

ARTICLE



Covariance between nonrelatives in maize

Rex Bernardo ¹✉

© The Author(s), under exclusive licence to The Genetics Society 2022

The covariance between relatives is a tenet in quantitative genetics, but the covariance between nonrelatives in crops has not been studied. My objective was to determine if a covariance between nonrelatives is present in maize (*Zea mays* L.). The germplasm comprised 272 maize lines that were previously genotyped with 28,626 single nucleotide polymorphism (SNP) markers. Pairs of unrelated lines were identified on the basis of their membership probabilities in five subpopulations. The covariance between nonrelatives was assessed as the regression of phenotypic similarity on SNP similarity between unrelated lines. Out of 77 regressions, seven were significant at a 5% false discovery rate: anthesis and silking dates in unrelated B73 and Oh43 lines; plant height and ear height in unrelated Oh43 and PH207 lines; oil in unrelated A321 and Mo17 lines; starch in unrelated B73 and PH207 lines; and protein in unrelated B73 and Mo17 lines. The latter covariance was negative, and this negative covariance between nonrelatives was attributed to the subpopulations having different linkage phases between the markers and underlying causal variants. Overall, the results indicated that a covariance between nonrelatives in maize is not ubiquitous but is sometimes present for specific traits and for certain groups of unrelated individuals. I propose that the covariance between nonrelatives and the covariance between relatives be combined into a generalized covariance between individuals, thus giving a unified framework for expressing the resemblance regardless of the degree of relatedness.

Heredity (2022) 129:155–160; <https://doi.org/10.1038/s41437-022-00548-8>

INTRODUCTION

The resemblance between relatives or that like begets like has been observed since time immemorial, and the foundations of quantitative genetics were first laid when Fisher (1918) deduced the correlation between relatives on the assumption that the unknown loci controlling a quantitative trait behave in a Mendelian fashion. Two individuals are related if they have an ancestor in common. At any locus that affects a given trait, related individuals could inherit copies of the same allele found in the common ancestor. These copies of the same ancestral allele, which are described as being identical by descent, then make the relatives resemble each other.

Nonrelatives are individuals that are known to not have or assumed to not have any common ancestors in their respective pedigrees. Alleles shared by nonrelatives are alike in state (i.e., physically the same allele) but are not identical by descent. The probability of identity by descent can be assessed in a probabilistic manner from pedigrees. For example, the probability of identity by descent of alleles is 1/4 between a parent and its offspring, 1/4 between full sibs, and 1/8 between half sibs (Falconer 1960). In contrast, prior to the availability of molecular markers, there had been no way to assess the extent to which nonrelatives have alleles in common. As such, the covariance between nonrelatives has been implicitly assumed as zero in classical quantitative genetics. But is it always really zero?

If markers are in linkage disequilibrium with causal variants for a given trait, regressing the phenotypic covariance between nonrelatives on marker similarity (Ritland 1996; Lynch 1999) may reveal a nonzero covariance between nonrelatives. For complex traits in

humans, the covariance among nonrelatives or distantly related individuals has been estimated in the context of genomewide prediction (de los Campos et al. 2013) or in resolving the issue of missing heritability in genomewide association studies (Yang et al. 2017). The results from these aforementioned studies in humans have indicated the presence of a covariance among nonrelatives.

Linkage disequilibrium is likely stronger in plants than in humans, yet empirical information is unavailable on the covariance between nonrelatives in crops. If a covariance among nonrelatives is frequent and strong enough, such a covariance would need to be considered when coupling modern genomics tools with quantitative genetics theory in plant breeding (Bernardo 2020). Furthermore, having a covariance between nonrelatives and between relatives would suggest the need for a broader theoretical framework for the covariance between individuals in general. My objective in this study was to determine if a covariance between nonrelatives is present for different traits in groups of unrelated maize lines.

MATERIALS AND METHODS

Maize lines, marker data, and phenotypic data

Phenotypic and single nucleotide polymorphism (SNP) marker data were from previous experiments conducted at the University of Minnesota (Schaefer and Bernardo 2013a, 2013b). The maize germplasm included 272 publicly developed lines and private lines whose US Plant Variety Protection certificates had expired. The lines were evaluated for anthesis date (growing degree days from planting to when 50% of the plants were shedding pollen), silking date (growing degree days from planting to when 50% of the plants had exposed silks), plant height (distance in cm from the soil surface

¹Department of Agronomy and Plant Genetics, University of Minnesota, 411 Borlaug Hall, 1991 Buford Circle, Saint Paul, MN 55108, USA. Associate editor: Chenwu Xu.

