Heredity (2010) 105, 333–340

www.nature.com/hdy

ORIGINAL ARTICLE

Variation explained in mixed-model association mapping

G Sun[1,5], C Zhu[1,5], MH Kramer[2], S-S Yang[3], W Song[3], H-P Piepho[4] and J Yu[1]

[1]*Department of Agronomy, Kansas State University, Manhattan, KS, USA;* [2]*USDA-ARS, Beltsville, MD, USA;* [3]*Department of Statistics, Kansas State University, Manhattan, KS, USA;* [4]*Institute of Crop Production and Grassland Research, Bioinformatics Unit, University of Hohenheim, Stuttgart, Germany*

Genomic mapping of complex traits across species demands integrating genetics and statistics. In particular, because it is easily interpreted, the R^2 statistic is commonly used in quantitative trait locus (QTL) mapping studies to measure the proportion of phenotypic variation explained by molecular markers. Mixed models with random polygenic effects have been used in complex trait dissection in different species. However, unlike fixed linear regression models, linear mixed models have no well-established R^2 statistic for assessing goodness-of-fit and prediction power. Our objectives were to assess the performance of several R^2-like statistics for a linear mixed model in association mapping and to identify any such statistic that measures model-data agreement and provides an intuitive indication of QTL effect. Our results showed that the likelihood-ratio-based R^2 (R^2_{LR}) satisfies several critical requirements proposed for the R^2-like statistic. As R^2_{LR} reduces to the regular R^2 for fixed models without random effects other than residual, it provides a general measure for the effect of QTL in mixed-model association mapping. Moreover, we found that R^2_{LR} can help explain the overlap between overall population structure modeled as fixed effects and relative kinship modeled though random effects. As both approaches are derived from molecular marker information and are not mutually exclusive, comparing R^2_{LR} values from different models provides a logical bridge between statistical analysis and underlying genetics of complex traits.
Heredity (2010) **105,** 333–340; doi:10.1038/hdy.2010.11; published online 10 February 2010

Keywords: QTL mapping; association mapping; genetic variation; maize

Researchers in many disciplines use linear regression models widely. The R^2 statistic, the coefficient of determination, is one of the most frequently used measures of prediction power and goodness-of-fit for simple linear regression models (Draper and Smith, 1981; Everitt, 2002). In the literature on genetics, researchers often report R^2 values of newly identified genetic loci in addition to effect sizes and *P*-values (Lettre *et al*., 2008; Weedon *et al*., 2008). For nonstandard linear regression models, however, several competing R^2-like statistics have been proposed to measure prediction power and goodness-of-fit (Buse, 1973; Magee, 1990; Xu, 2003; Kramer, 2005) but have not been used in genetics. Indeed, it is desirable to have a measure for general linear mixed models analogous in some ways to the R^2 of the linear regression model, which has a 'variation explained' interpretation.

Association mapping searches the association between genetic markers and complex traits (for example disease susceptibility) based on populations (Hirschhorn and Daly, 2005). It complements linkage analysis in mapping the genetic basis of complex traits. Mixed models have long been used in genetic research (Henderson, 1984; Lynch and Walsh, 1998), and the mixed-model association mapping methods were developed to account for complex population structure (Meuwissen *et al*., 2002; Yu *et al*., 2006; Malosetti *et al*., 2007). Although statistics like deviance and the Bayesian Information Criterion (BIC) (Schwarz, 1978) can be used to select models (Broman and Speed, 2002; Littell *et al*., 2006), many researchers desire a R^2-like statistic for mixed models because it can indicate the prediction power of various models containing different fixed and random effects and their associated variance–covariance structure. After identifying statistically significant genetic loci (Kennedy *et al*., 1992), many geneticists would ask how much of the phenotypic variation is explained by each quantitative trait locus (QTL) for the interpretation or comparison purpose. In other words, what is the relative degree of improvement of the model fit to the data that results by including this significant genetic effect. Moreover, R^2-like statistics complement statistical testing by providing practitioners with a more intuitive measurement than the *P*-value from other statistical tests (for example, likelihood-ratio (*LR*) test or F-test). Compared with statistics like deviance and BIC, R^2-like statistics offer an alternative, easier to grasp measurement for geneticists.

Several approaches can quantify the genetic relationship of a complex population in the context of association mapping using molecular marker information (Weir *et al*., 2006). The first approach was developed to examine population structure by estimating the probability of subgroup membership (Pritchard *et al*., 2000; Falush

Correspondence: Dr J Yu, Department of Agronomy, Kansas State University, 2004 Throckmorton Plant Sciences Center, Manhattan, KS 66506-5501, USA.
E-mail: jyu@ksu.edu
[5]*These authors contributed equally to this work.*
Received 13 October 2009; revised 22 December 2009; accepted 15 January 2010; published online 10 February 2010

et al., 2003). Recent research showed that principal component analysis (PCA) can also capture population differentiation (Price *et al*., 2006). A second approach focuses on the pairwise genetic relationship by estimating relative kinship (Loiselle *et al*., 1995; Ritland, 1996; Yu *et al*., 2006). As these two approaches are not orthogonal and the same marker data can be used to reflect population structure, principal components, and relative kinship, dependency among these different estimates is expected. Simultaneously fitting these estimates in the model, however, does not necessarily preclude the objective of controlling multiple levels of genetic relatedness within the association panel. In practice, the effects of controlling complex population structure with different estimates (population structure, principal components, and relative kinship) can vary by populations, traits, or both (Yu *et al*., 2006, 2009; Zhao *et al*., 2007; Zhu and Yu, 2009). A legitimate question, then, is whether a statistic like R^2 can be used to compare the different levels of control for genetic relationships.

