more than 30 distinct susceptibility
569 loci for Crohn's disease. *Nat Genet* **40,** 955-962.
570 Beavis, W. D. (1998). QTL analysis: power, precision and accuracy. *Molecular*
571 *Dissection of Complex Traits***,** 145 - 62.
572 Bolormaa, S., Hayes, B. J., Savin, K., Hawken, R., Barendse, W., Arthur, P. F.*, et al.*
573 (2011). Genome-wide association studies for feedlot and growth traits in
574 cattle. *Journal of Animal Science*.
575 Botstein, D., White, R. L., Skolnick, M. &Davis, R. W. (1980). Construction of a
576 genetic linkage map in man using restricted fragment length polymorphisms.
577 *American Journal of Human Genetics* **32,** 314-331.
578 Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., Lohmueller,
579 K. E.*, et al.* (2010). A simple genetic architecture underlies morphological
580 variation in dogs. *PLoS Biology* **8,** e1000451.
581 Charlier, C., Coppieters, W., Farnir, F., Grobet, L., Leroy, P. L., Michaux, C.*, et al.*
582 (1995). The *mh* gene causing double-muscling in cattle maps to bovine
583 Chromosome 2. *Mammalian Genome* **6,** 788-792.
584 Clark, A. G., Nielsen, R., Signorovitch, J., Matise, T. C., Glanowski, S., Heil, J.*, et al.*
585 (2003). Linkage disequilibrium and inference of ancestral recombination in
586 538 single-nucleotide polymorphism clusters across the human genome. *The*
587 *American Journal of Human Genetics* **73,** 285-300.
588 Coppieters, W., Riquet, J., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B.*, et al.*
589 (1998). A QTL with major effect on milk yield and composition maps to
590 bovine Chromosome 14. *Mammalian Genome* **9,** 540-544.
591 Daetwyler, H. D. (2009). *Genome-wide evaluation of populations.* Wageningen
592 University.
593 Daetwyler, H. D., Hickey, J., Henshall, J., Dominik, S., Gredler, B., van der Werf, J.
594 H. J.*, et al.* (2010). Accuracy of estimated genomic breeding values for wool
595 and meat traits in a multi-breed sheep population. *Animal Production Science*
596 **50,** 1004-1010.
597 Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M. &Robinson, J. A. B. (2008). A
598 genome scan to detect quantitative trait loci for economically important traits
599 in Holstein cattle using two methods and a dense single nucleotide
600 polymorphism map. *Journal of Dairy Science* **91,** 3225-3236.
601 Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A. &Goddard., M. E.
602 (2011). Imputation of genotypes from sparse to high density using long-range
603 phasing. *Genetics* **189,** 317-327.
604 Darvasi, A.Pisanté-Shalom, A. (2002). Complexities in the genetic dissection of
605 quantitative trait loci. *Trends in Genetics* **18,** 489-491.