✉email: bernardo@umn.edu

to the flag leaf), ear height (distance in cm from the soil surface to the ear leaf node), kernel oil concentration (g kg^{-1}), kernel protein concentration (g kg^{-1}), and kernel starch concentration (g kg^{-1}). Phenotyping was done in six location-year combinations in Minnesota in 2011 and 2012 (Schaefer and Bernardo 2013b). The phenotypic data analyzed herein were the least-squares means of each line for each trait across all six environments.

The lines were genotyped at 56,110 SNP loci on the Maize SNP50 BeadChip developed by Illumina (San Diego, California). Procedures for marker analysis of population structure were described by Schaefer and Bernardo (2013a) but are repeated here for convenience. Marker loci with a minor allele frequency less than 7% or with more than 10% missing data were disregarded, leading to 43,252 SNP loci. STRUCTURE software (Pritchard et al. 2000) was used to assess population structure among the lines. To help meet the assumption in STRUCTURE that the marker loci are in linkage equilibrium within subpopulations, a random subset of 3000 SNP loci was used in the model-based cluster analysis. This prior analysis by Schaefer and Bernardo (2013a) led to membership probabilities of each line for the following five subpopulations: A321 (Minnesota 13), B73 (Iowa Stiff Stalk Synthetic), Mo17, Oh43, and PH207 (Iodent) (Troyer 1999). There were 109 lines in the A321 subpopulation, 61 in the B73 subpopulation, 29 in the Mo17 subpopulation, 45 in the Oh43 subpopulation, and 28 in the PH207 subpopulation. For convenience, the five subpopulations are used herein to indicate line membership, so that 'a B73 line' refers to a line with primary membership in the B73 group rather than the line B73 itself.

Pairwise similarity and identifying pairs of unrelated lines

A set of 28,626 SNP loci with reduced multicollinearity was previously identified (Schaefer and Bernardo 2013b) via PLINK software (Purcell et al. 2007), according to a linkage disequilibrium maximum threshold of $r^2 = 0.9$ within a sliding window of 50 markers. The 28,626 SNP loci were then used for calculating the pairwise similarity among the lines. The numbers of SNP loci on each of the 10 maize chromosomes (in parentheses) were as follows: (1) 4182, (2) 1972, (3) 3618, (4) 3343, (5) 3367, (6) 2518, (7) 2493, (8) 2682, (9) 2316, and (10) 2135. Marker similarity between each of the 36,856 pairs of lines was calculated on an allelic basis. At a given SNP locus with alleles M and m , similarity was 1.0 between lines that both had the MM genotype; 1.0 between lines that both had the mm genotype; 0 between an MM line and mm line; and 0.5 when one or both lines had the Mm genotype (i.e., Mm versus Mm , Mm versus MM , or Mm versus mm). The within-locus similarity was summed across loci and divided by the total number of SNP loci.

Unrelated lines were identified in the following manner. Consider lines i and j and subpopulation k , and assume that p_{ik} was the membership probability of line i in subpopulation k whereas p_{jk} was the membership probability of line j in subpopulation k . The $p_{ik}p_{jk}$ value was calculated for each of the five subpopulations. Lines i and j were then considered unrelated when $p_{ik}p_{jk}$ did not exceed 0.01 for any of five subpopulations. In other words, pairs of unrelated lines were identified by finding those with membership-probability products that were less than 1% for all subpopulations. As indicated in the Discussion, this was likely a conservative approach for finding pairs of unrelated lines. Unrelated lines were identified across all subpopulations as well as for each of the 10 pairwise combinations of the five subpopulations (e.g., A321 lines versus B73 lines, A321 lines versus Mo17 lines, ... Oh43 lines versus PH207 lines).

Estimating the covariance between nonrelatives

The presence of a covariance between nonrelatives was assessed by regressing the cross products between trait values of nonrelatives on the marker similarity between the nonrelatives. This procedure required the absence of a nongenetic covariance between lines, and this requirement was met by randomization of the lines within each location in which phenotyping was done. Suppose line i was primarily a member of subpopulation k . For a given trait, the mean of line i was first modeled as $Y_{i(k)} = \mu + v_k + g_i + \text{error}$, where μ was the overall mean, v_k was the effect of subpopulation k (Yu et al. 2006), and g_i was the effect of line i . The $Y_{i(k)}$ value was corrected for the overall mean and subpopulation effect, i.e., $Y_{i(k)} = Y_{i(k)} - (\hat{\mu} + \hat{v}_k)$.

