nature genetics

A unified mixed-model method for association mapping that accounts for multiple levels of relatedness

Jianming Yu[1,9], Gael Pressoir[1,9], William H Briggs[2], Irie Vroh Bi[1], Masanori Yamasaki[3], John F Doebley[2], Michael D McMullen[3,4], Brandon S Gaut[5], Dahlia M Nielsen[6], James B Holland[4,7], Stephen Kresovich[1,8] & Edward S Buckler[1,4,8]

As population structure can result in spurious associations, it has constrained the use of association studies in human and plant genetics. Association mapping, however, holds great promise if true signals of functional association can be separated from the vast number of false signals generated by population structure[1,2]. We have developed a unified mixed-model approach to account for multiple levels of relatedness simultaneously as detected by random genetic markers. We applied this new approach to two samples: a family-based sample of 14 human families, for quantitative gene expression dissection, and a sample of 277 diverse maize inbred lines with complex familial relationships and population structure, for quantitative trait dissection. Our method demonstrates improved control of both type I and type II error rates over other methods. As this new method crosses the boundary between family-based and structured association samples, it provides a powerful complement to currently available methods for association mapping.

Population structure is universal among organisms[3,4]. It can arise naturally in the form of herds, colonies, ethnic groups or other types of aggregations, owing to geography, natural selection or artificial selection. For association mapping, a given sample may fall into one of five categories defined by population structure associated with local adaptation or diversifying selection and familial relatedness from recent coancestry (**Fig. 1**). Ideally, samples with minimal population structure or familial relatedness (area I) result in the greatest statistical power, provided that the trait of interest is well distributed (**Fig. 1**). Such samples, however, often prove very difficult to collect, are small in size and/or have a narrow genetic basis. Family-based samples (area II) have been exploited to avoid the effect of population structure[5–8], but these samples are also limited by sample size and allelic diversity and can be difficult to collect, particularly for late-onset human diseases (**Fig. 1**). For quantitative traits, the Quantitative Transmission Disequilibrium Test (QTDT) is one method widely used for association mapping with these family-based samples[7]. Samples of increased size as well as broader allelic diversity in a species often contain population structure (area IV) or include familial relationships within structured population (area III; **Fig. 1**). For these samples, Structured

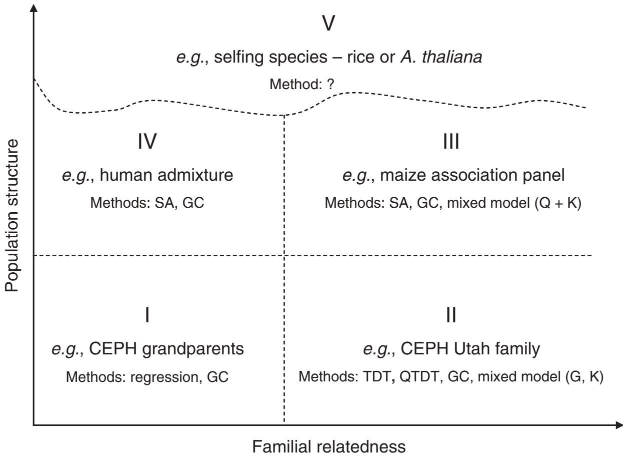

Figure 1 Different types of samples used for association mapping. Although all individuals are related via a large genealogical tree, the population structure axis depicts relationships among major subpopulations associated with local adaptation or diversifying selection. The familial relatedness axis depicts the relationships among individuals from recent coancestry. The dotted lines indicate that clear delineations may not exist between areas. Methods listed can be applied to different sample types, although they are designed for specific purposes (for example, SA was designed for samples with population structure, and TDT and QTDT were designed for family-based samples). These methods deal with the specific sample structure either directly (as in SA, QTDT, TDT and the mixed model) or indirectly by adjusting the test statistics from regression analysis (as in GC).

[1]Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, USA. [2]Department of Genetics, University of Wisconsin, Madison, Wisconsin 53706, USA. [3]Division of Plant Sciences, University of Missouri, Columbia, Missouri 65211, USA. [4]United States Department of Agriculture-Agricultural Research Service (USDA-ARS). [5]Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697, USA. [6]Department of Statistics and [7]Department of Crop Science, North Carolina State University, Raleigh, North Carolina 27695, USA. [8]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to E.S.B. (esb33@cornell.edu).