606 Darvasi, A., Weinreb, A., Minke, V., Weller, J. I. &Soller, M. (1993). Detecting
607 marker-QTL linkage and estimating QTL gene effect and map location using a
608 saturated genetic map. *Genetics* **134,** 943-951.
609 Dekkers, J. C. M.Dentine, M. R. (1991). Quantitative genetic variance associated with
610 chromosomal markers in segregating populations. *TAG Theoretical and*
611 *Applied Genetics* **81,** 212-220.
612 Flint, J., Valdar, W., Shifman, S. &Mott, R. (2005). Strategies for mapping and
613 cloning quantitative trait genes in rodents. *Nature Review Genetics* **6,** 271-286.
614 Gilmour, A. R., Gogel, B. J., Cullis, B. R. &Thompson, R. (2006). ASReml user
615 guide 2.0. VSN International Ltd., Hemel Hempsted, UK.
616 Goddard, M. E., Wray, N. R., Verbyla, K. &Visscher, P. M. (2009). Estimating
617 effects and making predictions from genome-wide marker data. *Statistical*
618 *Science* **24,** 514-529.
619 Goring, H. H. H., Terwilliger, J. D. &Blangero, J. (2001). Large upward bias in
620 estimation of locus-specific effects from genomewide scans. *American*
621 *Journal of Human Genetics* **69,** 1357-1369.
622 Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi,
623 R. E.*, et al.* (1983). A polymorphic DNA marker genetically linked to
624 Huntington's disease. *Nature* **306,** 234-238.
625 Haile-Mariam, M., Nieuwhof, G. J., Beard, K. T., Konstatinov, K. V. &Hayes, B. J.
626 (2012). Comparison of heritabilities of dairy traits in Australian Holstein-
627 Friesian cattle from genomic and pedigree data and implications for genomic
628 evaluations. *Journal of Animal Breeding and Genetics* **early view article**.
629 Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of
630 distances between the loci of linkaged factors. *Journal of Genetics* **8,** 299-309.
631 Haley, C. S.Knott, S. A. (1992). A simple regression method for mapping quantitative
632 trait loci in line crosses using flanking markers. *Heredity* **69,** 315-324.
633 Henderson, C. R. (1950). Estimation of genetic parameters. *Annuals of Mathematical*
634 *Statistics* **21,** 309.
635 Huisman, A. E.Brown, D. J. (2009). Genetic parameters for bodyweight, wool, and
636 disease resistance and reproduction traits in Merino sheep. 3. Genetic
637 relationships between ultrasound scan traits and other traits. *Animal*
638 *Production Science* **49,** 283-288.
639 Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B.*, et*
640 *al.* (2010). Variance component model to account for sample structure in
641 genome-wide association studies. *Nature Genetics* **42,** 348-354.
642 Kemper, K. E., Emery, D. L., Bishop, S. C., Oddy, H., Hayes, B. J., Dominik, S.*, et*
643 *al.* (2011). The distribution of SNP marker effects for faecal worm egg count
644 in sheep, and the feasibility of using these markers to predict genetic merit for
645 resistance to worm infections. *Genetics Research* **93,** 203-219.
646 MacLeod, I. M., Hayes, B. J., Savin, K. W., Chamberlain, A. J., McPartlan, H. C.
647 &Goddard, M. E. (2010). Power of a genome scan to detect and locate
648 quantitative trait loci in cattle using dense single nucleotide polymorphisms.
649 *Journal of Animal Breeding and Genetics* **127,** 133-42.
650 Martínez, O.Curnow, R. N. (1992). Estimating the locations and the sizes of the
651 effects of quantitative trait loci using flanking markers. *TAG Theoretical and*
652 *Applied Genetics* **85,** 480-488.
653 McKenzie, C. A., Abecasis, G. R., Keavney, B., Forrester, T., Ratcliffe, P. J., Julier,
654 C.*, et al.* (2001). Trans-ethnic fine mapping of a quantitative trait locus for

