Published January, 2001

CROP SCIENCE

Volume 41 January–February 2001 Number 1

PERSPECTIVES

What If We Knew All the Genes for a Quantitative Trait in Hybrid Crops?

Rex Bernardo*

ABSTRACT

Plant genomics programs are expected to decipher the sequence and function of genes controlling important traits. Most of the important traits in crops are quantitative and are controlled jointly by many loci. What if we knew all the genes for a quantitative trait in hybrid crops? Will genomics enhance hybrid crop breeding, which currently involves selection on the basis of phenotypes rather than gene information? With maize (*Zea mays* L.) as a model species, I found through computer simulation that gene information is most useful in selection when few loci (e.g., 10) control the trait. With many loci (≥50), the least squares estimates of gene effects become imprecise. Gene information consequently improves selection efficiency among hybrids by only 10% or less, and actually becomes detrimental to selection as more loci become known. Increasing the population size and trait heritability to improve the estimates of gene effects also improves phenotypic selection, leaving little room for improvement of selection efficiency via gene information. The typical reductionist approach in genomics therefore has limited potential for enhancing selection for quantitative traits in hybrid crops.

BREEDERS HAVE SUCCESSFULLY IMPROVED CROPS despite not knowing the genes affecting quantitative traits. The numbers of genes controlling quantitative traits in different crops are yet unknown, although rough estimates include 69 loci for oil and 173 loci for protein content in the maize kernel (Dudley and Lambert, 1992). Experiments in many plant species have indicated that few quantitative trait loci have large effects, whereas many loci have smaller effects (Kearsey and Farquhar, 1998). Will knowing all the genes for a quantitative trait in crops further enhance breeding progress?

Suppose the identity and function of quantitative trait loci become known through extensive analysis of sequence homology, map position, gene expression, or genetic pathways (Bowen and Luedtke, 1997; Somerville and Somerville, 1999). If inbreds differ at only a few loci with large effects, then information regarding gene function may be directly useful in selection, e.g., "cherry-pick" as many desirable genes as possible into one single-cross hybrid. It becomes increasingly difficult to accumulate all the desirable genes into one hybrid if the inbreds differ at an increasingly large number of loci. Consequently, the effects of the individual genes need to be quantified for the information to be useful in selection (Kennedy et al., 1992). In other words, a maize breeder would need to know how many grams per kilogram of oil each gene for kernel oil contributes.

Selection in hybrid crops, such as maize, oilseed rape (*Brassica napus* L.), hybrid rice (*Oryza sativa* L.), rye (*Secale cereale* L.), sorghum (*Sorghum bicolor* L. Moench), sugar beet (*Beta vulgaris* L.), and sunflower (*Helianthus annuus* L.), is performed among testcrosses of recombinant inbreds and among hybrids (Fehr, 1987, p. 2, 5–6). Best linear unbiased prediction on the basis of trait phenotypes (T-BLUP; Henderson, 1985) is particularly useful for selecting improved single-cross hybrids (Bernardo, 1996). Selection, however, can be on the basis of both trait values and known genes (via trait and gene best linear unbiased prediction, i.e., TG-BLUP) if some of the genes are known, or on gene information alone (via standard multiple regression) if all the genes are known (Kennedy et al., 1992). Details of these procedures are in the **Genetic Model and Simulation** section.

I found that the advantage of TG-BLUP and multiple regression over T-BLUP increased as the number of loci decreased. For a trait controlled by 10 loci, TG-BLUP and multiple regression were up to 37% more efficient than T-BLUP in identifying the best untested hybrids (Fig. 1A). Likewise, TG-BLUP and multiple regression were up to 60% more efficient than T-BLUP in selecting the best recombinant inbreds developed from an F_2 population (Fig. 1B). As expected, the efficiency of TG-BLUP or multiple regression increased as a greater proportion of the loci became known.

Surprisingly, exploiting gene information through TG-BLUP or multiple regression did not substantially

Dep. of Agronomy and Plant Genetics, Univ. of Minnesota, 411 Borlaug Hall, 1991 Buford Circle, St. Paul, MN 55108-6026. Received 3 April 2000. *Corresponding author (berna022@umn.edu).

Published in Crop Sci. 41:1–4 (2001).