When unrelated lines were considered while ignoring their subpopulation memberships, the cross product for unrelated lines i (in subpopulation k) and j (in subpopulation k') was calculated as $C_{ij} = [y_{i(k)} - \bar{y}] [y_{j(k')} - \bar{y}']$, where \bar{y} was the mean corrected value of all the lines used in calculating the C_{ij} values. When pairs of unrelated lines from specific subpopulations were considered (e.g., an A321 line and an unrelated B73 line), the cross product for unrelated lines i and j was calculated as $C_{ij} = [y_{i(k)} - \bar{y}_k] [y_{j(k')} - \bar{y}_{k'}]$, where \bar{y}_k was the mean corrected value of the set of subpopulation k lines that were included in calculating the C_{ij} values, and $\bar{y}_{k'}$ was the mean

Table 1. Marker similarity, calculated from 28,626 single nucleotide polymorphism loci, among related lines and among unrelated lines in maize.

Group	Pairs of lines	Mean similarity (range)
All related lines	9490	0.656 (0.590, 0.980)
Within A321	5886	0.632 (0.595, 0.977)
Within B73	1830	0.721 (0.611, 0.980)
Within Mo17	406	0.676 (0.590, 0.928)
Within Oh43	990	0.657 (0.604, 0.976)
Within PH207	378	0.731 (0.614, 0.947)
All unrelated lines	5278	0.586 (0.435, 0.646)
A321, B73	1214	0.593 (0.560, 0.638)
A321, Mo17	397	0.579 (0.541, 0.602)
A321, Oh43	33	0.622 (0.610, 0.637)
A321, PH207	482	0.621 (0.592, 0.644)
B73, Mo17	738	0.529 (0.435, 0.619)
B73, Oh43	757 ^a	0.586 (0.561, 0.623)
B73, PH207	787	0.597 (0.567, 0.646)
Mo17, Oh43	233	0.582 (0.539, 0.637)
Mo17, PH207	307	0.590 (0.534, 0.637)
Oh43, PH207	329	0.616 (0.596, 0.631)

^aThe A662-LH149 pair, which had an unusually high similarity of 0.662 for two unrelated lines, was excluded from the analysis.

corrected value of the set of subpopulation k' lines that were included in calculating the C_{ij} values.

The covariance between nonrelatives was then assessed via the regression of C_{ij} on S_{ij} , where S_{ij} was the marker similarity between unrelated lines i and j (Ritland 1996; Lynch 1999). For convenience and ease of interpretation, the S_{ij} values were converted to percentages and expressed as a deviation from the mean. The regression of C_{ij} on S_{ij} is equal to the covariance only when the S_{ij} values are standardized, and the initial analysis involved standardizing the S_{ij} values. Such analysis proved less informative than the regression of C_{ij} on nonstandardized S_{ij} values: when comparing results across different subpopulations, it was helpful to assess the change in C_{ij} per percentage-point change (rather than per standardized-unit change) in S_{ij} . Hence, the results reported herein are regressions instead of covariances but the two are sometimes used interchangeably given the objective of determining whether a covariance (as reflected by the regression of C_{ij} on S_{ij}) between nonrelatives is present in maize.

The p -values for the regression coefficients were calculated via z -tests and a false discovery rate of 0.05 was imposed for the multiple comparisons made (Benjamini and Hochberg 1995). In addition to regressing C_{ij} on the across-genome S_{ij} , the regression of C_{ij} on per-chromosome S_{ij} was calculated. This analysis was conducted to assess if any significant regression was due to similarity across most or all chromosomes, or was due to similarity on specific chromosomes.

For reference purposes, the covariance between relatives was also assessed. Most pairs of lines within a given subpopulation were expected to be related, and the regression of C_{ij} of S_{ij} was calculated for all pairs of lines within each of the five subpopulations. A false discovery rate of 0.05 was imposed on the within-subpopulation regression coefficients (Benjamini and Hochberg 1995).

RESULTS

Among the 272 maize lines, there were 5278 pairs of unrelated lines across the five subpopulations (Table 1). When the analysis was restricted to unrelated lines between two subpopulations, the fewest pairs of unrelated lines was 33 for the A321 and Oh43 groups; these 33 pairs of unrelated lines involved only seven A321 lines and five Oh43 lines. The largest number of pairs of unrelated lines was between the A321 and B73 groups, with 1214 unrelated pairs that involved 52 A321 lines and 27 B73 lines. The number of