655 circulating angiotensin I-converting enzyme (ACE). *Human Molecular*
656 *Genetics* **10,** 1077-1084.
657 Meuwissen, T. H. E., Hayes, B. J. &Goddard, M. E. (2001). Prediction of total genetic
658 value using genome-wide dense marker maps. *Genetics* **157,** 1819-1829.
659 Mortimer, S. I., van der Werf, J. H. J., Jacob, R. H., Pethick, D. W., Pearce, K. L.,
660 Warner, R. D.*, et al.* (2010). Preliminary estimates of genetic parameters for
661 carcass and meat quality traits in Australian sheep. *Animal Production Science*
662 **50,** 1135-1144.
663 Mrode, R. A. (2005). *Linear Models for the prediction of Animal Breeding Values*.
664 CABI Publishing, Oxfordshire, UK.
665 Patterson, N., Price, A. L. &Reich, D. (2006). Population structure and eigenanalysis.
666 *PLoS Genetics* **2,** e190.
667 Renwick, J. H. (1969). Progress in mapping human autosomes. *British Medical*
668 *Bulletin* **25,** 65-73.
669 Safari, E., Fogarty, N. M. &Gilmore, A. R. (2005). A review of genetic parameters
670 estimates for wool, growth, meat and reproduction traits in sheep. *Livestock*
671 *Production Science* **92,** 271-289.
672 Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal*
673 *Statistical Society: Series B (Statistical Methodology)* **64,** 479-498.
674 Sun, L., Dimitromanolakis, A., Faye, L. L., Paterson, A. D., Waggott, D. &Bull, S. B.
675 (2011). BR-squared: A practical solution to the winner's curse in genome-wide
676 scans. *Human Genetics* **129,** 545-552.
677 Tsui, L., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, J.*, et al.*
678 (1985). Cystic fibrosis locus defined by a genetically linked polymorphic
679 DNA marker. *Science* **230,** 1054-1057.
680 Visscher, P.Haley, C. (1996). Detection of putative quantitative trait loci in line
681 crosses under infinitesimal genetic models. *TAG Theoretical and Applied*
682 *Genetics* **93,** 691-702.
683 White, J. D., Allingham, P. G., Gorman, C. M., Emery, D., Hynd, P., Owens, J.*, et al.*
684 (2012). Design and phenotyping proceedures for recording wool, skin, parasite
685 resistance, growth, carcass yield and quality traits of the SheepGENOMICS
686 mapping flock. *Animal Production Science* **52,** 157-171.
687 Xiao, R.Boehnke, M. (2011). Quantifying and correcting for the winner's curse in
688 quantitative-trait association studies. *Genetic Epidemiology* **35,** 133-138.
689 Xu, S. (2003a). Estimating polygenic effects using markers of the entire genome.
690 *Genetics* **163,** 789-801.
691 Xu, S. (2003b). Theoretical basis of the Beavis effect. *Genetics* **165,** 2259-2268.
692 Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A., Nyholt, D. R.*, et al.*
693 (2010). Common SNPs explain a large proportion of the heritability for human
694 height. *Nature Genetics* **42,** 565-569.
695 Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham,
696 J. M.*, et al.* (2011). Genome partitioning of genetic variation for complex traits
697 using common SNPs. *Nature Genetics* **43,** 519-25.
698 Zöllner, S.Pritchard, J. K. (2007). Overcoming the Winner s Curse: Estimating
699 Penetrance Parameters from Case-Control Data. *American journal of human*
700 *genetics* **80,** 605-615.
701
702

703 # Figures

704

705 **Figure 1.** Comparison of the test statistics across the genome for linkage
706 (grey) and the association (black) analyses. Markers significant in both
707 analyses are highlighted in red ($P < 0.01$).

708

709 **Figure 2.** Comparison of test statistics for chromosomes 3 (*a*) and 6 (*b*) using
710 the linkage (grey) and association (black) analyses. Markers significant in
711 analyses are highlighted in red ($P < 0.01$).

712

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

718

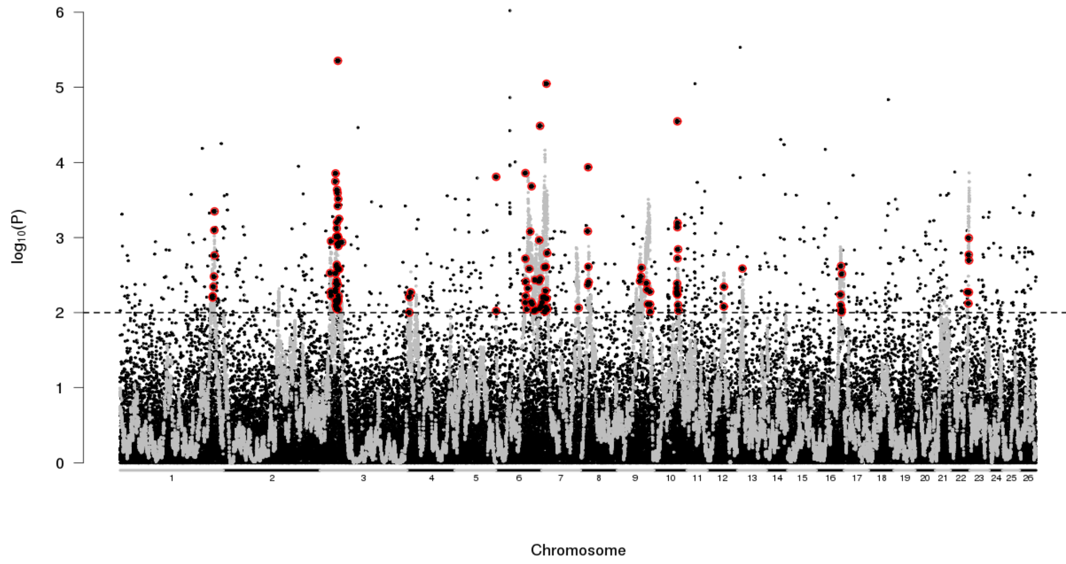
719 **Figure 4.** The absolute effect of SNP alleles when fitted as fixed (*a*) or
720 random (*b*) in the association analysis. Grey lines indicate the positions of the
721 largest effect in (*a*) or (*b*) with colours showing the linkage disequilibrium
722 (correlation) between these marked SNP and the surrounding markers.