Much of the literature on using R^2 for nonstandard linear models comes from statistics and econometrics, whereas such literature in the field of genetics is limited. Accordingly, our objectives were to assess the performance of several R^2-like statistics for a linear mixed model in association mapping and to identify a general R^2-like statistic that measures model-data agreement and provides an intuitive indication of the QTL effect. Although theoretical derivation or developing new statistics are beyond the scope of this study, we introduce four R^2-like statistics for nonstandard linear models, describe mixed-model association mapping, and test the performance of these four R^2-like statistics in the context of association mapping with computer simulations. We then apply these statistics to two empirical data sets.

Materials and methods

R^2 for fixed linear models

For the linear model with only fixed effects

$$y = X\beta + e \quad (1)$$

where y is an $n \times 1$ vector, X is an $n \times k$ matrix, β is a $k \times 1$ vector of unknown regression coefficients, and e is an $n \times 1$ vector consisting of *i.i.d*. normal variables with mean 0 and variance σ^2. Then the usual R^2 statistic is defined as

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

where $SSR = (\hat{y} - \bar{y})'(\hat{y} - \bar{y})$, $SSE = (y - \hat{y})'(y - \hat{y})$, $SSTO = (y - \bar{y})'(y - \bar{y})$, $(y - \bar{y})'(y - \bar{y}) = (\hat{y} - \bar{y})'(\hat{y} - \bar{y}) + (y - \hat{y})'(y - \hat{y})$, $\hat{y} = X\hat{\beta}$, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. As $0 \leqslant SSE \leqslant SSTO$, it follows that $0 \leqslant R^2 \leqslant 1$.

R^2 statistics for linear mixed models

The linear model with both fixed effects and random effects is

$$y = X\beta + Zu + e \quad (2)$$

where y is an $n \times 1$ observation vector, X is an $n \times k$ design matrix linked to the fixed effect, β is a $k \times 1$ vector of unknown regression coefficients of fixed effects, Z is an $n \times p$ design matrix linked to the random effects, u is a $p \times 1$ vector of random variables from a multivariate normal distribution (MVN) with zero means and variance–covariance matrix G (that is $u \sim$ MVN (0, G)), and e is an $n \times 1$ vector of random errors with zero means and variance–covariance matrix $I\sigma^2$ (that is $e \sim$ MVN (0, $I\sigma^2$)). Thus, y is MVN ($X\beta$, V) and $V = ZGZ' + I\sigma^2$. Several statistics have been proposed for mixed models (Table 1), and we describe them briefly in the following sections.

Two research groups (Cox and Snell, 1989; Magee, 1990) have independently proposed the likelihood-ratio-based R^2 (R^2_{LR}), a R^2-like statistic based on the *LR*:

$$R^2_{LR} = 1 - \exp(-\frac{2}{n}(\log L_M - \log L_0))$$

where $\log L_M$ is the maximum log-likelihood of the model of interest, $\log L_0$ is the maximum log-likelihood of the intercept-only model, n is the number of observations, and $\log L = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta) - \frac{n}{2}\log(2\pi)$. Please note that the calculation is based on maximum likelihood (ML), not restricted ML (REML). The same formula of R^2_{LR} was also suggested for the binary response models earlier by Maddala (1983). The *LR* statistic can be written as $LR = 2\log(L_M/L_0)$. The relationship between R^2_{LR} and *LR* is $R^2_{LR} = 1 - \exp(-LR/n)$. The R^2_{LR} statistic is appropriate when the concept of residual variance cannot

Table 1 Summary of different R^2 statistics for the linear mixed model

R^2 *statistics*	*Theoretical basis*	*Formula*	*References*
R^2_{LR}	Likelihood ratio	$1 - \exp(-\frac{2}{n}(LogL_M - LogL_0))$	(Magee, 1990)
R^2_W	Wald statistic	$R^2_{w1} = 1 - \frac{(y-\bar{y})'V^{-1}(y-\tilde{y})}{(y-\bar{y})'V^{-1}(y-\bar{y})}$	(Buse, 1973)
		$R^2_{w2} = 1 - \frac{(y-\tilde{y})'V^{-1}(y-\tilde{y})}{(y-\bar{y})'V^{-1}(y-\bar{y})}$	(Kramer, 2005)
r_c	Concordance correlation	$1 - \frac{(y-\hat{y})'(y-\hat{y})}{(y-\bar{y})'(y-\bar{y})+(\hat{y}-\bar{y})'(\hat{y}-\bar{y})+n(\bar{y}-\bar{y})^2}$	(Vonesh *et al*., 1996)
P_{rand}	Penalized quasi-likelihood function	$1 - \frac{(1/2\hat{\sigma})(y-\hat{y})'(y-\hat{y})+(1/2)\hat{u}'\hat{G}^{-1}\hat{u}}{(1/2\hat{\sigma})(y-\bar{y})'(y-\bar{y})}$	(Zheng, 2000)

$\bar{y} = X\hat{\beta}$, $\hat{y} = X\hat{\beta} + Z\hat{u}$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\tilde{y} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}$, $\bar{\bar{y}} = \frac{\xi' V^{-1} y}{\xi' V^{-1}\xi}$, $V = ZGZ' + I\sigma^2$, and $\xi' = (1,\ldots,1)$.

