Theor Appl Genet (1997) 94: 724–730

K. Chase · F. R. Adler · K. G. Lark

Epistat: a computer program for identifying and testing interactions between pairs of quantitative trait loci

Received: 16 January 1996 / Accepted: 27 September 1996

Abstract We describe a computer program, Epistat, which combines statistical methods and color-graphic displays to facilitate the analysis of interactions between pairs of quantitative trait loci (QTLs). Epistat organizes genetic-mapping data and quantitative-trait values into graphic displays which illustrate the individual effects of single loci as well as the interactions between any two loci. Keyboard commands allow the user to search the data set for individual QTLs and to test for interactions between QTLs. For a given trait, the program displays the effects of the alleles at each of two loci on the quantitative-trait value, as well as the effects of the interactions between these alleles. Log-likelihood ratios are used to compare the likelihood of explaining the effects by null, additive, or epistatic models. Examples of interactions in soybean are presented for near-infrared transmittance (NIT), seed number, and reproductive period. Epistat has been used to find numerous interactions between QTLs in soybean in which trait variation at one locus is conditional upon a specific allele at another.

Key words Computer program · Epistasis · Maximum likelihood · Molecular markers · QTLs

Introduction

Quantitative traits are traits in which several individual genes interact with the environment to produce the phenotype. Trait values are continuously distributed, with each individual gene explaining only part of the observed trait variation (Tanksley 1993). The genes responsible for trait variation, called quantitative trait loci (QTLs), are identified by their linkage to previously mapped qualitative genetic markers (Martin et al. 1989; Dudley 1993). For example, one can divide a recombinant inbred (RI) population into two sub-populations corresponding to the contrasting alleles at a qualitative marker tightly linked to a QTL. If the linkage is absolute, the mean difference between the two is an estimate of the effect of the two alleles segregating at the QTL. If the linkage is less than complete, this difference will be reduced by the degree of recombination between the marker locus and the QTL.

Although the effect of a QTL can be easily estimated, significant *interactions* between QTLs are usually difficult to identify (Tanksley 1993). It is often assumed that QTLs act additively; i.e., the variation attributable to a pair of loci is the sum of the variation explained by each of the individual loci. Such loci are assumed to act independently. If the effects of the alleles at one QTL are found to be dependent on the allele at another QTL, this assumption must be discarded. This interdependence between loci, or epistasis, can exaggerate or diminish the phenotype. For example, two QTLs which have no effect individually might produce an effect when combined. Alternatively, the effect of one QTL might be conditional upon the presence of a specific allele at another locus which by itself has no effect upon the trait.

The computer program Epistat was developed to help us identify and analyze interactions between loci in a RI soybean population (Lark et al 1995). This interactive program allows the researcher to quickly scan the data set for potential interactions and apply statistical tests to the candidate loci. It does this by dividing the homozygous RI segregant population into sub-populations based upon the four possible

Communicated by A. L. Kahler

K. Chase · F. R. Adler[1] · K. G. Lark (✉)
Department of Biology, University of Utah, Salt Lake City, UT 84112, USA

Present address:
[1] Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA

genotypic combinations of the parental alleles at two loci (Epistat is currently designed to use data on homozygous loci only and ignores heterozygous loci). The nature of the interaction between two loci is illustrated by graphical displays, which also allow the user to discern interesting patterns or anomalies in the data set.

Materials and methods

Data display

Epistat displays the genotypic subpopulations as cumulative distribution graphs. Cumulative distributions allow easy visual comparison of many aspects of the sub-populations. An example comparing two alleles *A* and *a* of a hypothetical locus is shown in Fig. 1. Each sub-population is the basis for one of the sigmoid-shaped curves. All of the individuals in the sub-population are ranked along the y axis according to their phenotypic value along the x axis. Thus all of the higher values are at the top of the curve and to the right, while all of the lower values are at the bottom to the left. The median of the sub-population is half way up the curve and for symmetric distributions is equal to the mean. The variance of the sub-population is the general slope of the curve. The mean effect of a locus on the phenotype is the distance between the two curves representing the two alleles.

Figure 2 is an example of Epistat's full display screen, showing two loci and their interactions. The allele at each homozygous locus is represented by a single symbol (e.g., *A* or *a*). Thus the genotype for two loci is represented by a pair of symbols (e.g., *Ab* or *aB*; acting as shorthand for *AAbb* or *aaBB*, etc.). The individual effects of the two loci are presented in the outside panels (1 and 4) and the interactions between the two loci are shown in the inner two panels (2 and 3).