723

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

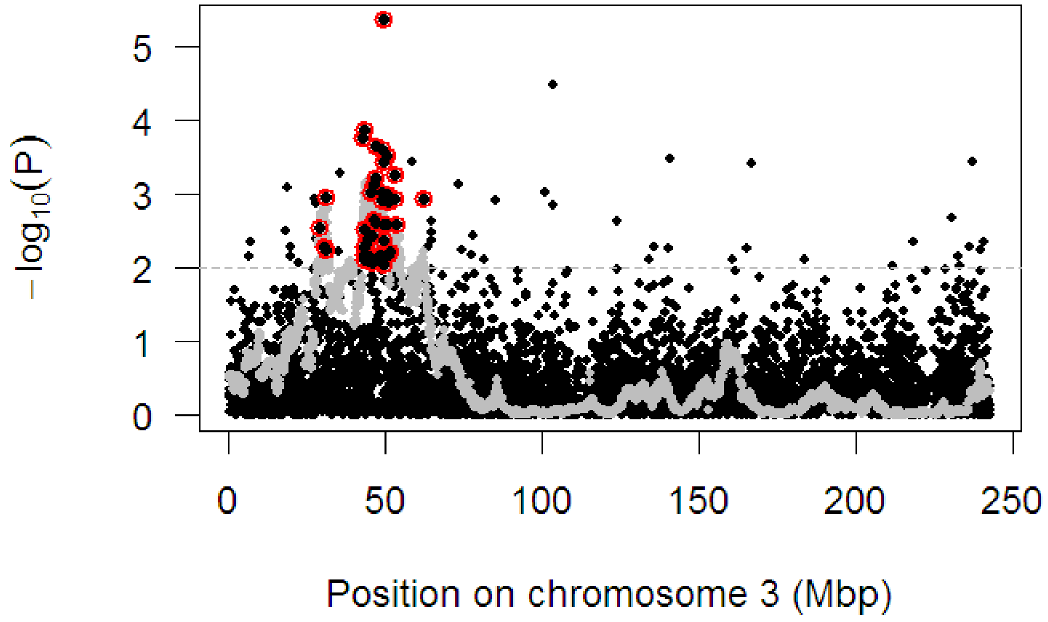
727

728 **Figure 6.** Marker effects (mm) estimated from linkage when alleles are fitted
729 as random (y-axis) or predicted from the sum of the association effects
730 accounting for recombination (x-axis). The association analysis either includes
731 all sires (*a*) or excludes the sire to be predicted (*b*).
732