$LogL_M$ denotes the logarithm of maximum likelihood of the model of interest and $LogL_0$ for the intercept-only model.

be easily defined and ML is the criterion of fitting the model of interest. It can be shown that when the model only has fixed effects, R^2_{LR} is reduced to the traditional R^2 statistic. For discrete models like logistic regression, a scaling procedure should be applied to ensure the resulting R^2_{LR} is bounded between 0 and 1 (Nagelkerke, 1991).

The generalized least square R^2 statistic, R^2_W is defined as (Buse, 1973):

$$R^2_W = 1 - \frac{(y-\hat{y})'V^{-1}(y-\hat{y})}{(y-\bar{\bar{y}})'V^{-1}(y-\bar{\bar{y}})}$$

where $\hat{y} = X\hat{\beta}$ is the best predictor of y, and $\bar{\bar{y}}$ is the weighted mean: $\bar{\bar{y}} = (\xi'V^{-1}y/\xi'V^{-1}\xi)$ with $\xi' = (1,\ldots,1)$. This original definition is denoted as R^2_{W1}. It can be shown that, with $\hat{y} = X\hat{\beta}$, there is a direct summation relationship:

$$(y-\bar{\bar{y}})'V^{-1}(y-\bar{\bar{y}}) = (y-\hat{y})'V^{-1}(y-\hat{y}) + (\hat{y}-\bar{\bar{y}})'V^{-1}(\hat{y}-\bar{\bar{y}})$$

Replacing $\hat{y}$ with $\hat{y} = X\hat{\beta} + Z\hat{u}$ in R^2_W yields the R^2_{W2} statistic, $R^2_{W2} = 1 - ((y-\hat{y})'V^{-1}(y-\hat{y}))/((y-\bar{\bar{y}})'V^{-1}(y-\bar{\bar{y}}))$ (Kramer, 2005). There is no direct summation relationship for components in R^2_{W2}. In addition, it is difficult to interpret the numerator term where V^{-1}, rather than $(I\sigma^2)^{-1}$, is used because the random term appears in both $(y-\hat{y}) = (y-(X\hat{\beta} + Z\hat{u}))$ and $V = ZGZ' + I\sigma^2$. Here $\hat{y}$ is marginal because the prediction only involves fixed effects, but $\hat{y}$ is conditional because the prediction is conditional on random effects (Vonesh *et al*., 1996; Vonesh and Chinchilli, 1997; Littell *et al*., 2006). Note that when the model has only fixed effects, both forms of R^2_W are reduced to the traditional R^2 statistic.

The r_c statistic is a goodness-of-fit measure originally derived for the generalized nonlinear mixed-effect model, following the unweighted concordance correlation coefficient (ρ_c) (Vonesh *et al*., 1996):

$$r_c = 1 - \frac{(y-\hat{y})'(y-\hat{y})}{(y-\bar{y})'(y-\bar{y}) + (\hat{y}-\tilde{y})'(\hat{y}-\tilde{y}) + n(\bar{y}-\tilde{y})^2}$$

where n is the number of observations, $\bar{y}$ is the mean of y, $\hat{y} = X\hat{\beta} + Z\hat{u}$, and $\tilde{y}$ is the mean of $\hat{y}$. With $\hat{y} = X\hat{\beta} + Z\hat{u}$, both fixed and random effects are used to measure goodness-of-fit and prediction power and r_c is conditional (Vonesh *et al*., 1996; Vonesh and Chinchilli, 1997). The r_c statistic can be interpreted as a measure of the degree of agreement between the observed values and the predicted values as ρ_c measures agreement between two random variables. The possible values of r_c lie in the range $-1 \leqslant r_c \leqslant 1$.

The P_{rand} statistic measures the proportional reduction in the penalized quasi-likelihood function assuming MVN random effects (Zheng, 2000):

$$P_{\text{rand}} = 1 - \frac{-PQL_M}{-PQL_0} = 1 - \frac{(1/2\hat{\sigma})(y-\hat{y})'(y-\hat{y}) + (1/2)\hat{u}'\hat{G}^{-1}\hat{u}}{(1/2\hat{\sigma})(y-\bar{y})'(y-\bar{y})}$$

where PQL_M denotes a penalized quasi-likelihood function for the model of interest, PQL_0 denotes a penalized quasi-likelihood function for the null model where the model contains only the intercept, $\hat{u}$ is the estimated best linear unbiased predictor of u, $\hat{y} = X\hat{\beta} + Z\hat{u}$ is the estimated best linear unbiased predictor of y, $\hat{G}$ is the ML estimate of G (the variance covariance matrix of u), and $\hat{\sigma}$ is the ML estimate of σ. The range of the statistic P_{rand} is 0–1 under these model assumptions. The larger the P_{rand} the better the prediction and the smaller the random effect. The penalty for random effects in P_{rand} is analogous to Akaike's Information Criterion and Schwarz's BIC. Note that when the model has only fixed effects, P_{rand} is reduced to the traditional R^2 statistic.