Figure 2 is an example of a proto-typical additive interaction. In panel 1, Epistat has divided the entire population into two sub-populations, one homozygous for the parental allele "*A*" and the other homozygous for the parental allele "*a*". The significant separation between the two curves indicates a large effect for the first locus (note the means indicated at the bottom of the panel). In panel 4, Epistat has divided the same population into two homozygous classes for the second locus; representing the parental allele "*B*" and the parental allele "*b*". The effect of the second locus, while not as large as the first, is still significant. In the two center panels, Epistat has sorted the entire population into the four genotypic sub-populations that represent the four homozygous allelic combinations at the two loci. Panel 2 compares the two sub-populations that contain the "*A*" allele (*AB* and *Ab*). The separation between the two curves is an indication of the effect of the second locus in the presence of the "*A*" allele. Because the curves are similar to those in panel 4, the effect is the same as the effect of the second locus by itself. Panel 3 compares the two sub-populations with the "*a*" allele (*aB* and *ab*). Again the curves are similar, meaning that the effect of the second locus is the same in the presence of the "*a*" allele as it is by itself. The similarity of the middle two panels thus shows the independence of these two loci, which results in an additive effect. The largest trait values occur for plants with the *ab* genotype, associated with the large allele from each locus. The deviation of the mean trait value of the sub-population from the mean of the total population can be explained as the sum of effects for each allele separately. As

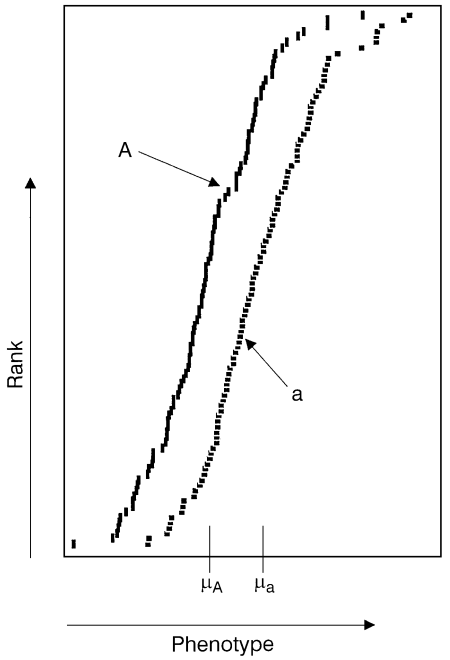

Fig. 1 Simplified graphic display of the effect the alleles of one locus on the phenotype. The individuals are ranked along the y axis according to their trait value plotted along the x axis. The magnitude of the separation between the two curves is an indication of the effect of the locus, *A* or *a*

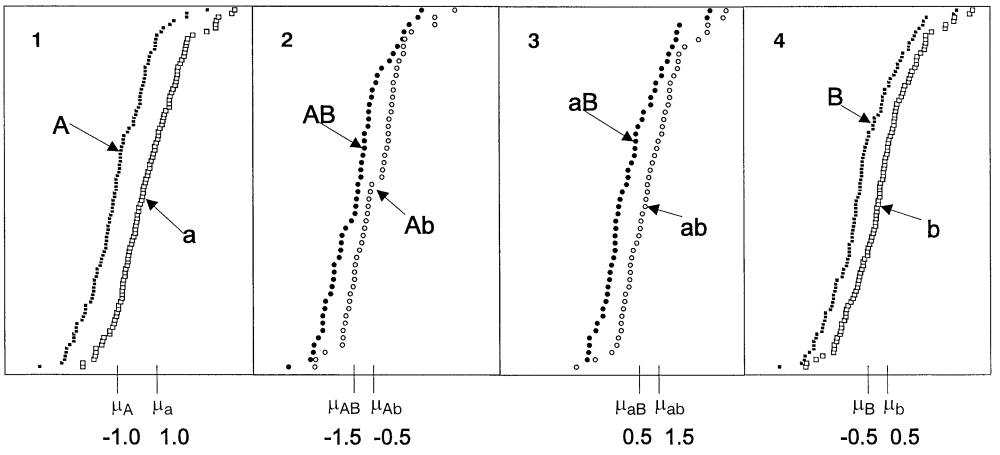

Fig. 2 Simplified graphic display of the type used in Epistat. Cumulative distributions of the quantitative trait values are presented ranked according to their trait values. *Panels 1 and 4*: the population is subdivided according to the alleles at locus 1 (*panel 1 A* or *a*) or at locus 2 (*panel 4 B* or *b*). *Panels 2 and 3*: the population is subdivided into four sub-populations corresponding to the four genotypes (*AB*, *Ab*, *aB*, *ab*). The mean (μ) of each genotypic sub-population is indicated below the panels

we will see, differences between the middle panels can both identify and characterize interactions.