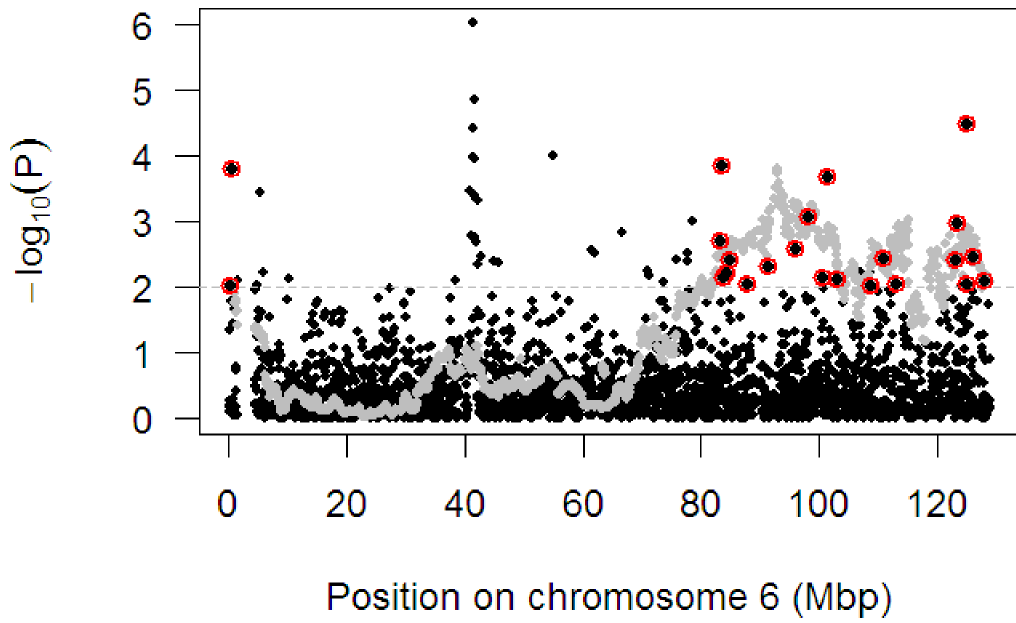


733

734 **Figure 1.**

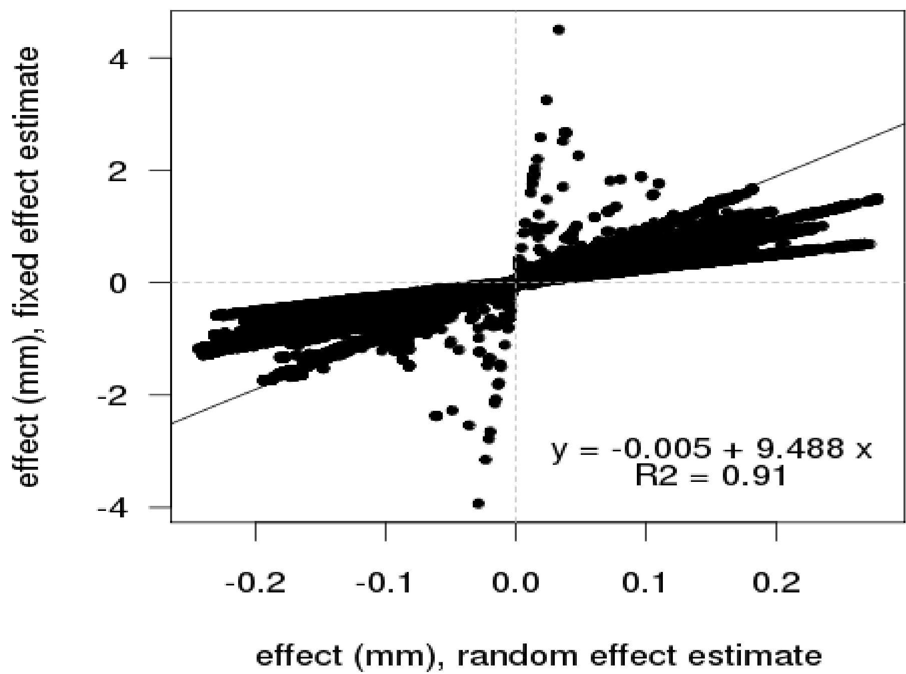


735