Abbreviations: T-BLUP, best linear unbiased prediction on the basis of trait phenotypes; TG-BLUP, best linear unbiased prediction on the basis on trait phenotypes and known genes.

enhance—and sometimes reduced—the selection efficiency when many loci (50 or 100) controlled the trait. Suppose 500 tested hybrids, a typical number of hybrids between two complementary heterotic groups in maize breeding programs (Bernardo, 1996), are available for estimating gene effects. Also suppose that a quantitative trait has a heritability of 0.20 and is controlled by 50 loci. For selecting among untested hybrids, a maximum increase in selection efficiency of 8% was achieved when the 30 loci with the largest effects were known (Fig. 1C). When all 50 loci were known, selection based on gene information was actually 5% less efficient than selection based on trait phenotypes alone. Heritability estimates for maize grain yield, on an entry mean basis, have ranged from 0.40 to 0.60 (Bernardo, 1996). When the trait had a heritability of 0.50 and was controlled by 100 loci, a maximum increase in selection efficiency of 7% was achieved when the 30 loci with the largest effects were known (Fig. 1D). Selection efficiency decreased by 13% when all 100 loci were known. It would therefore be more advantageous to ignore the genes with smaller effects, even if such genes are known.

Obtaining precise estimates of the effects of individual genes, by TG-BLUP or multiple regression, became more difficult as more genes became known. For example, the variance of gene effects at Locus 1 increased by 112 to 355% when the number of loci controlling the trait increased from 10 to 100 (i.e., with heritability of 0.20, and 10% of the loci being known; Fig. 2). Two factors contributed to the loss of precision in the estimates of gene effects: multicollinearity (i.e., lack of independence among the factors whose effects are being estimated), and inadequate sample size.

As the number of genes increases, the effects of the individual genes become associated with each other because of sampling (i.e., finite sample size) or linkage (i.e., finite genome size). Statistical procedures that reduce the effects of multicollinearity (e.g., ridge regression and orthogonalization; Draper and Smith, 1981, p. 258) can be used. The usefulness of these procedures in estimating gene effects needs further study. There