Models in association mapping

When both population structure (Q) and kinship (K) are included, the mixed model for the Q + K method is

$$y = \mu + Qv + Zu + e \quad (3)$$

where y is a vector of phenotype observation, μ is a vector of intercepts; v is a $k \times 1$ vector of population effects; u is a $p \times 1$ vector of random polygene background effects; e is a vector of random experimental errors; Q is an $n \times k$ matrix defining the subgroup membership, generated from population structure analysis of marker data, and Z is an $n \times p$ incidence matrix relating y to u. For $\text{Var}(u) = G = 2KV_g$, K is a $p \times p$ matrix of kinship coefficients, and V_g (a scalar) is the unknown genetic variance, $E(e) = 0$ and $\text{Var}(e) = I\sigma^2$.

Likewise, we can define the Q model without the Zu term; the K model without the Qv term; the P model with P (that is eigenvectors) from PCA replacing Q but no Zu term; and the P + K model with P replacing Q (Table 2). These models represent different combinations of methods that account for complex genetic relationships in the association mapping population (Yu *et al*., 2006; Weber *et al*., 2007; Zhao *et al*., 2007).

Computer simulation

To assess the performance of these R^2 statistics in the context of mixed-model association mapping, we generated genetic populations with both gross level population structure and familial relationships within subpopulations. This allowed us to investigate mixed models with both fixed effects for population structure and random effects for relative kinship. Detailed simulation procedures have been described earlier (Zhu and Yu, 2009). Briefly, the β distribution (Balding and Nichols, 1995; Nicholson *et al*., 2002; Marchini *et al*., 2004) was used to model the correlated allele frequencies. Once allele frequencies of each locus for each subpopulation were sampled under the β model, conditionally on Hardy–Weinberg and linkage equilibrium, we mimicked

Table 2 Models used in the data analysis

Model	*Description*	*Model form*
Intercept	Intercept only for comparison	$y = \mu + e$
P	Regression model with fixed principal component covariates	$y = \mu + Pv + e$
K	Mixed model with random kinship	$y = \mu + Zu + e$
P+K	Mixed model fixed principal component covariates and random kinship	$y = \mu + Pv + Zu + e$
Q	Regression model with fixed population structure covariates	$y = \mu + Qv + e$
Q+K	Mixed model fixed population structure covariates and random kinship	$y = \mu + Qv + Zu + e$

different populations consisting of subpopulations. Specifically, we carried out simulations that mimicked two types of population used in association studies (Yu *et al*., 2006; Zhu and Yu, 2009): samples with both population structure and familial relatedness (type IV) and samples with severe population structure and familial relationship (type V). As with earlier extensive simulations (Zhu and Yu, 2009), the population size was 216, and three subpopulations were simulated for type IV and V samples. For each sample type, a total of 500 independent data sets were generated for analysis with three different models, and the various R^2 statistics were obtained. Samples in which the Hessian matrix or the covariance matrix of the random effects (seven for type IV and four for type V) were not positive semidefinite were removed.

To generate genotypes and phenotypes, a linkage map of 2000 cM composed of 10 chromosome segments, each 200 cM in length, was considered. An additive genetic model with no dominance or epistasis was used. Of the 2000 single nucleotide polymorphism (SNP) locations, 25 were chosen at random to be quantitative trait nucleotide (QTN) locations. In all simulations, we set each QTN genotypic value with genotype *QQ* as 0.5, genotype *qq* as 0, and the overall mean at 10. The overall genotypic value of an individual was obtained as the sum of genotypic values across all QTN plus the overall mean. An individual phenotype was generated as the genotypic value plus a random variable sampled from a standard normal distribution. Heritability for each QTN varied around 2%, depending on the allele frequency at each specific QTN.

To verify the general agreement between the R^2_{LR} statistic and the detection of true QTNs, we plotted the values of R^2_{LR} for all SNPs with the P + K model from a random run of type IV samples.

Empirical data analysis

Data from two association mapping populations were used for empirical data analysis. Genotypes and three phenotypes (that is flowering time, ear height, and ear diameter) were chosen from 277 maize strains across 553 SNP as described earlier (Liu *et al*., 2003; Flint-Garcia *et al*., 2005; Yu *et al*., 2006). The Q matrix was computed by STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003) and the K matrix by SPAGeDi (Hardy and Vekemans, 2002). The P matrix was computed from EIGENSTRAT (Price *et al*., 2006), and three PCAs were used to be consistent with the Q matrix for degree of freedom in the model-fitting process. *Arabidopsis* genotypes and phenotypes were obtained from a published data set with 5419 SNPs and two flowering time measurements (SDV and JIC8W) (Zhao *et al*., 2007). These two traits passed our trait screening process and yielded meaningful variance component estimates for mixed-model analysis. The Q matrix contains eight subgroups, and the P matrix contains the first eight PCAs (Zhao *et al*., 2007). For R^2_{LR}, we modified the Venn diagrams to depict the overlapping but complementary nature of Q and K in capturing genetic relationships. The modification was to make the size of the circle proportional to the R^2_{LR} value for easier interpretation of the diagram.

Results

All R^2 statistics (Table 1) yielded values between zero and one when different models were used to analyze data from two association mapping sample types, except the R^2_{W1} statistic (Table 3). Notably, the zero values for R^2_{W1} under the K model were not unexpected because its definition excludes random effects in calculating the predicted value. However, including the random term in prediction (R^2_{W2}) yields values comparable to those of other R^2 statistics.

When only the fixed effect was involved (that is P model), four R^2 statistics (that is R^2_{LR}, R^2_{W1}, R^2_{W2}, and P_{rand}) yielded identical values (Table 3). This was expected because theoretical derivation showed that all three definitions reduce to the original R^2 form for the fixed linear model. Meanwhile, the r_c statistic yielded different values for the fixed-effect model P because its formula does not reduce to R^2 for the fixed linear model.