Received 7 July; accepted 13 October; published online 25 December 2005; doi:10.1038/ng1702

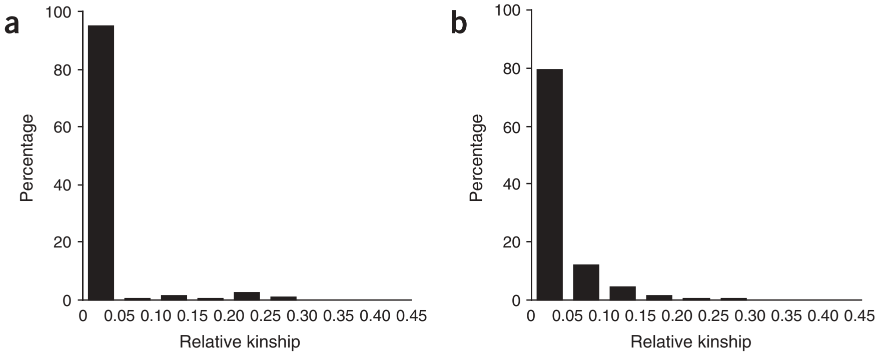

Figure 2 Distribution of pairwise relative kinship estimates in the CEPH sample and maize sample. The peaks around 0 indicate no relationship. (**a**) For the CEPH sample, the small peaks around 0.25 represent relationships among offspring within a family, parent-offspring or grandparent-parent, and the small peak around 0.125 represent relationships of grandparent-offspring. (**b**) For the maize sample, the continuously descending estimates of relative kinship agree with the complex familial relationships and population structure among these maize inbred lines. For simplicity, we show only percentages of the relative kinship estimates of the maize sample from 0 to 0.45.

Association (SA) and Genomic Control (GC) are common methods used in human and plant studies to control for the false positives (type I errors) caused by this population structure[9–13]. With GC, random markers are used to estimate and adjust the inflation of test statistics generated by population structure, assuming such a structure has a similar effect on all loci. SA analysis uses random markers to estimate population structure and then incorporates this into further statistical analysis. For samples in area III, however, accounting only for population structure may lead to either inadequate control for false positives or a loss in power owing to familial relatedness (**Fig. 1**). It remains to be seen which methods will prove most useful when evaluating samples with very high levels of population structure along with diverse levels of familial relatedness (area V; **Fig. 1**).

In this paper, we present a new method for association mapping that is applicable to samples from areas II and III (**Fig. 1**). For the family-based sample, we applied our method to microarray data from the baseline expression levels of genes in immortalized B cells from 14 families of Centre d'Etude du Polymorphisme Humain (CEPH) Utah pedigree[14]. In this study, six gene expression phenotypes were regarded as phenotypic traits in mapping expression quantitative trait loci (eQTL). For the sample containing complex familial relationships and population structure, we examined three quantitative traits measured on 277 diverse maize inbred lines, representing the diversity present in public breeding programs around the world[15]. As maize is a highly outbred species, the population differentiation (F_{st}) among the major subgroups in our sample ranged from 0.047 (SSR) to 0.073 (SNP), similar to that of a recent human study[11] ($F_{st} = 0.013$ for Chinese-Japanese and $F_{st} = 0.145$ for between continents). When we included a minor bottlenecked subgroup, the overall F_{st} rose to 0.106 (SSR) and 0.118 (SNP).

Our association mapping approach integrates genomic tools to uncover population structure and familial relationships with the traditional mixed-model framework that has long been used by animal geneticists[16–18]. One obvious obstacle in applying a mixed-model method beyond a few domesticated animal species is that pedigree records are often unknown or inaccurate. Genomic tools now allow us to detect both population structure (Q) and relative kinship (K) within a sample. Marker-based relative kinship estimates have proven useful for quantitative inheritance studies in different populations[19,20]. This K estimate approximates identity by descent by adjusting the probability of identity by state between two individuals with the average probability of identity by state between random individuals. For the CEPH sample, we replaced the pedigree-based coancestry matrix (G) in a traditional mixed model with the K matrix to define the degree of genetic covariance among individuals. No population structure was detected, and Q was not included in the mixed-model analyses. For the maize sample, we fit both Q and K into the mixed model to account for multiple levels of relatedness.