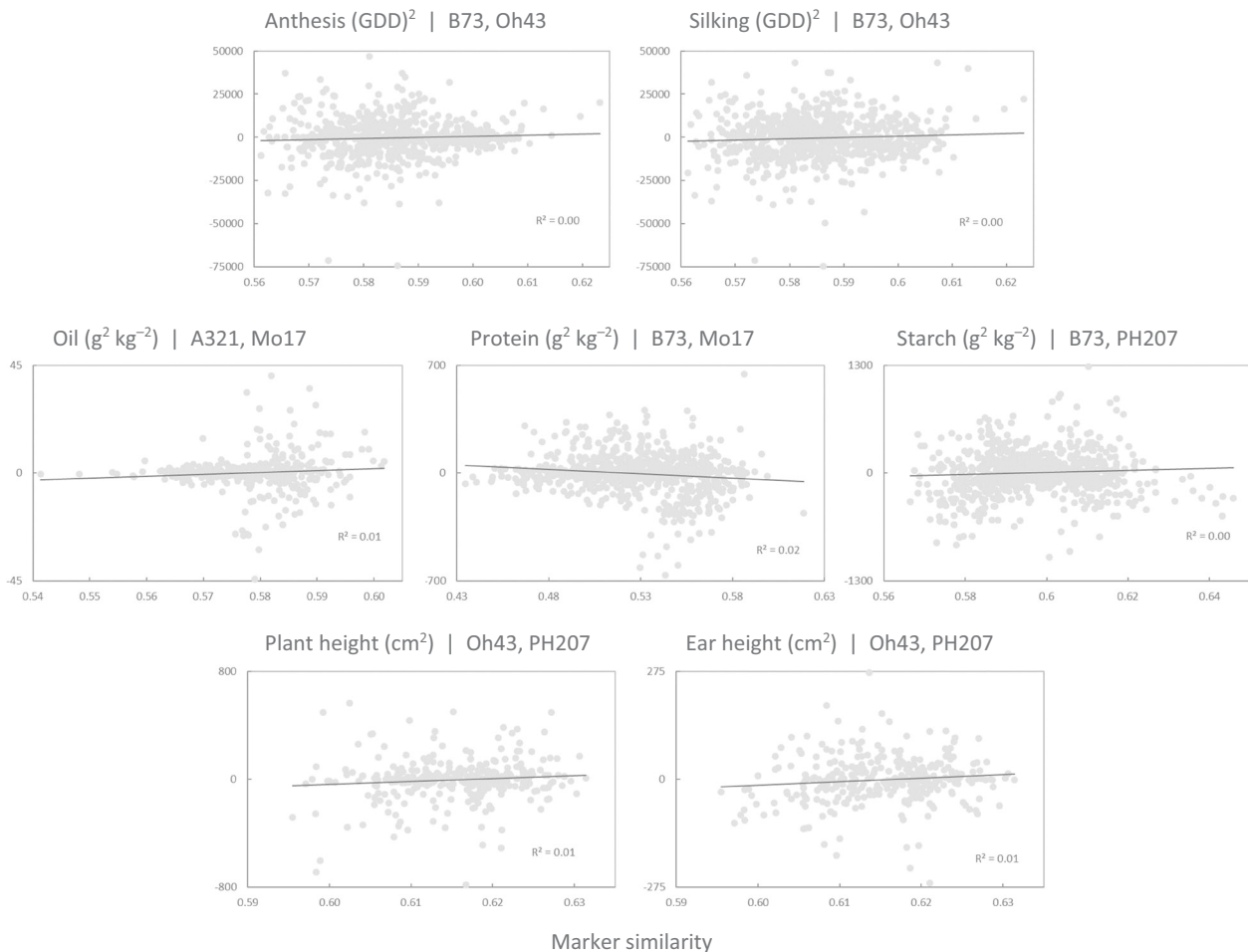


Fig. 1 Seven instances of a significant covariance between nonrelatives in maize, as assessed by the regression of cross-products between nonrelatives (Y-axis) on across-genome marker similarity.

pairs of related lines ranged from 378 within the PH207 group and 5886 within the A321 group (Table 1).

The marker similarity among the 5278 pairs of unrelated lines (across all subpopulations) ranged from 0.435 to 0.646 and had a mean of 0.586 (Table 1). The mean similarity was lowest (0.529) between unrelated B73 and Mo17 lines and was highest (0.622) between unrelated A321 and Oh43 lines. The range in similarity was widest between unrelated B73 and Mo17 lines (0.435 to 0.619) and was narrowest between unrelated A321 and Oh43 lines (0.610 to 0.637). Marker similarity was higher between related lines (mean of 0.656) than between unrelated lines (mean of 0.586) (Table 1). Within-group similarity was highest among B73 lines (mean of 0.729), and the highest similarity (0.980) was between B73 itself and line F42 within the B73 subpopulation.