Comparing an R^2 statistic among P, K, and P + K models showed differences between having a variable missing and having it added. Notably, R^2_{LR} for the model with added variables (P + K model) was consistently higher than for the model with fewer variables (P or K model) without exception, but this was not the case for other R^2 statistics (Table 3). Moreover, the standard deviation of R^2_{LR} was either equal to or smaller than that of other R^2 statistics. Also, the range of R^2 statistics was 0–1 except when the Hessian matrix or the

Table 3 Performance of R^2 statistics from different models under two association sample types

Sample type	R^2 *statistics*	*P+K*	*K*	*P*	*Count_1*[a]	*Count_2*[b]
IV	R^2_{LR}	0.617 (0.044)	0.578 (0.024)	0.488 (0.071)	0	0
	R^2_{W1}	0.211 (0.064)	0	0.488 (0.071)	0	468
	R^2_{W2}	0.524 (0.058)	0.509 (0.032)	0.488 (0.071)	19	9
	r_c	0.893 (0.083)	0.862 (0.085)	0.737 (0.091)	133	82
	P_{rand}	0.753 (0.072)	0.688 (0.079)	0.488 (0.071)	69	0
V	R^2_{LR}	0.705 (0.047)	0.598 (0.035)	0.680 (0.144)	0	0
	R^2_{W1}	0.333 (0.065)	0	0.680 (0.144)	0	452
	R^2_{W2}	0.594 (0.055)	0.504 (0.050)	0.680 (0.144)	0	411
	r_c	0.909 (0.082)	0.773 (0.077)	0.793 (0.169)	61	77
	P_{rand}	0.799 (0.156)	0.704 (0.089)	0.680 (0.144)	16	0

Numbers were calculated from 500 simulation runs.[2]tandard deviation is given in parenthesis.
[a]The number of times when the R^2 statistic for the K model is greater than for the P+K model.
[b]The number of times when the R^2 statistic for the P model is greater than for the P+K model.

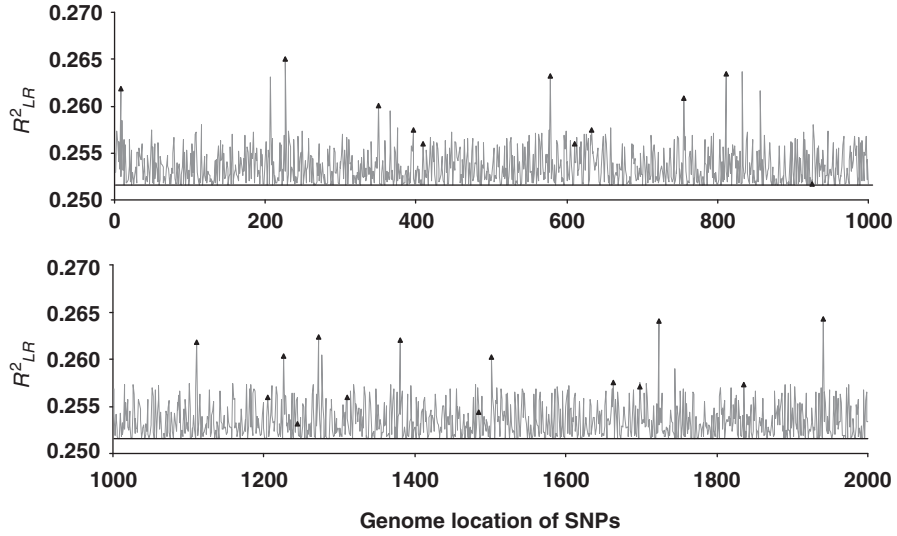

Figure 1 The R^2_{LR} values of the mixed model including each SNP across the genome. Triangles indicate the R^2_{LR} values and positions of the QTNs simulated and the straight line under the curve is the baseline R^2_{LR} value of the mixed model without SNP. Note that the way the computer simulation was carried out does not allow all QTNs to have a high R^2_{LR} value, mimicking the complex scenarios that are typical in association mapping studies.

Table 4 Analysis results of different R^2 statistics obtained by analyzing the maize traits with different models

R^2 *statistics*	*Flowering time*					*Ear height*					*Ear diameter*				
	Q	*Q+K*	*K*	*P+K*	*P*	*Q*	*Q+K*	*K*	*P+K*	*P*	*Q*	*Q+K*	*K*	*P+K*	*P*
R^2_{LR}	0.341	0.411	0.342	0.397	0.295	0.158	0.249	0.208	0.246	0.144	0.025	0.133	0.132	0.139	0.040
R^2_{W1}	0.341	0.118	0.000	0.090	0.295	0.158	0.054	0.000	0.050	0.144	0.025	0.001	0.000	0.008	0.040
R^2_{W2}	0.341	0.723	0.791	0.743	0.295	0.158	0.721	0.774	0.724	0.144	0.025	0.847	0.849	0.843	0.040
r_c	0.509	0.910	0.950	0.921	0.456	0.273	0.885	0.919	0.888	0.252	0.048	0.949	0.950	0.947	0.077
P_{rand}	0.341	0.812	0.867	0.828	0.295	0.158	0.815	0.860	0.818	0.144	0.025	0.835	0.837	0.832	0.040

Table 5 Analysis results of different R^2 statistics obtained by analyzing the *Arabidopsis* traits with different models