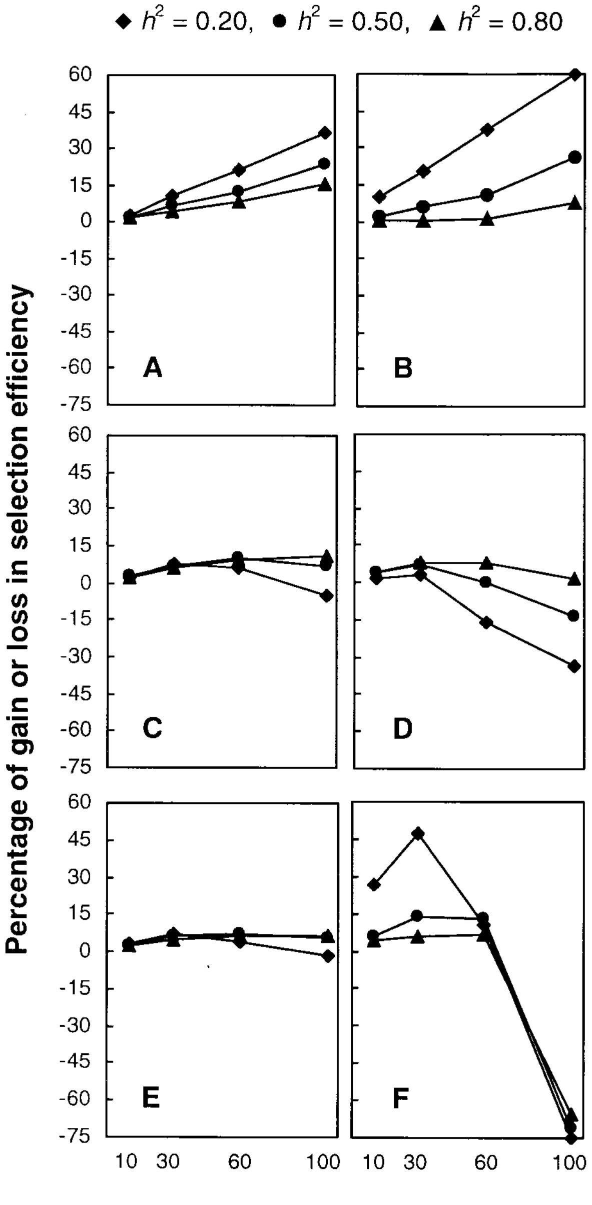

Fig. 1. Gain or loss in efficiency (%) of selection based on gene information compared with phenotypic selection. Trait heritability (h^2) was 0.20 (diamond), 0.50 (circle), and 0.80 (triangle). Selection was among untested single-cross hybrids or among recombinant inbreds, with different numbers of loci (l) and tested hybrids (n) from which gene effects were estimated. (A) Hybrids, l = 10, n = 500. (B) Inbreds, l = 10, n = 500. (C) Hybrids, l = 50, n = 500. (D) Hybrids, l = 100, n = 500. (E) Hybrids, l = 100, n = 2000. (F) Inbreds, l = 100, n = 2000.

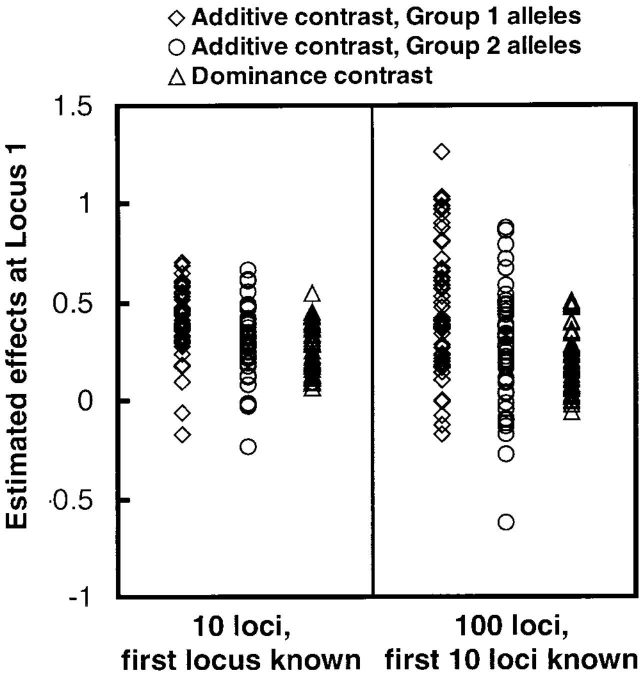

Fig. 2. Variability in the estimates of gene effects at Locus 1. Gene effects were expressed in terms of three orthogonal contrasts for the: testcross additive effect of Group 1 alleles (diamond); testcross additive effect of Group 2 alleles (circle); and dominance effect (triangle). The trait was controlled by 10 or 100 loci, of which 10% were known. Trait heritability was 20% and gene effects were estimated from 2000 tested hybrids.

are two reasons, though, for speculating that these procedures will have limited usefulness over TG-BLUP or multiple regression. First, multicollinearity hinders the estimation of the effects of individual predictor variables (e.g., individual genes), but does not necessarily hinder the estimation of overall response (e.g., sum of effects across genes) (Neter and Wasserman, 1974, p. 345). Second, procedures that correct for multicollinearity would seem to have little effect unless the number of hybrids from which gene effects are estimated is large. Suppose 100 known loci control the trait, and phenotypic data are available from 500 tested hybrids. Three contrasts are needed to specify gene effects (Fig. 2). Estimating 300 effects (i.e., 100 loci multiplied by three contrasts per locus) from 500 observations would likely remain difficult regardless of the estimation procedure used.

A straightforward way of improving the estimates of gene effects is to increase the number of hybrids from which the effects are estimated. The catch is that this approach also improves the effectiveness of T-BLUP itself, leaving little room for further improvement via gene information. Suppose gene effects, for a trait controlled by 100 loci and with heritability of 0.50, are estimated from 2000 instead of 500 tested hybrids. The maximum increase in selection efficiency via gene information was 7% among untested hybrids (Fig. 1E). The maximum increase in selection efficiency was 14% among F_2-derived inbreds (Fig. 1F). When all 100 loci were known, however, the selection efficiency among inbreds decreased by 71%. The probable reason for this large decrease is that the use of the two best inbreds as parents of the F_2 population tended to cause homozygosity of the desirable allele at loci with large effects, i.e., Locus 1 had the largest effect, Locus 100 the smallest. The average homozygosity in the F_2 population was 70% for the first 10 loci, and 47% for the first 30 loci. The multiple regression procedure therefore relied on the gene effects at minor loci, which were difficult to estimate.