With seven traits and 10 pairs of subpopulations plus the overall set of unrelated lines, there were 77 regressions of cross products between unrelated lines (C_{ij}) on marker similarity (S_{ij}). Out of the 77 regressions, the following seven were significant at a false discovery rate of 0.05: anthesis and silking dates in unrelated B73 and Oh43 lines; plant height and ear height in unrelated Oh43 and PH207 lines; oil in unrelated A321 and Mo17 lines; protein in unrelated B73 and Mo17 lines; and starch in unrelated B73 and PH207 lines (Fig. 1). The strongest covariance was for protein in unrelated B73 and Mo17 lines, for which the regression coefficient was -5.8 g kg^{-1} per percentage change in marker similarity and the correlation between C_{ij} and S_{ij} was -0.14 . The second strongest covariance was for ear height in unrelated Oh43 and PH207 lines, for which the regression coefficient was 8.9 cm per percentage change in marker similarity

and the correlation between C_{ij} and S_{ij} was 0.11. The weakest significant covariance was for starch in unrelated B73 and PH207 lines, for which the regression coefficient was 11.6 g kg^{-1} per percentage change in marker similarity and the correlation between C_{ij} and S_{ij} was 0.06.

For protein in unrelated B73 and Mo17 lines, the regression coefficients of C_{ij} on per-chromosome S_{ij} had low p -values across all 10 chromosomes (Table 2). In contrast, for ear height in unrelated Oh43 and PH207 lines, the p -values for regression coefficients were 0.02–0.07 for chromosomes 1, 2, 6, 7, and 8 but were 0.19–0.34 for chromosomes 3, 4, 5, 9, and 10.

The overall covariance between relatives was significant ($P = 0.05$) for all traits except ear height (Table 3). Within the A321 group, the covariance between relatives was significant for all seven traits. But within the Mo17 group, none of the regression coefficients of C_{ij} on S_{ij} was significant. Within the B73 group, the regression coefficients were nonsignificant for ear height and oil but were significant for the five other traits (Table 3, Fig. 2). Within-group covariances were significant for the two flowering-date traits in all subpopulations except Mo17. Regression coefficients were much smaller for oil than for protein and starch.

DISCUSSION

Weak covariance between nonrelatives for certain traits and subpopulations

The results showed that in maize, a covariance between nonrelatives is not ubiquitous but is sometimes present for specific

Table 2. Per-chromosome regression of nonrelative cross-products on marker similarity^a for protein in unrelated B73 and Mo17 lines and for ear height in unrelated Oh43 and PH207 lines.

Chromosome(s) used to calculate marker similarity	Protein (g ² kg ⁻²) in unrelated B73 and Mo17 lines		Ear height (cm ²) in unrelated Oh43 and PH207 lines	
	Regression	p value	Regression	p value
1	-2.0	0.0196	7.5	0.0230
2	-1.7	0.0026	-3.0	0.0589
3	-4.2	0.0010	1.5	0.1929
4	-5.0	0.0000	0.7	0.2875
5	-4.6	0.0001	0.9	0.3423
6	-1.9	0.0601	2.8	0.0695
7	-3.0	0.0018	2.4	0.0623
8	-5.3	0.0002	1.8	0.0583
9	-1.6	0.0724	0.7	0.3386
10	-2.5	0.0041	-0.9	0.3308
All	-5.8	0.0001	8.9	0.0227

^aMarker similarity was expressed as a percentage rather than a proportion in calculating the regression.

Table 3. Regression of relative cross-products on marker similarity^a among related lines for different traits in maize.

Group	Number of lines	Anthesis (growing degree days) ²	Silking (growing degree days) ²	Plant height (cm ²)	Ear height (cm ²)	Oil (g ² kg ⁻²)	Protein (g ² kg ⁻²)	Starch (g ² kg ⁻²)
A321	109	553	686	6.1	1.6	0.3	2.6	5.0
B73	61	109	114	6.1	0.6 ^b	0.0 ^b	11.3	22.3
Mo17	29	74 ^b	25 ^b	0.6 ^b	-0.2 ^b	0.1 ^b	0.3 ^b	0.4 ^b
Oh43	45	600	709	2.6 ^b	1.3 ^b	0.4	2.9	7.7
PH207	28	499	461	4.0 ^b	1.0 ^b	0.0 ^b	0.0 ^b	1.0 ^b
All	272	196	225	2.9	0.5 ^b	0.1	3.4	7.0

^aMarker similarity was expressed as a percentage rather than a proportion in calculating the regression.

^bNot significant at a false discovery rate of 0.05; all other values were significant.

traits and for certain groups of unrelated individuals. Such a covariance was generally weak and most often positive, yet a negative covariance between nonrelatives is also possible when it is assessed via a random set of genomewide markers. The covariance among nonrelatives was due to similarity across the maize genome in some cases and to similarity at specific chromosomes in other cases. While the results were from a single panel of maize lines, the lines represented the key germplasm groups in US maize (Troyer 1999, Schaefer and Bernardo 2013a) and they were enough to show that a covariance between nonrelatives is sometimes present in maize. The presence of a covariance among nonrelatives challenges a long-held, implicit assumption in quantitative genetics and requires a rethinking of how the covariance between individuals of varying levels of relatedness is best expressed.