R^2 *statistics*	*SDV*					*JIC8W*				
	Q	*Q+K*	*K*	*P+K*	*P*	*Q*	*Q+K*	*K*	*P+K*	*P*
R^2_{LR}	0.369	0.404	0.172	0.474	0.473	0.528	0.538	0.085	0.565	0.564
R^2_{W1}	0.369	0.281	0.000	0.423	0.473	0.528	0.495	0.000	0.531	0.564
R^2_{W2}	0.369	0.315	−0.289[a]	0.516	0.473	0.528	0.610	0.175	0.627	0.564
r_c	0.539	0.738	0.580	0.706	0.643	0.691	0.803	0.448	0.786	0.721
P_{rand}	0.369	0.615	0.471	0.562	0.473	0.528	0.680	0.360	0.655	0.564

[a]As a result of $(y-\hat{y})'V^{-1}(y-\hat{y}) > (y-\bar{\bar{y}})'V^{-1}(y-\bar{\bar{y}})$.

covariance matrix of the random effects was not positive semidefinite, with the resulting negative value for P_{rand} removed in calculating the mean and standard deviation.

After determining the suitable candidate R^2 statistic for model comparison in mixed-model association mapping, we further demonstrated changes in R^2_{LR} as the QTNs and other SNPs across the genome entered the mixed model individually (Figure 1). To do this, we used a type IV association mapping sample. As expected, R^2_{LR} values with the SNP/QTN term were equal to or greater than the baseline R^2_{LR} value from the model without the SNP/QTN term. As the variation due to individual QTNs varied depending on allele frequency, not all QTNs yielded a high R^2_{LR} when their effects were included in the model. On the other hand, some SNPs can show a high R^2_{LR} even when they were not the causal loci, revealing the challenges faced in association mapping.

For the maize data, only R^2_{LR} consistently yielded a higher value for models with more variables (Q + K or P + K) than models with fewer variables (Q, K, or P) across three traits (Table 4). Next, for models with only fixed effects (that is Q or P), r_c values were different from the other four statistics, which agrees with the theoretical expectation and the simulation results. Furthermore, for *Arabidopsis* data, R^2_{LR}, r_c, and P_{rand} yielded a higher value for models with more variables, but this was not the case for R^2_{W1} or R^2_{W2} (Table 5).

In the modified Venn diagram, R^2_{LR} shows the overlap between the two methods in accounting for genetic

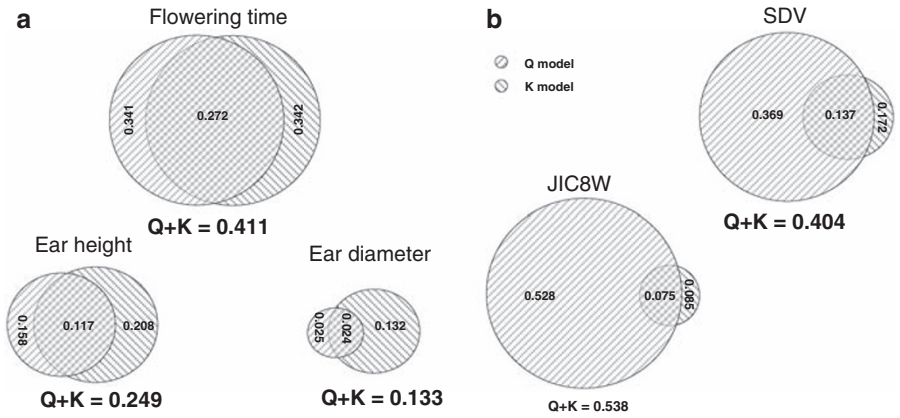

Figure 2 Modified Venn diagrams for R^2_{LR} values from different models obtained for (**a**) maize and (**b**) Arabidopsis traits. The number in each circle is the R^2_{LR} value of either the Q or K model, the R^2_{LR} value of the Q + K model is given under the jointed circles, and the number in the jointed area indicates the overlap between two complementary methods (i.e., Q and K) in controlling genetic relationship.

relationships: population structure (Q) captures general grouping patterns and relative kinship (K) is a polygene background control (Figure 2). The relative importance of Q and K in model fitting varied for different quantitative traits, which was expected given the theory (Tables 4 and 5). The complementary nature of P and K can also be seen in the modified Venn diagram. Obviously, the relative contribution of Q, P, and K to the mixed-model analysis varied across different data sets or different traits. For example, both Q and P made a small contribution in the analysis of maize ear diameter, but including K only improved the model fit by a negligible amount, as shown by a small increase in R^2_{LR}.

Discussion

Various R^2-like statistics for mixed models revealed the mixed perspectives on how the goodness-of-fit of the mixed models should be measured. For instance, the R^2_{LR} statistic, based on the *LR* test (Magee, 1990), considers the change of likelihood between models with different fixed and random effects simultaneously. However, the R^2_W statistic, based on the Wald statistic (Buse, 1973), measures the agreement between observations and the generalized least square predictors without considering random effects. The modified form, R^2_{W2}, which considers both random and fixed effects, would be a better choice than R^2_{W1} for analyzing genetic relationships but needs further study. Next, the r_c statistic, based on the concordance correlation (Vonesh *et al*., 1996), indicates agreement between observations and the unweighted predicted values with both fixed and random effects, whereas the P_{rand} statistic, based on the penalized quasi-likelihood function (Zheng, 2000), measures the proportional reduction in penalized quasi-likelihood function. When only fixed effects are included in the model, three R^2 statistics, but not r_c, reduce to the simple form for fixed linear models. By definition, all R^2 statistics other than R^2_{W1} would be suitable for genomic mapping with different fixed and random terms controlling genetic relationships. The zero value of R^2_{W1} for the K model prevents its use in mixed-model association analysis. In comparing R^2_{LR} and R^2_{W2} for mixed-model analysis of a randomized complete block design and a design with spatially autocorrelated residuals (Kramer, 2005), the R^2 values of these two statistics increased when random effects were added to the model or when the correlated error structure was considered.