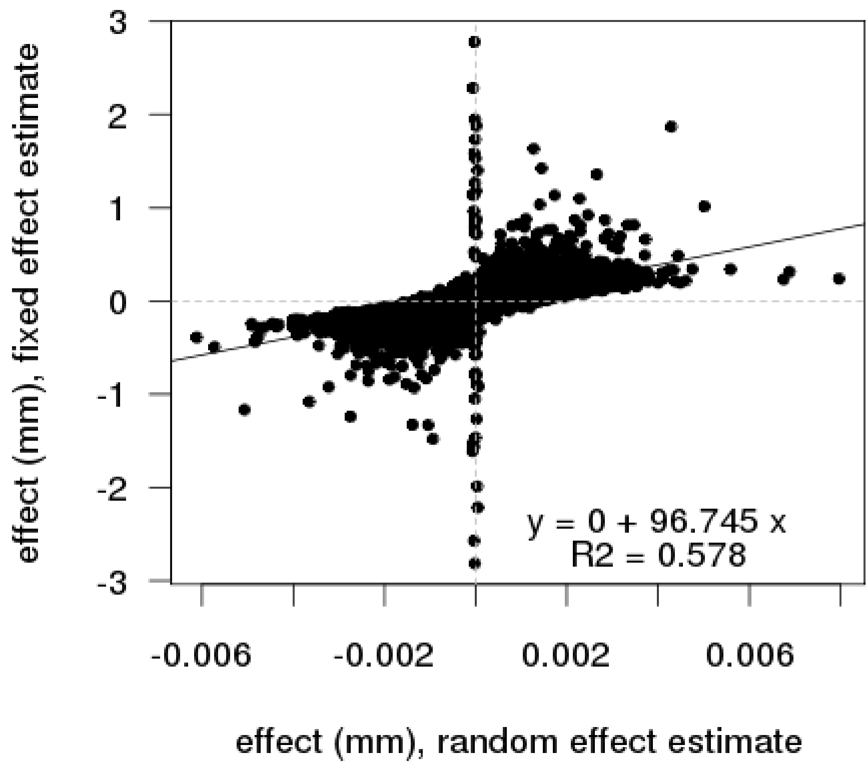


736

737 **Figure 2.**

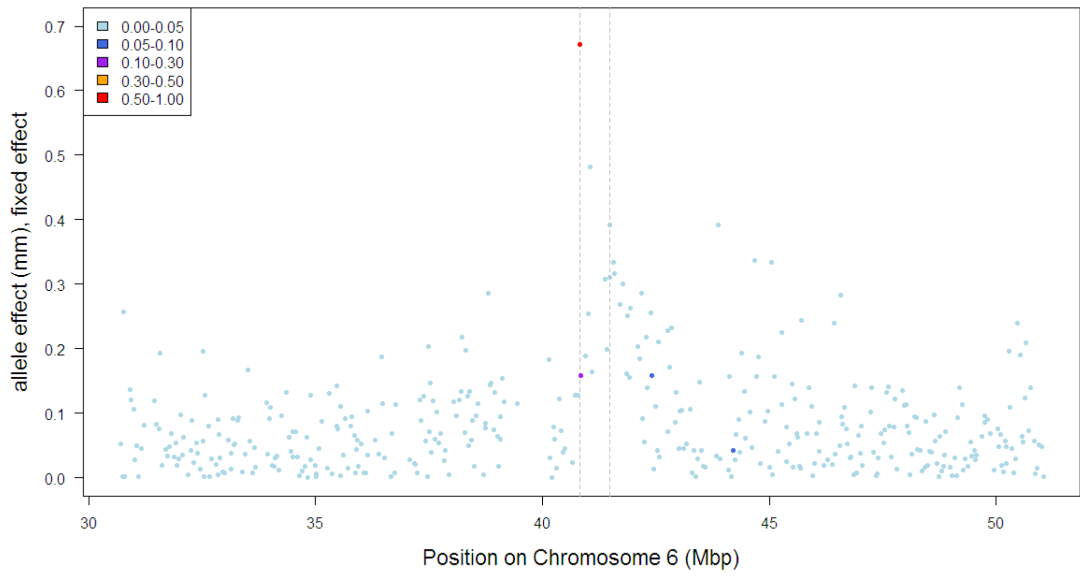


738

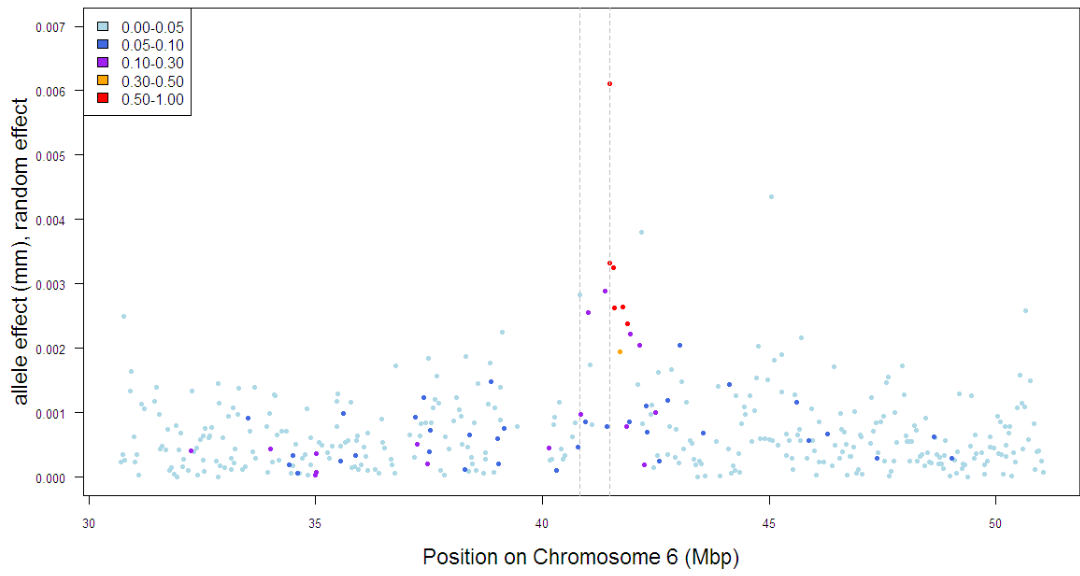