We randomly tested the expression level of six genes with good heritability estimates: *HSD17B12*, *TUBB2A*, *CTSH*, *RPS26*, *UBE2L3* and *SSR1*. The K matrix agreed with the family structure of the data. Although 94% of the pair-wise kinship estimates were close to 0, the small peaks around 0.25 and 0.125 represented the relationships within families (**Fig. 2a**). The mixed model with either the K or G matrix fits the data equally well. The K model showed a significant improvement in model fit over the simple model in which family structure is ignored (**Table 1**).

Overall, the K model, GC and QTDT showed good control of type I error rate (**Fig. 3a–c** and **Supplementary Fig. 1** online). The K model gave slightly liberal results and GC gave slightly conservative results for *TUBB2A*, whereas the results from QTDT were liberal for *TUBB2A* but slightly conservative for *HSD17B12*. The statistical power simulation was conducted by adding a genetic effect to each marker and then testing whether it could be detected by different models. The adjusted

Table 1 Goodness of fit of three different models in explaining phenotypic variation of human gene expressions

	HDS17B12		*TUBB2A*		*CTSH*		*RPS26*		*UBE2L3*		*SSR1*	
	–2 log likelihood	BIC	–2 log likelihood	BIC	–2 log likelihood	BIC	–2 log likelihood	BIC	–2 log likelihood	BIC	–2 log likelihood	BIC
Simple model	116.1[a]	126.6	210.3[a]	220.8	208.1[a]	218.6	184.9[a]	195.4	72.3[a]	82.8	14.0[a]	24.6
Coancestry	70.0	85.8	174.9	190.7	137.7	153.5	56.5	72.3	43.6	59.4	–3.2	12.6
K model	67.8	83.6	173.9	189.7	140.7	156.5	58.7	74.5	45.0	60.8	–3.2	12.6

[a]Model comparison based on χ^2 test indicates whether the K model significantly improves the model fit at $P < 0.001$; BIC, Bayesian Information Criterion (smaller is better); all CEPH Utah family members were used in the analyses of three models; the simple model was included only for the purpose of illustrating the effect of ignoring family relationships, as it is not a standard practice.