As the direct summation of sum of square of model and sum of square of residual to equal sum of square of the corrected total does not necessarily exist in generalized linear mixed models, the term 'Pseudo-R^2' was suggested to differentiate the above proposed statistics from the classical R^2 (Schabenberger and Pierce, 2002). We, however, adopted the general definition of the R^2 statistic (Buse, 1973; Magee, 1990; Nagelkerke, 1991), rather than the specific definition for a fixed linear model, in the text. Here, we stress that the 'proportion of variation explained' in linear mixed models should not be interpreted to mean that there is always an exact summation. In this study, we focused on comparing four different R^2 statistics for their potential in mixed-model association mapping. All these statistics contain similar components, involving differences between the observed values and the predicted values (either directly in R^2_W, r_c and P_{rand} or indirectly in R^2_{LR}). In particular, the R^2_{LR} statistic has several appealing properties (Nagelkerke, 1991). First, it reduces to the classical R^2 for fixed models and is asymptotically independent of the sample size. Second, it is dimensionless and permits an interpretation based on proportion of variation explained. Furthermore, using R^2_{LR}, to compare models with the same random components (that is K with Q + K or P + K) can be interpreted as comparing the fit of various nested models. On the other hand, comparing models with different fixed and random components provides a measure of model-data agreement under the ML framework, which satisfies a criterion proposed earlier: R^2 values for different models fitting the same data should be directly comparable (Kvalseth, 1985).

Ultimately, because it is easily computed and its monotonic nondecreasing property, R^2_{LR} is our choice to measure the goodness-of-fit of the model to the data. Expanding the mixed model to include other genetic and nongenetic factors should not complicate the calculation and interpretation of R^2_{LR} because it is directly computed from the maximum log-likelihood of the full model and the reduced model. In simulation studies, an R^2 measure computed as the squared correlation between simulated and model predicted genetic values may be used (Piepho

and Möhring, 2007). Other R^2 statistics based on the ratio of variance component for residuals between two models have also been proposed (Xu, 2003). A recent study, however, found that these latter statistics performed poorly because the R^2 values varied so little that identifying the most parsimonious model was difficult (Oreliena and Edwards, 2008). Extending R^2_{LR} to the REML approach needs further study because comparing models with different fixed or random terms is only valid under the ML framework (Littell *et al*., 2006). The relationship between model fit and model selection, particularly in genomic mapping, is beyond the scope of this study (Broman and Speed, 2002; Sillanpaa and Corander, 2002; Yi *et al*., 2005). We have no intention of using R^2_{LR} to conduct model selection because the monotonic nondecreasing property of R^2_{LR} does not indicate a better model as additional fixed or random effects are added. Instead, we stress that the R^2_{LR} statistic provides an additional measurement for results interpretation.

For mixed models with random components (K, P + K, or Q + K), variance component estimation was conducted independently before the solutions for mixed models were used to compute different R^2 statistics. On the basis of the definition of R^2_{LR}, the convergence process of ML of a model containing additional effects other than intercept and residual can also be viewed as a process to maximize R^2_{LR} but not the other R^2 statistics. Clearly, R^2_{LR} can quantify the goodness-of-fit of different models regardless of the statistical properties of the models (Cameron and Windmeijer, 1996). In an earlier study, we showed that the likelihood-based model-fitting approach can quantify the robustness of genetic relationships derived from molecular marker data (Yu *et al*., 2009). Essentially, kinship construction with subsets of the whole marker panel and subsequent model testing with multiple phenotypic traits can be viewed as a process to test the model-data fit of different variance–covariance matrices. With an adequate number of molecular markers, an accurate genetic relationship among individuals (that is variance covariance matrices) can be obtained, and the change in the value of R^2_{LR} becomes minimal.

Comparing the values of R^2_{LR} for Q, K, and Q + K, as shown with modified Venn diagrams, can help us understand the genetics behind two overlapping methods in accounting for genetic relationships. With complex genetic relationships among individuals in many association mapping panels (Meuwissen *et al*., 2002; Yu *et al*., 2006; Zhao *et al*., 2007; Zhu and Yu, 2009), various competing but mostly complementary methods to capture these relationships were developed. Thus, the contribution to the model-data agreement from either Q and P (population structure and PCA) or K (kinship) can be determined from the R^2_{LR} when each is fitted alone. Next, the overall contribution and overlap can be shown by comparing the R^2_{LR} values of Q + K (or P + K) with the values from models with individual components (that is Q, P, or K). Finally, although it is not a statistic with a significance test, R^2_{LR} does provide an indication of a variable's importance in model fitting, for example, SNP, Q, P, or K (Kvalseth, 1985). With an established base model (Yu *et al*., 2006), the changes in R^2_{LR} values resulted from adding individual molecular marker provide information on the relative importance of different markers in further explaining the total variation.