Statistical methods

We used likelihood ratios (Bickel and Doksum 1977) to statistically compare an epistatic and an additive model. The estimated mean (m), the sample mean ($\hat{\mu}$) the sample size (n) and the sample variance ($\hat{\sigma}^2$) for a genotype are denoted by the appropriate parameter with a genotypic subscript, (e.g., $\hat{\mu}_{Ab}$ for plants homozygous for A at the first locus and b at the second). The epistatic model assumes no constraints on the means and computes the log-likelihood ratio associated with the maximum-likelihood estimator of each mean. The additive model requires that $m_{AB} + m_{ab} = m_{Ab} + m_{aB}$ and computes the log-likelihood ratio associated with the maximum-likelihood estimator of the means subject to this constraint. We assume that the distributions are normal and that the variances are given by the uncorrected sample variances: $\hat{\sigma}^2_{Ab}$, etc. The log-likelihood function for each model is:

$$\begin{aligned}\log(L) = & -n_{AB}(m_{AB} - \hat{\mu}_{AB})^2/2\hat{\sigma}_{AB^2} \\ & - n_{Ab}(m_{Ab} - \hat{\mu}_{Ab})^2/2\hat{\sigma}_{Ab^2} \\ & - n_{aB}(m_{aB} - \hat{\mu}_{aB})^2/2\hat{\sigma}_{aB^2} \\ & - n_{ab}(m_{ab} - \hat{\mu}_{ab})^2/2\hat{\sigma}_{ab^2} + \text{constant}.\end{aligned}$$

Under the epistatic model, the maximum-likelihood estimators of the means are the sample means. Under the additive constraint, the maximum-likelihood estimators are:

$$m_{AB} = \hat{\mu}_{AB} - \delta c_{AB}/c,\ m_{Ab} = \hat{\mu}_{Ab} + \delta c_{Ab}/c,$$

$$m_{aB} = \hat{\mu}_{aB} + \delta c_{aB}/c,\ m_{ab} = \hat{\mu}_{ab} - \delta c_{ab}/c,$$

where $c_{AB} = \hat{\sigma}_{AB^2}/n_{AB}$ etc., $c = c_{AB} + c_{Ab} + c_{aB} + c_{ab}$ and $\delta = (\hat{\mu}_{AB} + \hat{\mu}_{ab}) - (\hat{\mu}_{Ab} + \hat{\mu}_{aB})$; δ represents the "deviation from additivity" and would be 0 if the sample means satisfied the additive constraint.

The log-likelihood ratio for the additive model is computed as the log of the ratio of the additive model and the epistatic model. A small log-likelihood ratio indicates that the data can be effectively explained by the additive model, while a large log-likelihood ratio indicates that the data are inconsistent with the additive constraint. With normal distributions and equal sample sizes, and assuming additivity, the log-likelihood ratio will exceed a value X with a probability $2(1 - \Phi(\sqrt{2}X))$ where Φ is the cumulative distribution function of the standard normal (Lark et al. 1995).

To test the effect of the alleles at a single locus, we compare a null model that assumes that the means are equal with a model allowing for an effect. For example, to examine the effect of alleles "a" and "A", we compare the log likelihood with the constraint $m_a = m_A$ to the unconstrained log likelihood. For each model, the log-likelihood function is:

$$\log(L) = -n_A(m_A - \hat{\mu}_A)/2\hat{\sigma}_{A^2} - n_a(m_a - \hat{\mu}_a)/2\hat{\sigma}^2_a.$$

In the unconstrained case, the maximum-likelihood estimators of the means are the sample means. Under the equality constraint, the estimator of $m_a = m_A$ is the mean of the entire sample (Bickel and Doksum 1977). The same method is used to evaluate the effect of alleles "a" and "A" in a given allelelic sub-population "B" or "b" by restricting attention to the appropriate sub-population.