The selection efficiencies of TG-BLUP or multiple regression were generally higher for recombinant inbred testcrosses than for untested hybrids, especially when gene effects were estimated at only a few loci. Selection in self-pollinated species, such as soybean [*Glycine max* (L.) Merr.] and oat (*Avena sativa* L.), is practiced among recombinant inbreds but not among hybrids. I therefore speculate whether genomics information would be more useful for a quantitative trait in self-pollinated crops than in hybrid crops. I also speculate whether genomics information for a quantitative trait would be more useful in animals than in hybrid crops. Compared with crop breeding, animal breeding programs are characterized by larger population sizes and individuals that are more distantly related (van Zyl, 1998). The larger population sizes suggest that gene effects can be estimated with greater precision in animals; the weaker genetic relationships suggest that increasing the population size will not lead to as large an increase in the effectiveness of T-BLUP in animals as in crop species. These two factors indicate that genomics may be more useful in selection for a quantitative trait in animals than in crops.

Plant genomics programs, in which heavy public and private research investments have been made (Service, 1998; Pennisi, 1998), will undoubtedly reveal useful biological information regarding the genetic basis of quantitative traits. However, the results indicated that genomics is of limited value in selection for quantitative traits in hybrid crops. Epistatic interactions, which were assumed absent in this study, would make the estimation of gene effects even more difficult. It is unknown whether methods other than TG-BLUP or multiple regression would substantially enhance the usefulness of gene information in selection. Perhaps the practical value of knowing all the genes in hybrid crops would be in creating new genetic variation. If the identity and function of important genes for a quantitative trait become known, then new genetic variation can be created by overexpressing genes, targeted mutagenesis, or searching for novel genes in other germplasm sources (Tanksley and McCouch, 1997). But after new genetic variation has been assembled in a breeding population, selection based primarily on trait phenotypes would be the preferred approach for improving inbreds and hybrids.

GENETIC MODEL AND SIMULATION

I wrote a Fortran program to simulate inbreds and single-cross hybrids, and to compare the T-BLUP, TG-BLUP, and multiple regression procedures. The simulation experiment was repeated 50 times. The 50 repeats differed at random in the arrangement of loci into linkage groups, genotypes of inbreds, phenotypic values of hybrids, and sets of tested and untested hybrids.

Heterotic Groups and Single-Cross Hybrids

Single-cross hybrids were made between an inbred from one heterotic group (i.e., Group 1) and an inbred from a complementary heterotic group (i.e., Group 2). In contrast, new recombinant inbreds are developed from crosses between inbreds from the same heterotic group. These new recombinant inbreds are then evaluated by crossing them to a tester from the opposite heterotic group. Each heterotic group comprised 76 inbreds. Four were founder inbreds, 18 were second-cycle inbreds, 27 were third-cycle inbreds, and 27 were fourth-cycle inbreds. The founder inbreds were unrelated within and between heterotic groups. Within each heterotic group, three second-cycle inbreds were randomly derived from the F_2 population of each of the six possible crosses among founder inbreds. One third-cycle inbred was randomly derived from the F_2 population of each of the 27 crosses between unrelated second-cycle inbreds. Suppose Inbreds 1 to 4 were founder inbreds. Inbred 5 was a second-cycle inbred developed from the cross between Inbreds 1 and 2, whereas Inbred 20 was a second-cycle inbred developed from the cross between Inbreds 3 and 4. Inbred 5 × Inbred 20 was then one of the 27 crosses between unrelated second-cycle inbreds. Finally, one fourth-cycle inbred was randomly derived from each of the F_2 populations obtained by chain crossing the 27 third-cycle inbreds.