In summary, we demonstrated through simulated association mapping samples and empirical data analyses that the *LR*-based R^2 statistic has several desirable properties useful in mixed-model association mapping. Applying genomic technologies in complex trait dissection has generated vast amounts of data, the analysis of which requires a joint effort in genetics and statistics. There are many challenges in this multidisciplinary research (Hirschhorn and Daly, 2005; Weir *et al*., 2006; McCarthy *et al*., 2008; Zhu *et al*., 2008), but such research also provides great opportunities for further collaboration among researchers from different disciplines with different specialties.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This project is supported by the National Research Initiative (NRI) Plant Genome Program of the USDA Cooperative State Research, Education and Extension Service (CSREES) (2006-03578), the National Science Foundation (DBI-0820610), and the Targeted Excellence Program of Kansas State University. Hans-Peter Piepho is supported by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr 'Synbreed—Synergistic plant and animal breeding'.

References

Balding DJ, Nichols RA (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**: 3–12.

Broman KW, Speed TR (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J R Stat Soc B* **64**: 641–656.

Buse A (1973). Goodness of fit in generalized least-squares estimation. *Am Stat* **27**: 106–108.

Cameron AC, Windmeijer FAG (1996). R-squared measures for count data regression models with applications to health-care utilization. *J Bus Econ Stat* **14**: 209–220.

Cox DR, Snell EJ (1989). *Analysis of Binary Data*, 2nd edn. Chapman and Hall: London.

Draper NR, Smith H (1981). *Applied Regression Analysis*, 2nd edn. John Wiley & Sons: New York, NY.

Everitt BS (2002). *Cambridge Dictionary of Statistics*, 2nd edn. Cambridge University Press: Cambridge, UK.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE *et al.* (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* **44**: 1054–1064.

Hardy OJ, Vekemans X (2002). SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. *Mol Eco Notes* **2**: 618–620.

Henderson CR (1984). *Application of Linear Models in Animal Breeding*. University of Guelph: Ontario.

Hirschhorn JN, Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95–108.

Kennedy BW, Quinton M, van Arendonk JA (1992). Estimation of effects of single genes on quantitative traits. *J Anim Sci* **70**: 2000–2012.

Kramer M (2005). R^2 statistics for mixed models. 2005 Proceedings of the Conference on Applied Statistics in Agriculture, Manhattan, KS, pp 148–160.
Kvalseth TO (1985). Cautionary note about R^2. *Am Stat* **39**: 279–285.
Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S *et al.* (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40**: 584–591.
Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2006). *SAS for Mixed Models*, 2nd edn. SAS Press: Cary, NC, USA.
Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003). Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* **165**: 2117–2128.
Loiselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). *Am J Bot* **82**: 1420–1425.
Lynch M, Walsh JB (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc.: Sunderland, MA.
Maddala GS (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press: Cambridge, UK.
Magee L (1990). R2 measures based on Wald and likelihood ratio joint significance tests. *Am Stat* **44**: 250–253.
Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. *Genetics* **175**: 879–889.
Marchini J, Cardon LR, Phillips MS, Donnelly P (2004). The effects of human population structure on large genetic association studies. *Nat Genet* **36**: 512–517.
McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP *et al.* (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**: 356–369.
Meuwissen TH, Karlsen A, Lien S, Olsaker I, Goddard ME (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**: 373–379.
Nagelkerke NJD (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**: 691–692.
Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánssonand K, Donnelly P (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J R Stat Soc B* **64**: 695–715.
Oreliena JG, Edwards LJ (2008). Fixed-effect variable selection in linear mixed models using R^2 statistics. *Comput Stat Data Anal* **52**: 1896–1907.
Piepho HP, Möhring J (2007). Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**: 1881–1888.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.
Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
Ritland K (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res* **67**: 175–186.
Schabenberger O, Pierce FJ (2002). *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press: Boca Raton, FL.
Schwarz G (1978). Estimating dimension of a model. *Ann Stat* **6**: 461–464.
Sillanpaa MJ, Corander J (2002). Model choice in gene mapping: what and why. *Trends Genet* **18**: 301–307.
Vonesh EF, Chinchilli VM (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measures*. Marcel Dekker: New York.
Vonesh EF, Chinchilli VM, Pu K (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics* **52**: 572–587.
Weber A, Clark RM, Vaughn L, Sanchez-Gonzalez Jde J, Yu J, Yandell BS *et al.* (2007). Major regulatory genes in maize contribute to standing variation in teosinte (Zea mays ssp. parviglumis). *Genetics* **177**: 2349–2359.
Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M *et al.* (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40**: 575–583.
Weir BS, Anderson AD, Hepler AB (2006). Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* **7**: 771–780.
Xu R (2003). Measuring explained variation in linear mixed effects models. *Stat Med* **22**: 3527–3541.
Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170**: 1333–1344.
Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.
Yu J, Zhang Z, Zhu C, Tabanao DA, Pressoir G, Tuinstra MR *et al.* (2009). Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. *Plant Genome* **2**: 63–77.
Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C *et al.* (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**: e4.
Zheng B (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Stat Med* **19**: 1265–1275.
Zhu C, Gore MA, Buckler ES, Yu J (2008). Status and prospects of association mapping in plants. *Plant Genome* **1**: 5–20.
Zhu C, Yu J (2009). Nonmetric multidimensional scaling corrects for population structure in whole genome association studies. *Genetics* **182**: 875–888.