739

740 **Figure 3.**

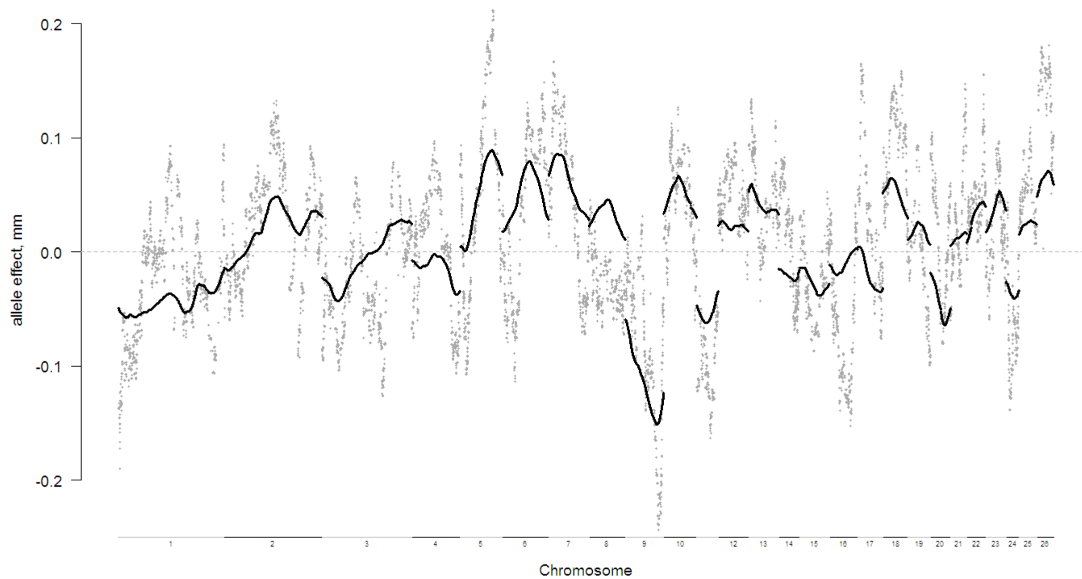


741



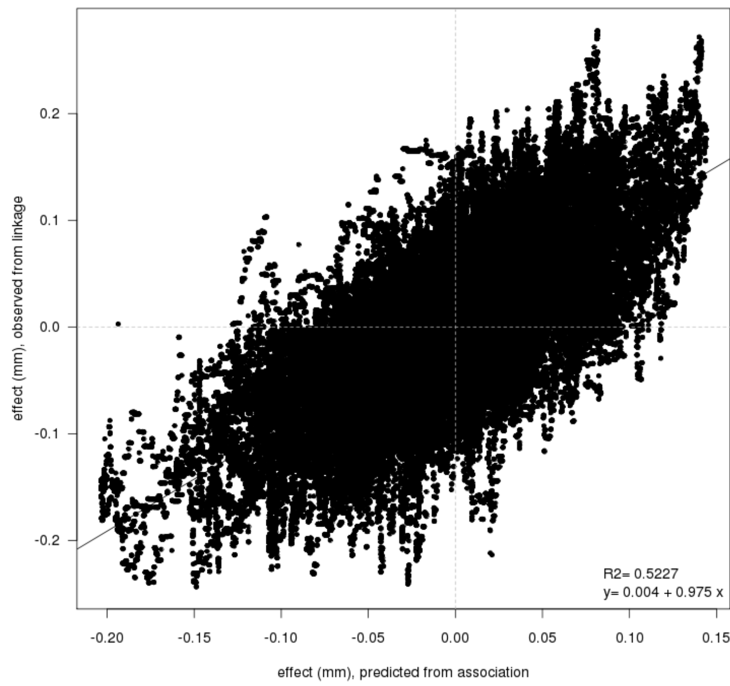