There were $76 \times 76 = 5776$ possible Group 1 × Group 2 single-cross hybrids. A total of $n = 500$ or 2000 hybrids were assumed to have been tested (i.e., have phenotypic data), whereas the performance of the $(5776 - n)$ untested hybrids was evaluated by T-BLUP, TG-BLUP, or multiple regression.

Gene Effects and Phenotypic Values

Each locus had four alleles (+, +′, −, −′). Group 1 had the + and − alleles at odd-numbered loci, and the +′ and −′ alleles at even-numbered loci. In contrast, Group 2 had the + and − alleles at even-numbered loci, and the +′ and −′ alleles at odd-numbered loci. The allele frequency among founder inbreds in each heterotic group was ½ at each locus. The effects of the $l = 10$, 50, or 100 loci were exponential, which approximated an L-shaped distribution of the quantitative effects of segregating loci in metabolic pathways (Bost et al., 1999). Genotypic values of homozygotes at the kth (= 1 to l) locus were (Bernardo, 1999): 0.98^k for $(+/+)_k$; $1/2(0.98^k)$ for $(+'/+')_k$; $-1/2(0.98^k)$ for $(-/-)_k$; and $-(0.98^k)$ for $(-'/-')_k$. Complete dominance of the more favorable allele was present, whereas epistasis was assumed absent (Dudley, 1984). Linkage among the loci was generated by randomly locating the l loci on 10 chromosomes. The chromosome sizes corresponded to those in a published maize linkage map (Senior et al., 1996).

The phenotypic value of a hybrid was equal to the sum of genotypic values at each locus plus a random nongenetic effect. Nongenetic effects were normally and independently distributed with a mean of zero. The variance of nongenetic effects corresponded to a heritability of $h^2 = 0.2$, 0.5, or 0.8.

T-BLUP, TG-BLUP, and Multiple Regression

The proportion of the known loci was $p = 0$ in T-BLUP; $p = 0.1$, 0.3, or 0.6 in TG-BLUP; and $p = 1$ in multiple regression. The covariance between single-cross hybrids (Stuber and Cockerham, 1966) was calculated at the $(1 - p)l$ loci that were assumed unknown in T-BLUP and TG-BLUP. Testcross additive, dominance, and residual variances were assumed unknown and were estimated with an EM-type algorithm (Henderson, 1985) for the linear model for the performance of the tested single-cross hybrids (Bernardo, 1996). The linear model included $\boldsymbol{\beta}$, a $(1 + 3pl) \times 1$ vector of fixed effects. In T-BLUP, the only element of $\boldsymbol{\beta}$ was the grand mean. In TG-BLUP and multiple regression, the elements of $\boldsymbol{\beta}$ were the grand mean plus three orthogonal contrasts for the quantitative effects of each of the pl known loci: (i) testcross additive effect of the alleles from Group 1; (ii) testcross additive effect of alleles from Group 2; and (iii) dominance effects among genotypes (Kempthorne, 1957, p. 376). In T-BLUP and TC-BLUP, the performance of the untested hybrids was predicted as (Bernardo, 1996) $\mathbf{y_U} = \mathbf{K}\boldsymbol{\beta} + \mathbf{C_{UT}}\,\mathbf{C_{TT}}^{-1}(\mathbf{y_T} - \mathbf{X}\boldsymbol{\beta})$, where: $\mathbf{K}$ = design matrix relating $\mathbf{y_U}$ to $\boldsymbol{\beta}$; $\mathbf{C_{UT}}$ = matrix of genetic covariances between the untested hybrids and the tested hybrids; $\mathbf{C_{TT}}$ = phenotypic variance-covariance matrix among the tested hybrids; $\mathbf{y_T}$ = performance of tested hybrids; and $\mathbf{X}$ = design matrix relating $\mathbf{y_T}$ to $\boldsymbol{\beta}$. When all the loci were known, $\mathbf{y_U}$ was simply equal to $\mathbf{K}\boldsymbol{\beta}$.