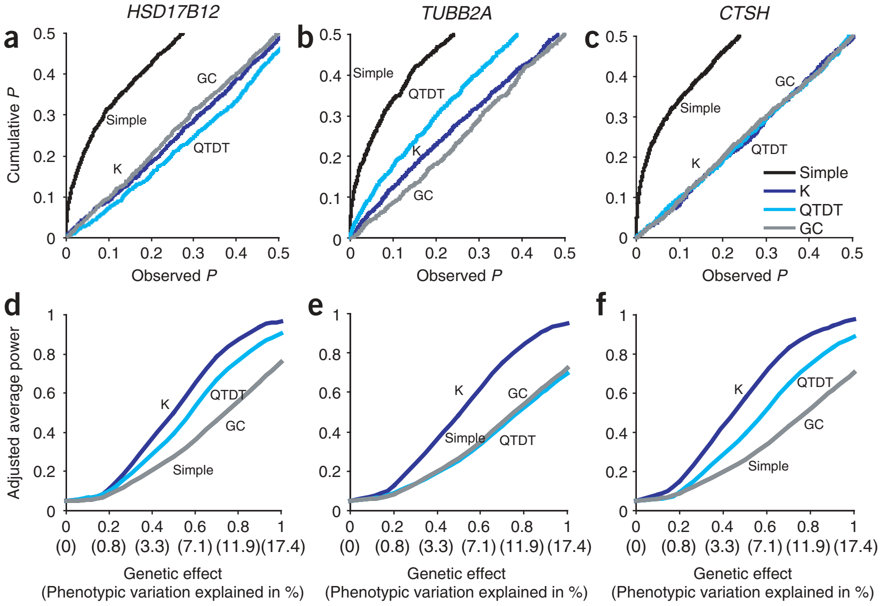

Figure 3 Model comparison with human gene expression phenotypes. (**a**–**c**) Evaluation of the model type I error rates using random SNPs for gene expression phenotypes for *HSD17B12* (**a**), *TUBB2A* (**b**) and *CTSH* (**c**). The cumulative distributions of observed *P* values are presented for the simple model, the K model, QTDT and the simple model with genomic control (GC). Under the expectation that random SNPs are unlinked to the polymorphisms controlling these traits (H_0: no SNP effect), approaches that appropriately control for type I errors should show a uniform distribution of *P* values (a diagonal line in these cumulative plots). The simple model was included only for the purpose of illustrating the effect of ignoring family relationships, as it is not a standard practice. (**d**–**f**) The adjusted average power of the models for *HSD17B12* (**d**), *TUBB2A* (**e**) and *CTSH* (**f**). A genetic effect was added to each random SNP (QTN effect), where $k = 0.1$, 0.2, 0.5, 0.7, 0.9 and 1.0 times the standard deviation of the phenotypic mean of a trait. Each model was adjusted based on its empirical type I error rate. The adjusted average power for GC is the same as that of the simple model with the empirical threshold *P* value. For convenience of comparison, we list the point value of phenotypic variation explained by a QTN at the allele frequency of $p = 0.3$.

average power was consistently higher for the K model than for QTDT, GC and the simple model (**Fig. 3d–f** and **Supplementary Fig. 1**). Differences in the adjusted average power between the K model and QTDT were smaller for *HSD17B12*, *CTSH* and *RPS26* than for *TUBB2A*, *UBE2L3* and *SSR1*. A previous study has shown that the gene expression levels of *HSD17B12*, *RPS26* and *CTSH* are regulated by *cis*-acting determinants[14]. The heritability estimates for *HSD17B12* ($h^2 = 0.58$), *CTSH* (0.64) and *RPS26* (0.80) gene expression were higher than those for *TUBB2A* (0.34), *UBE2L3* (0.52) and *SSR1* (0.21).

Table 2 Goodness of fit of four different models in explaining phenotypic variation of maize quantitative traits

	Flowering time		Ear height		Ear diameter	
	−2 log likelihood	BIC	−2 log likelihood	BIC	−2 log likelihood	BIC
Simple model	1,655.6[a]	1,666.7	2,332.3[a]	2,343.4	1,306.0[a]	1,316.9
Q model	1,546.1[a]	1,568.4	2,286.9[a]	2,309.2	1,300.1[a]	1,321.9
K model	1,545.9[a]	1,562.6	2,270.6[a]	2,287.3	1,272.8 (NS)	1,289.2
Q + K model	1,516.7	1,544.6	2,256.6	2,284.5	1,272.5	1,299.8

[a]Model comparison based on χ^2 test indicates whether the Q + K model significantly improves the model fit at $P < 0.001$; NS, not significant; BIC, Bayesian Information Criterion (smaller is better); the simple model was included only for the purpose of illustrating the effect of ignoring population structure and family relationships, as it is not a standard practice.

However, because our test of statistical power assumed complete linkage disequilibrium (LD) between markers and Quantitative Trait Nucleotides (QTN), whereas QTDT was designed to simultaneously test linkage and association, QTDT may be advantageous in a situation of low to moderate LD.

For the maize sample, although 80% of the pairwise kinship estimates were close to 0, the remaining estimates were distributed from 0.05 to 1.0, as expected from the complex familial relationships and population structure (**Fig. 2b**). The main difference in kinship estimates between the human and maize samples is that the maize sample had more first cousin–level relationships. In most cases, the Q + K model showed a significant improvement in goodness of fit compared with the other models (**Table 2**).

For all three maize traits, the Q + K model resulted in the best approximation to the expected cumulative distribution of *P* values, followed by the K model, the Q model and, lastly, the simple model (**Fig. 4a–c**). In general, GC performed well, except for ear diameter. As expected, correction for deviation from the uniform distribution of *P* values by fitting Q in the model (that is, SA alone) was greatest for flowering time, followed by ear height and, finally, ear diameter. Correction by the K model was always better than the Q model. We found that 37.6% of the SNPs were associated with flowering time at $P < 0.05$ by the simple model, compared with 14.1% by the Q model, 6.1% by the K model and only 6.0% by the Q + K model. For all three traits, the models with Q- or K-based control had higher power than did the simple model and GC (**Fig. 4d–f**). For flowering time and ear height, the Q + K model had the highest power. For ear diameter, the K model yielded a slightly higher power than the Q + K model did, which agreed with our model fitting results.