In this study, unrelated lines were identified not from pedigree records (which were often incomplete) but on the basis having a joint probability of membership less than 0.01 for each of the five subpopulation assignments given by Schaefer and Bernardo (2013a). Some pairs of lines that were unrelated by pedigree were excluded from the sets of unrelated lines. For example, the French lines F2 and F7 were both developed by self-pollination from the Lacaune landrace (Tenaillon and Charcosset 2011) and, according to their pedigrees, are unrelated to the US line B73. The membership probabilities in the B73 subpopulation were 1.0 for B73, 0.03 for F2, and 0.08 for F7. Given that the product of their membership probabilities for the B73 subpopulation exceeded

0.01, F2 and F7 were excluded from the set of lines unrelated to B73 even if their pedigree records indicated otherwise.

The foregoing point suggested that the assessment of the covariance between nonrelatives in this study may have been too conservative. For example, the nonzero membership probabilities of F2 and F7 in the B73 subpopulation indicated some level of alikeness in state with B73. If F2, F7, and other lines unrelated by pedigree were included among the lines unrelated to B73, the range in S_{ij} values among the resulting unrelated lines could have increased. Such expansion in the range of the x-axis could then have led to a stronger regression coefficient. This conservative criterion aside, a key point worth considering is the very definition of two individuals being unrelated.

In particular, the results herein underscored how the definition of relatedness differs from an evolutionary genetics perspective versus a quantitative genetics perspective. If a particular crop species emerged from only one domestication event, as might have been the case with maize (Matsuoka et al. 2002) and rice (*Oryza sativa* L.) (Molina et al. 2011; Huang et al. 2012), then all individual plants of that species would be considered as related because of their singular ancestry. On the other hand, inferences in classical quantitative genetics rely on having a real or conceptual population that is in Hardy-Weinberg equilibrium and in which the individuals are assumed non-inbred and unrelated (Falconer 1960). If the common ancestry was in the distant past, then the probability of identity by descent between two individuals will

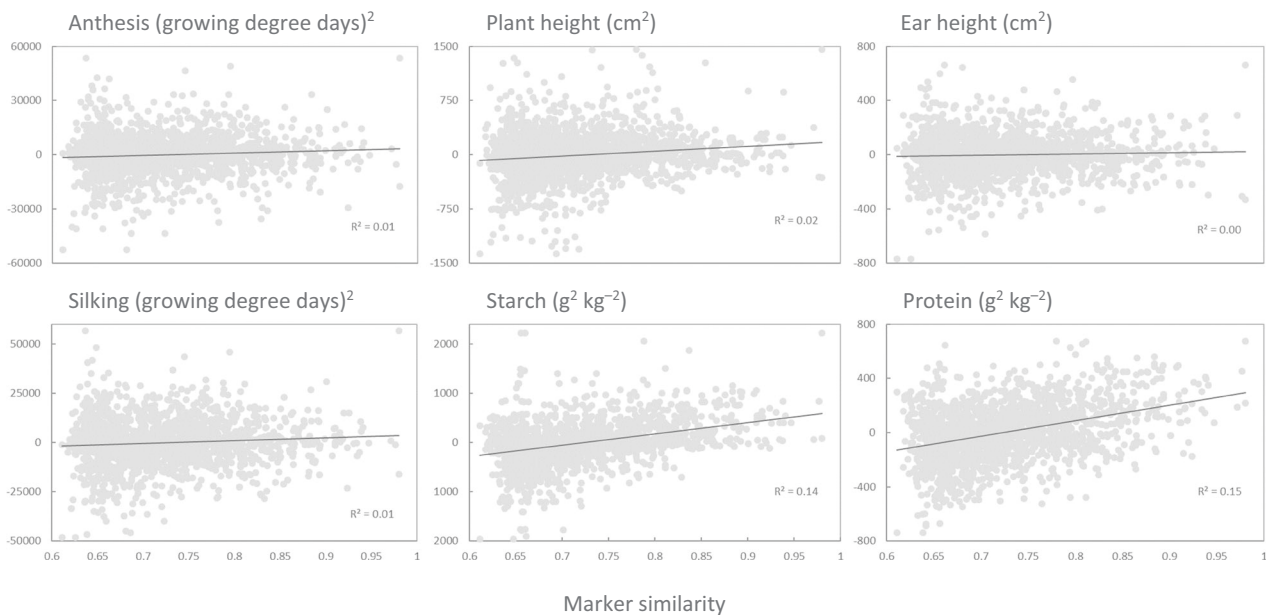


Fig. 2 Regression of nonrelative cross-products on across-genome marker similarity for different traits in the B73 group. Results are not shown for oil concentration, for which the regression was near zero and nonsignificant and the scatterplot was similar to that for ear height.

be close to zero and will have no meaningful contribution to the covariance between relatives. Relatedness in this study was viewed according to this quantitative genetics perspective.