Monte Carlo simulations

Because some genotypes and traits do not satisfy the normality and variance assumptions, we conducted Monte Carlo simulations (Manly 1991). Each Monte Carlo simulation was specific to a given trait and pair of loci. Sub-populations were randomly chosen without replacement from the total population and log-likelihood ratios for the additive model were calculated based on these sub-populations. A total of 1,000,000 trials were done and a tally of the log-likelihood ratios in the range of 0–20 was maintained.

Monte Carlo simulations were used to verify the predicted probability of exceeding a given log-likelihood ratio. We compared the distributions from 50 different Monte Carlo simulations to the predicted distribution. The fraction of trials that exceeded a given log-likelihood ratio closely matched that predicted by theory, with only a few cases exceeding the theoretical prediction by a factor of about ten at very high log-likelihood ratios. When the sub-population sizes are moderate to large ($n > 30$) the log-likelihood ratio appears to provide a robust test for significance, even for data that deviate from normality.

Sources of data

Seed number and reproductive-period data for the examples discussed in this paper were obtained from a Minsoy × Noir1 recombinant inbred population originally described by Mansur and Orf (1995). Data for qualitative genetic markers (e.g., seed color and RFLP markers) and quantitative traits (e.g., seed number and reproductive period) were taken from the data base described by Mansur et al. (1996). Near-infrared transmittance (NIT) data and RI phenotypes for the I and R loci were kindly provided by James Specht, University of Nebraska. The NIT data were generated by Warren Rayford (ARS-USDA, Peoria, Illinois) who analyzed seed harvested from RI lines grown in Nebraska during the summer of 1995. Near-infrared transmittance data are ordinarily used to determine the oil content of yellow seed, a procedure not recommended for heavily pigmented seed (Rayford, personal communication).

Results

The effects of seed coat color on near-infrared transmittance (NIT) data

A well-understood epistatic interaction in soybeans involves the I and R loci affecting seed coat color. The I locus controls the presence (*ii*) or absence (*II*) of pigment and the R locus results in black (*RR*) or brown (*rr*) pigmentation, when pigmentation is present (Williams 1952). The effect of an allele at the R locus is thus conditional upon which allele is present at the I locus.

Soybean seeds are assayed for oil content using near-infrared transmittance (NIT) techniques. However, transmittance values can be strongly affected by the presence of black pigment in the seed coat (Rayford, personal communication). This means that the NIT values in this RI population constitute a quantitative, polygenic trait, with the seed coat-color loci acting as two of the perhaps many loci that account for the variation in the trait. Because these RI lines can be genotyped at the I and R loci by direct examination of the seed, and because many soybean breeders and geneticists are familiar with these markers, we have chosen these two qualitative markers as an illustration of the use of Epistat (Fig. 3). The parental genotypes for the Minsoy-Noir1 RI population are: Minsoy (*IIrr*) which has a non-pigmented yellow seed coat, and Noir1 (*iiRR*) which has a black seed coat. There are thus only