As noted above, our new approach uses a relative kinship matrix estimated from marker data. As such, it is able to overcome the limitations of previous association studies in plants and many other organisms, where direct calculation of coancestry coefficients proved impractical owing to incomplete pedigree records or inaccurate owing to biases resulting from inbreeding, selection and drift. These biases can be especially strong in plant and animal breeding programs. Although the K model provided similar results as the G model in the CEPH sample, it would likely outperform the latter when errors in pedigree records, self-reported ancestry or segregation distortions exist.

A second benefit of our approach is that the Q + K model is able to systematically account for multiple levels of relatedness among individuals. Essentially, the genetic consequence of local adaptation or diversifying selection among the different maize populations—

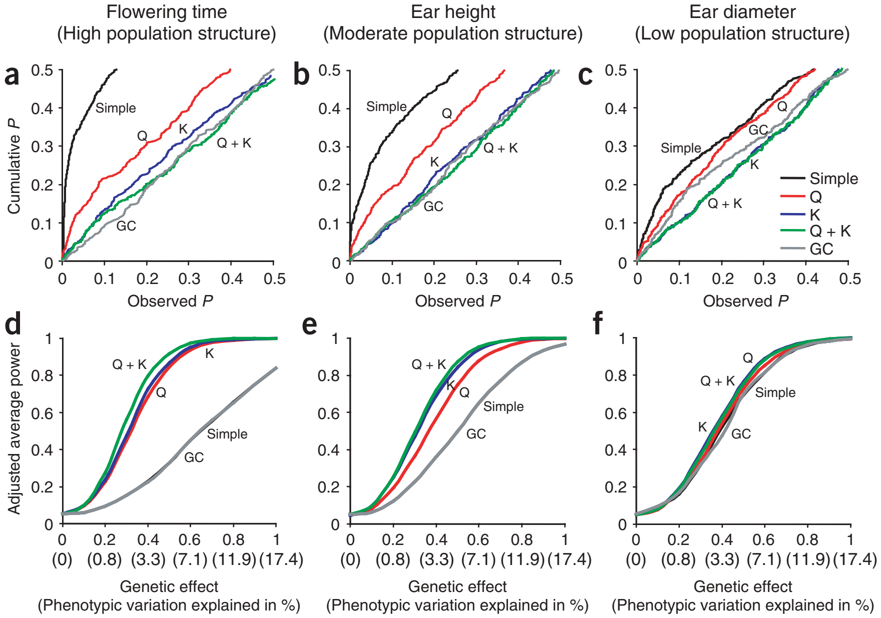

Figure 4 Model comparison with maize quantitative traits. (**a**–**c**) Evaluation of the model type I error rates using random SNPs for flowering time (**a**), ear height (**b**) and ear diameter (**c**). The cumulative distributions of observed *P* values are presented for the simple model, the Q model, the K model, the Q + K model and the simple model with genomic control (GC). Under the expectation that random SNPs are unlinked to the polymorphisms controlling these traits (H_0: no SNP effect), approaches that appropriately control for type I errors should show a uniform distribution of *P* values (a diagonal line in these cumulative plots). The simple model was included only for the purpose of illustrating the effect of ignoring population structure and family relationships, as it is not a standard practice. (**d**–**f**) The adjusted average power of the models for flowering time (**d**), ear height (**e**) and ear diameter (**f**). A genetic effect was added to each random SNP (QTN effect), where $k = 0.1$, 0.2, 0.5, 0.7, 0.9 and 1.0 times the standard deviation of the phenotypic mean of a trait. Each model was adjusted based on its empirical type I error rate. The adjusted average power for GC is the same as that of the simple model with the empirical threshold *P* value. For convenience of comparison, we list the point value of phenotypic variation explained by a QTN at the allele frequency of $p = 0.3$.

arguably the product of allelic differences at a few genes with relatively large phenotypic effects—was accounted for by Q in a gross manner, whereas relatedness among individuals within and between groups was accounted for by K on a finer scale. Thus, the two approaches for uncovering population structure are complementary. Unlike GC, which uses constant inflation factors for test statistics, the mixed model directly adjusts each test statistic internally by accounting for multiple levels of relatedness[17]. Consequently, our approach simultaneously improves the detection of QTN and the estimation of their effects and results in improved control for both type I and type II error rates.