The results indicated that a covariance between nonrelatives can be detected even when the range in S_{ij} values among unrelated lines is small. For example, the regression of C_{ij} for plant height and ear height on S_{ij} between unrelated Oh43 and PH207 lines was significant despite the range in S_{ij} values being only 0.035 (Table 1). On the other hand, the wide range (0.184) and low mean (0.52) of S_{ij} values between unrelated B73 and Mo17 lines reflected an ascertainment bias in the development of the Maize SNP50 BeadChip. This Illumina SNP chip was developed using B73 as the reference genome, and heterozygosity of the SNP markers in the B73 \times Mo17 cross was a criterion used to evaluate the utility of the SNP chip (Illumina 2012). It was therefore unsurprising that the mean similarity between unrelated B73 lines and Mo17 lines was lower than the mean similarity for the other pairs of subpopulations.

As expected, most of the covariances among relatives were significant with the notable exceptions of the Mo17 subpopulation, for which the regression of C_{ij} on S_{ij} was nonsignificant for all traits, and the PH207 subpopulation, for which the regression was nonsignificant for ear height and the three kernel composition traits (Table 3). The Mo17 and PH207 subpopulations had the fewest lines and they also tended to have narrowest ranges in line means for the traits studied (Schaefer and Bernardo 2013b).

Implications

Several previous studies that implicitly recognized a covariance between nonrelatives focused on how common causal variants contribute to a correlation between nonrelatives. For prediction of height among unrelated human subjects, the use of markers with the lowest p -values (from genomewide association analysis) led to higher prediction accuracies compared with using all of the markers (de los Campos et al. 2013). Multiple genomewide association studies (summarized by Visscher 2008) for height in humans have involved estimating heritability using markers with significant effects in different human populations. The current study differed from these aforementioned studies in that no attempt was made to estimate the similarity among unrelated individuals from a subset of markers with low p -values. That being said, the importance of specific chromosomes with regard to the

covariance among nonrelatives was evident in the regression of C_{ij} on per-chromosome S_{ij} (Table 2).

Previous analyses involving the regression of C_{ij} on S_{ij} in unrelated human subjects (Kemper et al. 2021) equated the regression coefficient to the additive genetic variance (Ritland 1996; Lynch 1999). In contrast, the regression coefficients in this study were used to assess covariances but not genetic variances which, unlike covariances, are positive by definition. The negative regression of C_{ij} for protein on S_{ij} in unrelated B73 and Mo17 lines (Fig. 1) was a unique result that indicated that, unlike the covariance between relatives which is always positive, the covariance between nonrelatives can be negative when it is assessed via genomewide markers.

The negative covariance for protein in unrelated B73 and Mo17 lines was likely due to differences in linkage phases between the two subpopulations; this same phenomenon has been recognized as a reason for a low accuracy of genomewide prediction between different populations (de Roos et al. 2009). Suppose the marker alleles at a locus are denoted by M and m whereas the causal alleles are denoted by Q and q . Furthermore, suppose that most of the gametes are in coupling phase in one subpopulation (MQ and mq) whereas most of the gametes are in repulsion phase in the second subpopulation (Mq and mQ). In this situation, similarity at the marker locus would be associated with dissimilarity at the causal locus, thus leading to a negative covariance. Thus, while any covariance among nonrelatives should be positive when similarity is directly assessed at the causal loci themselves, a negative covariance may arise if genomewide markers are used as proxies for the unknown causal variants.

This study focused on the initial step of determining whether a covariance between nonrelatives exists in maize, and investigations of the practical significance of this finding are deferred to future studies. Because the covariances between nonrelatives in this study were mostly weak, they could probably be safely ignored without much practical consequence. Such an approach is currently being used in genomewide prediction, for which the typical procedure is to either (1) capture identity by descent and exclude non-identity by descent using markers (Bernardo 1993) or (2) include both identity by descent and likeness in state without making a distinction between the two (de los Campos et al. 2013, Lorenz and Smith 2015). However, if the covariance between nonrelatives is

substantial (e.g., negative covariance for protein in unrelated B73 and Mo17 lines in Fig. 1), it might be advantageous to explicitly account for it when expressing the covariance between individuals.