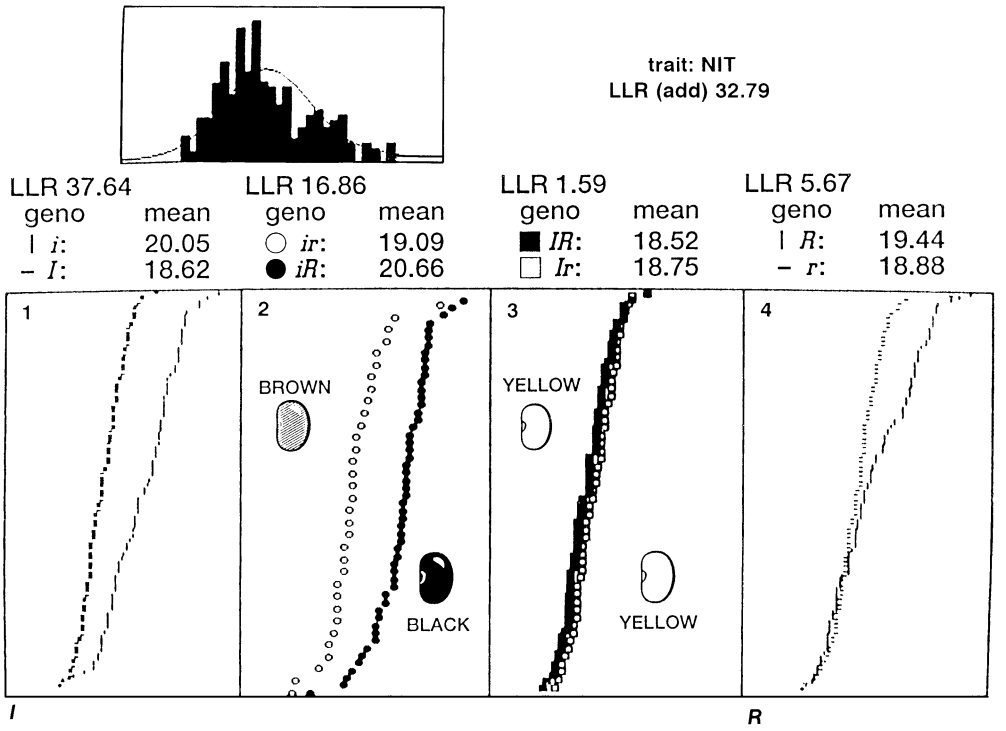

Fig. 3 Epistat display for the trait "NIT" and the loci *I* and *R*. The frequency distribution of the trait is shown in the upper left, superimposed on the normal distribution for the trait matching the sample mean and standard deviation. Cumulative distributions of different genotypic sub-populations are shown below. *Panel 1*: the population is subdivided into the genotypes *I* (–) and *i* (|); *panel 2 iR* (*black* ●) and *ir* (*brown* ○); *panel 3*: *IR* (*yellow* ■) and *Ir* (*yellow* □); and *panel 4*: *R* (*yellow and black* |) and *r* (*yellow and brown* –). Above each panel, values of the means for the two genotype sub-populations are shown. The log-likelihood ratio (*LLR*) against the null model is shown for each panel above the means. In the center is shown the LLR against the additive [e.g. *LLR* (add) 32.79]

three RI phenotypes: yellow (*IIRR* or *IIrr*), black (*iiRR*), or brown (*iirr*). All four genotypes are present in the RI population in approximately equal proportions with one half of the phenotypic segregants being yellow, one quarter brown, and one quarter black. Near-infrared transmittance by seeds with black coats is higher than that of seeds with yellow or brown coat but does not distinguish yellow from brown seeds.

Figure 3 shows an Epistat display of the effects of the *I* and *R* loci on the NIT values. In the upper left panel Epistat displays a frequency distribution bar graph for the NIT trait, with the curve showing a normal distribution with matching mean and variance. The four numbered lower panels display the cumulative distributions for the genotypic sub-populations.

The two outside panels illustrate the individual effects of the *I* and *R* loci on NIT values. Panel 1 distinguishes yellow seeds *(I)*from colored seeds *(i)*, demonstrating that the *I* locus is a QTL with a large effect on the NIT phenotype. In panel 4 only the upper half of the *R* genotypes are distinguishable from their *r* counterparts, being the plants with black seed coat *(Ri)* and the highest NIT values. The remaining *RI* genotypes are indistinguishable from the *ri* and *rI* plants. Although the difference between the two curves in panel 4 is visually striking, the likelihood score (5.67) is not as conspicuous.

The two center panels illustrate the interaction (epistasis) between the *I* and *R* loci; most simply demonstrated as the marked difference between the distributions in panel 2 and those in panel 3. When the genotype is *i* (seed-coat pigmented) the difference between the *R* (black) and *r* (brown) alleles is clear (panel 2). When the genotype is *I*, color is not expressed, the phenotypes of *Ir* and *IR* are the same (panel 3) and the distributions are indistinguishable. These two panels show that expression of the *R* locus is conditional on the *I* genotype. The enormous additive log-likelihood ratio of 32.79 confirms the epistasis between the two loci. Viewing the actual distributions, however, provides a wealth of information that is not supplied by the statistics. There is one plant in panel 2 with an *ir* genotype which has a much greater NIT value than the rest of the population, indicative either of an error in scoring or another, as yet undiscovered, source of variation for the NIT value. The *i* and *I* populations in panel 1 diverge more at the higher NIT values than at the lower values indicating another source of variation for the higher values (namely the *R* locus). The *R* locus distinguishes one half of the population but not the other, a possible indication of conditional epistasis (in this case the effects of the *I* locus).

The first example has made use of a quantitative trait that is affected by well-documented genes for seed coat color. In the next two examples the nature of the individual genes is unknown.