The mixed-model approach has the added advantage of being flexible, as it can be applied to both family-based and population-based samples[21]. For samples such as CEPH without population structure, dropping Q reduces the model to a single population-based association analysis with polygene control. If, however, random mating within each subpopulation can be safely assumed or determined through the model fitting, dropping K reduces the model to a regression-based structured association analysis. The fixed SNP effect can be replaced with either a fixed or a random haplotype effect[18,22]. The robustness of our new method in withstanding inaccurate or insufficiently determined population structure and relative kinship estimates would be population dependent or sample dependent and requires further investigation. As in other mixed-model and variance components applications, small sample size may hinder an accurate and meaningful estimate of the polygenic component[16,17]. This, however, should not be problematic for association studies similar to the examples outlined here, in which marker number and sample size are large enough to obtain accurate estimates of both Q and K and the polygene component. Building on previous research in human[5–10], animal[16–18] and plant[13] systems, our new method initiates a systematic approach to future applications of association mapping for samples with multiple levels of relatedness in many species.

METHODS

CEPH sample. The CEPH sample comprised members of 14 CEPH Utah families (CEPH 1333, 1340, 1341, 1345, 1346, 1347, 1362, 1408, 1416, 1418, 1421, 1423, 1424 and 1454)[14]. A total of 1,384 autosomal SNP markers with a minimum frequency of 0.1 in 194 individuals, whose lymphoblastoid cells were phenotyped for gene expression, were used in the current analysis. Gene expression data was $\log_2$ transformed. Details of the genotyping and phenotyping procedure can be found at the SNP Consortium Linkage Map Project database.

Maize sample. We recently assembled a maize association panel with 277 inbred lines[15]. These lines generally belong to one of four groups recognized by plant breeders: stiff stalk, non–stiff stalk, tropical/subtropical and mixed. Field tests were conducted in Clayton, North Carolina, USA (summer nursery) and Homestead, Florida, USA (winter nursery) in 2002, and the trait mean of the two field tests was used in the current study. Three traits on which population structure has different levels of effect were selected from among the 60 traits measured: (i) flowering time, which is strongly correlated to population structure ($R^2 = 0.35$), (ii) ear height, which is moderately correlated to population structure ($R^2 = 0.16$) and (iii) ear diameter, which has no correlation to population structure ($R^2 = 0.02$)[15]. Flowering time was measured as the number of days to pollen shed, ear height as the distance from the ground to the major ear-bearing node and ear diameter as the diameter of an ear at the midsection.

SNP discovery was performed using a diverse set of 14 maize inbreds and 16 teosinte (*Zea mays* ssp. *parviglumis*) inbreds[23]. We used 553 SNPs with a minimum frequency of 0.1 in the 277 maize inbreds for the current analysis. They were designed from 413 randomly selected genes out of the ~10,000 maize ESTs in the MMP-DuPont set[24]. SNP assay development and scoring was performed by Genaissance Pharmaceuticals using the Sequenom MassARRAY System[25]. Replicated assays indicate that the genotyping error rate is ~0.3%.

Candidate QTN simulation. The genetic effect (a) of a simulated candidate QTN was assigned as $k = 0.1$, 0.2, 0.5, 0.7, 0.9 and 1.0 times the standard deviation of the phenotypic mean. The percentage (π) of the total phenotypic variation explained by this genetic effect can be estimated as

$$\pi = p(1-p)k^2/(p(1-p)k^2 + 1 - 1/n) \approx 1/(1+1/(p(1-p)k^2) \qquad (1)$$

where n is the sample size, and p is the sample frequency of the polymorphism at the QTN[26]. At 0.5 times the standard deviation, this genetic effect explains 2–6% of the phenotypic variation, depending on p at the QTN. For convenience of comparison, we listed the percentage of phenotypic variation explained by a QTN at the allele frequency of $p = 0.3$.