At a single locus, the expectation of S_{ij} is $f_{ij} + \theta_{ij}$, where f_{ij} is probability that a random allele from i and a random allele from j are identical by descent (i.e., coefficient of coancestry), and θ_{ij} is the probability that a random allele from i and a random allele from j are alike in state, given that they are not identical by descent (Cox et al. 1985; Lynch 1988). In other words, unless unrelated individuals do not share any marker alleles, S_{ij} includes a portion due to relatedness (f_{ij}) and a portion due to nonrelatedness (θ_{ij}). This point also implies that whereas the covariance between nonrelatives was studied herein in isolation by focusing only on unrelated lines, a covariance between nonrelatives can play a role even among related lines. If dominance is absent, the covariance due to identity by descent is equal to $2f_{ij}V_A$, where V_A is the additive genetic variance (Falconer 1960). In an analogous manner, we define the portion of the covariance due to nonrelatedness as equal to $2\theta_{ij}Cov_U$, where Cov_U is the covariance between nonrelatives.

Suppose three pairs of lines all have $S_{ij} = 0.75$. If, in accordance with current practice, no distinction is made between the covariances due to relatedness and due to nonrelatedness, the value of the covariance will be identical for these three pairs of lines. In contrast, suppose a distinction is made between relatedness and nonrelatedness, and that the common S_{ij} of 0.75 corresponds to (1) $f_{ij} = 0.15$ and $\theta_{ij} = 0.60$ for the first pair of lines, (2) $f_{ij} = 0.20$ and $\theta_{ij} = 0.55$ for the second pair of lines, and (3) $f_{ij} = 0.25$ and $\theta_{ij} = 0.50$ for the third pair of lines. Furthermore, suppose that V_A (100) is larger than Cov_U (10). In this hypothetical example, the covariances between lines (calculated as $2f_{ij}V_A + 2\theta_{ij}Cov_U$) are now unequal and are 42 for the first pair, 51 for the second pair, and 60 for the third pair.

Accounting for the covariances due to relatedness and nonrelatedness could therefore alter the estimated covariances between individuals and, consequently, affect genomewide prediction and other procedures that rely on the covariance between individuals. This topic, along with how to partition S_{ij} into f_{ij} and θ_{ij} , is the focus of a follow-up study.

DATA AVAILABILITY

The maize datasets analyzed in this study can be accessed in Dryad (<https://doi.org/10.5061/dryad.4f4qrjfk>).

REFERENCES

- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 57:289–300
- Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. *Theor Appl Genet* 85:1055–1062
- Bernardo R (2020) Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity* 125:375–385
- Cox TS, Kiang YT, Gorman MB, Rodgers DM (1985) Relationship between coefficient of parentage and genetic similarity indices in the soybean. *Crop Sci* 25:529–532
- de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553
- de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet* 9:e1003608. <https://doi.org/10.1371/journal.pgen.1003608>
- Falconer DS (1960) *Introduction to Quantitative Genetics*. Oliver and Boyd: London.
- Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52:399–433

- Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490:497–501
- illumina (2012) MaizeSNP50 BeadChip. <https://www.illumina.com/products/by-type/microarray-kits/maize-snp50.html>
- Kemper KE, Yengo L, Zheng Z, Abdellouai A, Keller MC, Goddard ME et al. (2021) Phenotypic covariance across the entire spectrum of relatedness for 86 billion pairs of individuals. *Nat Commun* 12:1050. <https://doi.org/10.1038/s41467-021-21283-4>
- Lorenz A, Smith KP (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci* 55:2657–2667
- Lynch M (1988) Estimation of relatedness by DNA fingerprinting. *Mol Biol Evol* 5:584–599
- Lynch M (1999) Estimating genetic correlations in natural populations. *Genet Res Camb* 74:255–264
- Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez G, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proc Natl Acad Sci USA* 99:6080–6084
- Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A et al. (2011) Molecular evidence for a single evolutionary origin of domesticated rice. *Proc Natl Acad Sci USA* 108:8351–8356
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
- Ritland K (1996) A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* 50:1062–1073
- Schaefer CM, Bernardo R (2013a) Population structure and single nucleotide polymorphism diversity of historical Minnesota maize inbreds. *Crop Sci* 53:1529–1536
- Schaefer CM, Bernardo R (2013b) Genomewide association mapping of flowering time, kernel composition, and disease resistance in historical Minnesota maize inbreds. *Crop Sci* 53:2518–2529
- Tenaillon MI, Charcosset A (2011) A European perspective on maize history. *Comptes Rendus Biologies* 334:221–228
- Troyer AF (1999) Background of U.S. hybrid corn. *Crop Sci* 39:601–626
- Visscher PM (2008) Sizing up human height variation. *Nat Genet* 40:489–490
- Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017) Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* 49:1304
- Yu J, Pressoir G, Briggs WH, Bi IV, 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

ACKNOWLEDGEMENTS

The phenotypic and marker data analyzed herein were from the doctoral thesis experiments of my former Ph.D. student Christopher M. Schaefer. Thank you, Chris, for a most excellent dataset.

AUTHOR CONTRIBUTIONS

The single author thought of the problem, analyzed the data, interpreted the results, and wrote this manuscript.

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Rex Bernardo.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.