Seed number

Seed number is a polygenic trait with normally distributed values (Fig. 4). A significant QTL for seed number has been shown to be linked to the RFLP-locus A109a

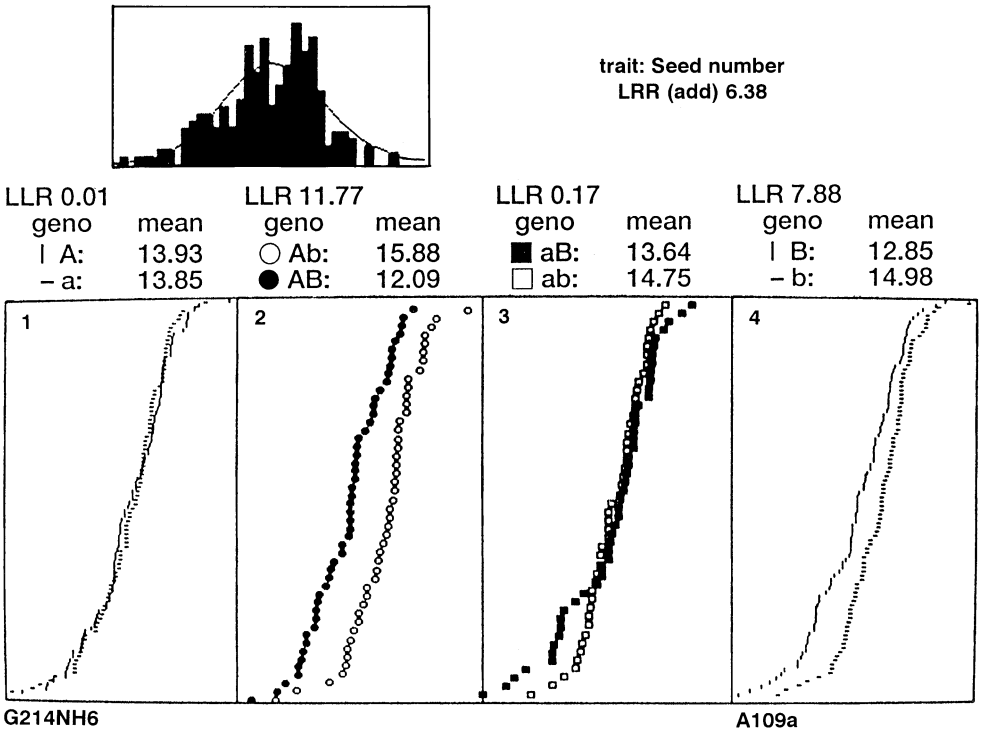

Fig. 4 Epistat display for the trait "seed number" and the RFLP loci G214NH6 and A109a. For a general description of the output see Fig. 3. The genotypes represented are: *A* (G214NH6; Noir I), *a* (G214NH6; Minsoy), *B* (A109a; Noir I) and *b* (A109a; Minsoy)

(Mansur et al. 1996). As shown in panel 4, those plants with the "*B*"allele have a lower seed number than those with the "*b*" allele. This locus was identified by searching all loci for single gene effects. The other locus, G214*N*H6, does not account for any significant variation in seed number as shown in panel 1. Epistat selected this locus in an automated search for loci that had a significant epistatic interaction with A109a. The G214*N*H6 locus interacts strongly with A109a as shown in the middle two panels. The statistical support for the interaction is shown above the panels as the log likelihood ratio of the epistatic and additive models (6.38).

The nature of this interaction can be deduced from the cumulative distributions. When the RI genotypes homozygous for the "*A*" allele of G214*N*H6 are sorted into sub-populations corresponding to the "*B*" and "*b*" homozygotes at the A109a locus the two sub-populations differ significantly in seed number (panel 2). In contrast, when the RI genotypes homozygous for the "*a*" allele of G214*N*H6 are sorted into sub-populations according to the alleles at the A109a locus (panel 3), the sub-populations do not differ in seed number. The effects of the alleles at locus A109a are thus conditional upon which allele is present at locus G214*N*H6.

The interesting anomaly at the bottom of panel 3 (Fig. 4), in which the two curves diverge for the low seed number values, prompted an investigation. It was found that the RI lines with very low seed numbers were genotypes with partially sterile phenotypes (variable penetrance). An anomaly of this nature would be difficult to detect with statistical methods, but is readily apparent in Epistat's visual display.

Reproductive period

Figure 5 shows two loci that regulate the reproductive period, here defined as the time between flowering and maturity (Mansur et al. 1996). Epistat found the loci shown by searching all pairs of loci for a log-likelihood ratio indicative of epistasis. All pairs of loci which exceeded a pre-established log-likelihood ratio were written to a file for subsequent detailed analysis.