The genetic effect was assigned to all random SNPs, one at a time, and each model was run to determine whether the effect could be detected with the empirical threshold P value. The proportion of the detected QTN summed across all SNPs, or adjusted average power, was used as a measurement of the control on the type II error rate by each model.

Statistical tests. For the CEPH data with all 194 individuals from 14 families, Merlin[27] software was first used to estimate identity by descent with all SNPs on each chromosome, and this information was then used in QTDT[7] software to implement association tests for the 1,384 SNPs. STRUCTURE[9] did not detect any population structure other than known family structure, which agreed with results from previous study[28]. Heritability for each gene expression was calculated with Merlin.

For the maize data, STRUCTURE with 89 microsatellites showed that the likelihood for model parameter $k = 3$ was much higher than $k = 2$ and comparable with $k = 4$ or higher[15]. Combining this with knowledge of the breeding history of these inbred lines, we chose $k = 3$ and defined a mixed group to include lines with all three genome percentages under 0.8. The relative kinship (K) matrix was calculated on the basis of 553 SNPs using the software package SPAGeDi[29]. Negative values between individuals were set to 0, as this indicates that they are less related than random individuals[29]. Essentially, the degree of genetic covariance caused by polygenic effects was defined to be 0 for a pair of individuals that are not related but positive for a pair of individuals that are related. This threshold is similar to the pedigree-based coancestry matrix in which individuals with unknown relationship are set to 0. We replaced the pedigree-based coancestry matrix of the traditional mixed model with this K matrix to define the degree of genetic covariance between pairs of individuals. Model comparison of K matrix with other marker-based genetic similarity matrices, simple matching coefficients, Jaccard similarity coefficients and Dice coefficients indicated a better fit of the K matrix on the basis of Bayesian Information Criterion (BIC) values. We also experimented with different thresholds; these tests suggested that a threshold is needed, although the current approach may not always be the optimal solution for every population. We also implemented STRUCTURE analyses with SNP data and SPAGeDi with microsatellite data and obtained similar Q and K. All models with Q and K calculated based on either SNPs or microsatellites were tested for model fit, and the results consistently showed a better fit for the Q + K model over either Q or K alone. For consistency with previously published work[15], our results for Q were calculated on the basis of microsatellites and for K on the basis of SNPs.

We programmed a macro in SAS[30] to iteratively analyze the data with Proc Mixed. To compare the goodness of fit of the different models, we used maximum likelihood methods to obtain the –2 residual log likelihood and BIC, which accounts for the number of parameters in the model, because both fixed and random effects were involved in the model comparison. The test of SNP/QTN effect was carried out by F test, with denominator degrees of freedom per the Satterthwaite method, after the convergence of REML. Because LD decays rapidly with these samples, few, if any, of the random SNP should associate with the traits. Consequently, these random SNPs provided an empirical null distribution to compare these methods on the control of type I error. With this empirical distribution, an inflation factor (λ) was estimated, and genomic control for the simple model was conducted by dividing the F test statistic by this inflation factor[10].

The mixed model equation for our Q + K method is expressed as

$$\mathbf{y} = X\boldsymbol{\beta} + S\boldsymbol{\alpha} + Q\mathbf{v} + Z\mathbf{u} + \mathbf{e} \qquad (2)$$