F$_2$-Derived Recombinant Inbreds

The Group 2 inbred with the best mean testcross performance when crossed to all Group 1 inbreds was used as the tester. The pair of Group 1 inbreds (A_i and A_j) with the best performance when crossed with the tester was chosen based on their known genotypic values across all loci. A total of 200 random recombinant inbreds were developed from the $(A_i \times A_j)F_2$ population. Estimates of $\boldsymbol{\beta}$ were obtained from the analysis of the tested hybrids. In T-BLUP and TG-BLUP, the testcross performance of an inbred was predicted as $y_{(RITester)} = \mathbf{W}\boldsymbol{\beta} + h^2(y_P - \mathbf{W}\boldsymbol{\beta})$, where: $\mathbf{W}$ = incidence vector relating the inbred testcross to $\boldsymbol{\beta}$; and y_P = observed testcross performance of the inbred. When $p = 1$, $y_{(RITester)}$ was equal to $\mathbf{W}\boldsymbol{\beta}$.

Selection Efficiency

I compared T-BLUP, TG-BLUP, and multiple regression by calculating, for each procedure, the correlation between the predicted performance and true genetic performance of the 200 recombinant inbred testcrosses, as well as the correlation between the predicted and true performance of the $(5776 - n)$ untested hybrids. The mean correlation across the 50 repeats was calculated. The selection efficiency (Falconer, 1981, p. 149, 175) of TG-BLUP or multiple regression over T-BLUP was calculated as the correlation for TG-BLUP or multiple regression divided by the correlation for T-BLUP.

REFERENCES

Bernardo, R. 1996. Best linear unbiased prediction of maize single-cross performance. Crop Sci. 36:50–56.

Bernardo, R. 1999. Marker-assisted best linear unbiased prediction of single-cross performance. Crop Sci. 39:1277–1282.

Bost, B., C. Dillmann, and D. de Vienne. 1999. Fluxes and metabolic pools as model traits for quantitative genetics. I. The L-shaped distribution of gene effects. Genetics 153:2001–2012.

Bowen, B., and R. Luedtke. 1997. Understanding maize through genomics. p. 29–43. *In* Proc. Am. Seed Trade Assoc. Corn Sorghum Res. Conf., Chicago, IL. 10–11 Dec. 1997. Am. Seed Trade Assoc., Washington, D.C.

Draper, N.R., and H. Smith. 1981. Applied regression analysis. 2nd ed. John Wiley & Sons, New York.

Dudley, J.W. 1984. A method for identifying lines for use in improving parents of a single cross. Crop Sci. 24:355–357.

Dudley, J.W., and R.J. Lambert. 1992. Ninety generations of selection for oil and protein in maize. Maydica 37:81–87.

Falconer, D.S. 1981. Introduction to quantitative genetics. 2nd ed. Longman, London.

Fehr, W.R. 1987. Principles of cultivar development. Vol. 2, Crop species. Macmillan, New York.

Henderson, C.R. 1985. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. J. Anim. Sci. 60:111–117.

Kearsey, M.J., and A.G.L. Farquhar. 1998. QTL analysis; where are we know? Heredity 80:137–142.

Kempthorne, O. 1957. An introduction to genetic statistics. John Wiley & Sons, New York.

Kennedy, B.W., M. Quinton, and J.A.M. van Arendonk. 1992. Estimation of effects of single genes on quantitative traits. J. Anim. Sci. 70:2000–2012.

Neter, J., and W. Wasserman. 1974. Applied linear statistical models. Richard D. Irwin, Inc., Homewood, IL.

Pennisi, E. 1998. A bonanza for plant genomics. Science 282:652–654.

Senior, M.L., E.C.L. Chin, M. Lee, J.S.C. Smith, and C.W. Stuber. 1996. Simple sequence repeat markers developed from maize sequences found in the GENBANK database: Map construction. Crop Sci. 36:1676–1683.

Service, R.F. 1998. Chemical industry rushes toward greener pastures. Science 282:608–610.

Somerville, C., and S. Somerville. 1999. Plant functional genomics. Science 285:380–383.

Stuber, C.W., and C.C. Cockerham. 1966. Gene effects and variances in hybrid populations. Genetics 54:1279–1286.

Tanksley, S.D., and S.R. McCouch. 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. Science 277:1063–1066.

van Zyl, C.M. 1998. Estimation of genetic parameters for production traits of corn and dual-purpose sheep. Ph.D. diss. (Diss. Abstr. B 59/04, p. 1481). Univ. of Nebraska, Lincoln, NE.