Although neither locus has a significant independent effect, the two-locus interaction accounts for a large amount of variation. Panels 1 and 4 demonstrate that the two alleles at each locus have little effect on reproductive period. In contrast, a difference in reproductive period is evident for the pairs of genotypes in panels 2 and 3. When both loci are derived from the same parent (Minsoy or NoirI) the reproductive period is long (means = 65.71 and 65.72) and about the same. When the alleles at the two homozygous loci differ (one from Minsoy and one from Noir I) the reproductive periods are much shorter (63.31 and 62.95) and differ. This form of epistasis is quite distinct from the previous examples and represents the result expected if co-adapted loci are uncoupled by outcrossing and subsequent segregation. Finding this form of epistasis required searching 24,200 different pairs of loci. False

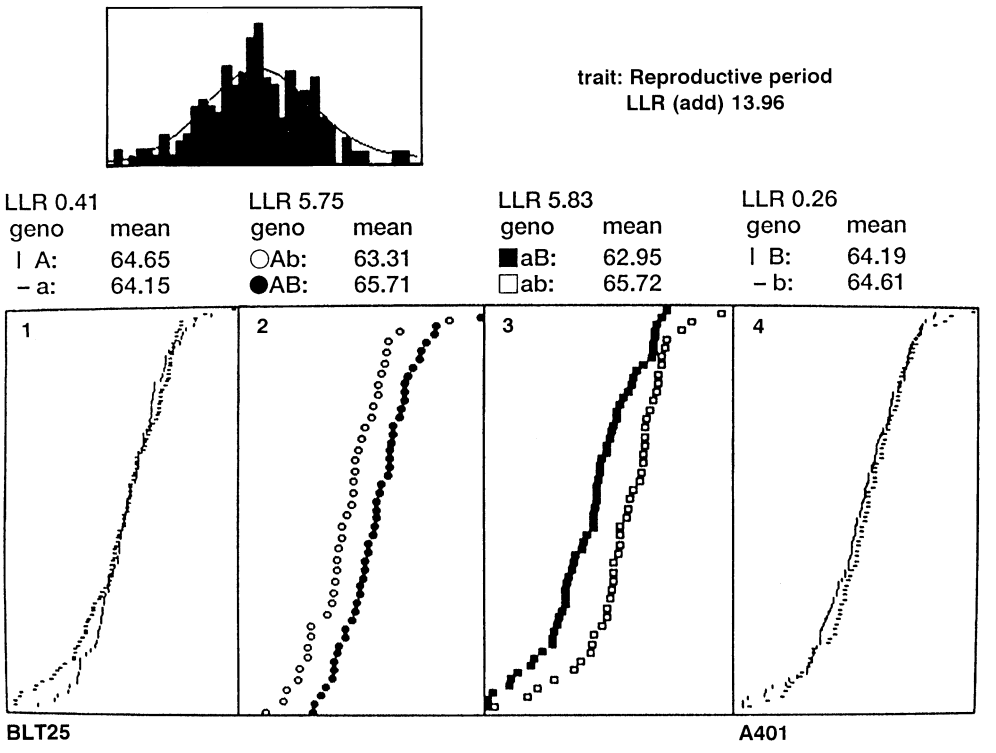

Fig. 5 Epistat display for the trait "reproductive period" and the RFLP loci BLT25 and A401

positives are always possible with such numbers but the statistical significance of a log-likelihood ratio of 13.96 is high. Epistat was used to run a Monte-Carlo simulation which provided a P-value of 3×10^{-7} (which corresponds to a P-value of 7.3×10^{-3} after correction for the number of pairs)

Discussion

We have described a computer program, Epistat, designed to graphically display and statistically evaluate interactions between paired QTLs. The program subdivides a population of homozygous genotypes into the four genetically distinct sub-populations determined by the two possible alleles present at each of two homozygous loci. The trait values for these sub-populations are displayed as cumulative distributions. Likelihood methods are used to evaluate the significance of QTLs that are linked to qualitative markers, and to identify and evaluate the significance of the interactions between pairs of QTLs.

The log-likelihood ratio quickly approximates the significance of interactions that have been identified. Monte Carlo methods can be used to generate a probability distribution of log-likelihood ratios that can be used to evaluate the significance of the interaction within the context of the number of loci analyzed (i.e., 200 or 20,000 possible interactions in our soybean population). The likelihood method has proven robust in that various data sets produce consistent probability distributions for likelihood values (see Materials and methods), even when distributions deviate significantly from normality.