Equation (2) is an expanded version of a traditional mixed model in which all fixed effects are modeled in the $X\boldsymbol{\beta}$ term[16]. In our expanded form, $X\boldsymbol{\beta}$ simply represents those fixed effects other than the SNP under testing and the population structure; $\mathbf{y}$ is a vector of phenotypic observation; $\boldsymbol{\beta}$ is a vector of fixed effects other than SNP or population group effects; $\boldsymbol{\alpha}$ is a vector of SNP effects (QTN); $\mathbf{v}$ is a vector of population effects; $\mathbf{u}$ is a vector of polygene background effects; $\mathbf{e}$ is a vector of residual effects; Q is a matrix from STRUCTURE relating $\mathbf{y}$ to $\mathbf{v}$; and X, S and Z are incidence matrices of 1s and 0s relating $\mathbf{y}$ to $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\mathbf{u}$, respectively. The variances of the random effects are assumed to be $\mathrm{Var}(\mathbf{u}) = 2KV_g$, and $\mathrm{Var}(\mathbf{e}) = RV_R$, where K is an $n \times n$ matrix of relative kinship coefficients that define the degree of genetic covariance between a pair of individuals; R is an $n \times n$ matrix in which the off-diagonal elements are 0 and the diagonal elements are the reciprocal of the number of observations for which each phenotypic data point was obtained; V_g is the genetic variance; and V_R is the residual variance. Best linear unbiased estimates (BLUE) of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\mathbf{v}$ (fixed effects) and best linear unbiased predictions (BLUP) of $\mathbf{u}$ (random effects) were obtained by solving the mixed-model equations[16,17]. We are implementing the new method in our publicly available software TASSEL.

URLs. SNP Consortium Linkage Map Project database, http://snp.cshl.org/linkage_maps/; TASSEL software, http://www.maizegenetics.net.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank past and present members of the Buckler lab for their assistance in phenotypic data collection. We thank P.J. Bradbury, Z. Zhang, R.L. Quass and E.J. Pollak for their insights and discussion regarding the mixed model. We thank N. Stevens for technical editing of the manuscript. This work was supported by the US National Science Foundation and the USDA-ARS. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
3. Hartl, D.L. & Clark, A.G. *Principles of Population Genetics* (Sinauer, Sunderland, Massachusetts, 1997).
4. Hey, J. & Machado, C.A. The study of structured populations-new hope for a difficult and divided science. *Nat. Rev. Genet.* **4**, 535–543 (2003).
5. Allison, D.B. Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**, 676–690 (1997).
6. Fulker, D.W., Cherny, S.S., Sham, P.C. & Hewitt, J.K. Combined linkage and association analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259–267 (1999).
7. Abecasis, G.R., Cardon, L.R. & Cookson, W.O.C. A general test for association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
8. Cardon, L.R. A sib-pair regression model of linkage disequilibrium for quantitative traits. *Hum. Hered.* **50**, 350–358 (2000).
9. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
10. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
11. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
12. Pritchard, J.K. & Donnelly, P. Case-control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227–237 (2001).
13. Thornsberry, J.M. *et al.* *Dwarf8* polymorphisms associate with variation in flowering time. *Nat. Genet.* **28**, 286–289 (2001).
14. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
15. Flint-Garcia, S.A. *et al.* Maize association population: a high resolution platform for QTL dissection. *Plant J.* **44**, 1054–1064 (2005).
16. Henderson, C.R. *Application of Linear Models in Animal Breeding* (Univ. of Guelph, Ontario, 1984).
17. Kennedy, B.W., Quinton, M. & van Arendonk, J.A.M. Estimation of effects of single genes on quantitative trait. *J. Anim. Sci.* **70**, 2000–2012 (1992).
18. George, A.W., Visscher, P.M. & Haley, C.S. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**, 2081–2092 (2000).
19. Loiselle, B.A. *et al.* Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). *Am. J. Bot.* **82**, 1420–1425 (1995).
20. Ritland, K. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res.* **67**, 175–186 (1996).
21. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
22. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**, 1198–1211 (1998).
23. Wright, S.I. *et al.* The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314 (2005).

24. Gardiner, J. *et al.* Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. *Plant Physiol.* **134**, 1317–1326 (2004).
25. Jurinke, C., van den Boom, D., Cantor, C.R. & Koster, H. The use of MassARRAY technology for high throughput genotyping. *Adv. Biochem. Eng. Biotechnol.* **77**, 57–74 (2002).
26. Long, A.D. & Langley, C.H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
27. Abecasis, G.R. *et al.* Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
28. Zhang, W. *et al.* Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl. Acad. Sci. USA* **101**, 18075–18080 (2004).
29. Hardy, O.J. & Vekemans, X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2**, 618–620 (2002).
30. SAS Institute. *SAS/STAT User's Guide* Version 8 (SAS Institute, Cary, North Carolina, 1999).