742

743 **Figure 4.**

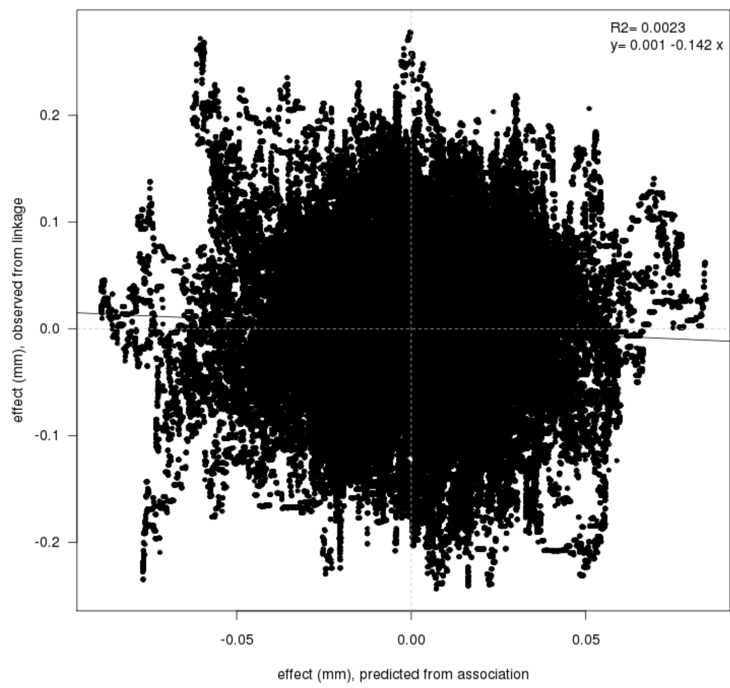


744

745 **Figure 5.**



746



747

748 **Figure 6.**