Several other tools for data analysis are included in the program as user options. Some of these have already been described, such as displays of the frequency distribution of trait values. In addition, the following procedures have been included to facilitate data analysis:

(1) Visual generation of many random cumulative graphs (Monte Carlo) for simultaneous comparison with data. The random samples are graphed over the existing curves. Whenever one graph overlaps another the color of the overlapping region is drawn in a "hotter" color. This gives a general feel for the probability of having a curve in a given region of the panel. The colder regions (black being the coldest) represent the lower probabilities and the hotter regions represent higher probabilities.
(2) More complete Monte Carlo simulations to estimate the probability of exceeding the observed likelihood value. The user may specify the number of simulations to run; however, this may be a very lengthy operation. Typically for us 1,000,000 simulations require close to 2 h.
(3) Construction of expected normal cumulative distributions (from data parameters) for comparison with observed data. The theoretical cumulative curve is drawn over the existing curve giving a good idea of how well the data fits the expected.

(4) Output of the data base to a "mapmaker.raw" or "SAS" format.
(5) Display of the relation between two traits as a scatter plot with a numerical value for the correlation coefficient. All of the individuals are displayed with the value for the first trait along the y axis and the value from the second trait along the x axis. The user may select any of the traits for either of the axes.

The following automated searches are also available:
(6) A search of all pair-wise interactions in the data base which occur at a significance above a user-set threshold. Again this may be a lengthy operation depending upon the number of pairs to be checked; typically for us 20 min to search 20 000 pairs.
(7) A limited form of this search in which one locus is designated and all loci with which it interacts (above a pre-set likelihood threshold) are identified.
(8) Identification of linked (within 10 cM) pairs of loci.
(9) Search for data duplication (either a duplication of segregant lines, or of marker data).
(10) Identification of allele reversal (incorrect entry) for any pair of linked loci (within 10 cM).

Visual presentation of data is an important aspect of this program. At all times, the user remains in contact with the entire data set. Suggestive aberrations in the distributions, as in Figs. 3 and 4, are readily seen. Such visual displays can guide further analysis or lead to the design of future experiments.

Epistat runs in DOS 3.0 or higher. We use Epistat on several different systems from 386 to Pentium-100. The only features that require rapid processors are the utomated searches and the Monte Carlo simulations. For us a search of 20 000 pairs requires 20 min and a Monte Carlo simulation with 1 000 000 trials requires close to 2 h. For more information about how to get a copy of this program send e-mail to Chase@Bioscience.utah.edu

Acknowledgements This work was supported by grants to K.G.L. from the National Institute of Health (GM 42337). We thank J. Specht and W. Rayford for making NIT data available for our use. We also thank J. Specht and J. Orf for valuable comments during the preparation of the manuscript. Finally we thank Alex Kahler and three anonymous reviewers for their comments and a helpful critique of the original manuscript.

References

Bickel PJ, Doksum KA (1977) Mathematical statistics. Holden Day, Oakland, California

Dudley JW (1993) Molecular markers in plant improvements: manipulation of genes affecting quantitative traits. Crop Sci 33:660–668

Lark KG, Chase K, Adler F, Mansur LM, Orf J (1995) Interactions between quantitative trait loci in which trait variation at one locus is conditional upon a specific allele at another. Proc Natl Acad Sci USA 92:4656–4660

Manly BFJ (1991) Randomization and Monte Carlo methods in biology. Chapman and Hall, London New York Tokyo Melbourne Madras

Mansur LM, Orf JH (1995) Evaluation of soybean (*Glycine max* L.) recombinant inbreds for agronomic performance in Northern U.S. and Chile. Crop Sci 35:422–425

Mansur LM, Orf JH, Lark KG (1996) Genetic mapping of agronomic traits using recombinant inbred lines of soybean [*Glycine max* (L.) Merr.]. Crop Sci 36:1327–1336

Martin B, Nienhus J, King B, Shaffer A (1989) Restriction fragment length polymorphism associated with water-use efficiency in tomato. Science 243: 1725–1728

Tanksley SD (1993) Mapping polygenes. Annu Rev Genet 27:205–223

Williams LF (1952) The inheritance of certain black and brown pigments in the soybean. Genetics 37:208–215