e-Xtra*

In the Eye of the Beholder: The Effect of Rater Variability and Different Rating Scales on QTL Mapping

Jesse A. Poland and Rebecca J. Nelson

First author: Department of Plant Breeding and Genetics, and second author: Department of Plant Pathology and Plant Microbe Biology and Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853.
Current address of J. A. Poland: United States Department of Agriculture–Agricultural Research Service, 4011 Throckmorton Plant Science Center, Manhattan, KS 66506-5502.
Accepted for publication 3 October 2010.

ABSTRACT

Poland, J. A., and Nelson, R. J. 2011. In the eye of the beholder: The effect of rater variability and different rating scales on QTL mapping. Phytopathology 101:290-298.

The agronomic importance of developing durably resistant cultivars has led to substantial research in the field of quantitative disease resistance (QDR) and, in particular, mapping quantitative trait loci (QTL) for disease resistance. The assessment of QDR is typically conducted by visual estimation of disease severity, which raises concern over the accuracy and precision of visual estimates. Although previous studies have examined the factors affecting the accuracy and precision of visual disease assessment in relation to the true value of disease severity, the impact of this variability on the identification of disease resistance QTL has not been assessed. In this study, the effects of rater variability and rating scales on mapping QTL for northern leaf blight resistance in maize were evaluated in a recombinant inbred line population grown under field conditions. The population of 191 lines was evaluated by 22 different raters using a direct percentage estimate, a 0-to-9 ordinal rating scale, or both. It was found that more experienced raters had higher precision and that using a direct percentage estimation of diseased leaf area produced higher precision than using an ordinal scale. QTL mapping was then conducted using the disease estimates from each rater using stepwise general linear model selection (GLM) and inclusive composite interval mapping (ICIM). For GLM, the same QTL were largely found across raters, though some QTL were only identified by a subset of raters. The magnitudes of estimated allele effects at identified QTL varied drastically, sometimes by as much as threefold. ICIM produced highly consistent results across raters and for the different rating scales in identifying the location of QTL. We conclude that, despite variability between raters, the identification of QTL was largely consistent among raters, particularly when using ICIM. However, care should be taken in estimating QTL allele effects, because this was highly variable and rater dependent.

Beauty is in the eye of the beholder, as the old saying goes. One would hope, however, that for visual evaluation of quantitative disease resistance (QDR), the results and inferences would not be specific to the rater who assessed the disease severity or the scale that was used for rating. Due to the importance of QDR, hundreds of studies have been published on the mapping of quantitative trait loci (QTL) for disease resistance in plants (27,30,32,33). The plant populations used in these studies have been assessed almost exclusively using visual estimation of disease severity. Typically, populations were assessed by one or a few raters using several replications repeated over years, seasons, or environments (1,2). The results of those studies are only as accurate as the methods used to assess resistance and little is known about how rater variability and different rating scales influence detection of QTL.

QDR is an important objective in crop breeding programs because this type of resistance tends to be more durable than resistance conditioned by genes of large effect (27,29). For some pathosystems, QDR is the only available type of resistance. The development of cultivars with high levels of QDR remains challenging due to the polygenic nature of the phenotype and the small phenotypic effects of individual genes. Therefore, progress in resistance breeding requires accurate methods for assessing QDR. In the field of quantitative genetics, accurate disease assessments are important in providing precise measures of QTL positions and effects. Accurate identification of disease resistance QTL is essential for implementation of marker-assisted selection.

Accuracy and precision, two different measures of visual assessment of disease severity, have been analyzed in many studies (5,13,20,23,24). Accuracy is generally defined as the closeness of the visual estimate to the true level of disease. Precision is a measure of the repeatability of visual evaluations (22). Multiple different terms have been used to describe accuracy and precision in disease severity assessment and are described in detail by Bock et al. (7). Here, we used the term "precision" to describe the repeatability of disease assessments between and among raters, or among the ratings of a given individual. Our concern was the influence of different raters' perceptions on QTL mapping outcomes; therefore, we did not utilize an objective rating technique such as image analysis to determine "true" disease levels and do not make inferences about accuracy. We use the term "consistency" as a qualitative descriptor for the agreement in position and effect size of QTL identified through QTL mapping.

Many issues concerning visual ratings have been addressed in previous studies. By comparing visual evaluation of diseased leaves with image analysis of the same leaves, Parker et al. (25) found that levels of *Septoria tritici* leaf blotch and powdery mildew were not accurately or consistently estimated, even by experienced individuals. Overestimation was evident at low disease levels. The correlation between estimates of yield loss and disease severity for corky root in lettuce was found to vary with different qualitative and quantitative scales (24). The qualitative

Corresponding author: R. J. Nelson; E-mail address: rjn7@cornell.edu

*The *e*-Xtra logo stands for "electronic extra" and indicates that the online version contains three supplemental figures.

doi:10.1094/PHYTO-03-10-0087

scales were found to be most precise, whereas the quantitative scale correlated best with yield loss. For Phomopsis leaf blight of strawberry, the correlation among individual estimations and the actual disease severity was high for six different individuals using either the Horsfall-Barratt scale (14) (0.85 to 1.00) or direct percentage estimation (0.92 to 0.99), though the direct percentage estimation was more accurate (20).

Horsfall and Barratt (14) originally proposed a modified ordinal disease rating scale that was based on logarithmic increases in disease severity. Each unit increase in the ordinal scale represented a doubling in the disease severity. This scale was proposed because it was believed that the reference stimulus (i.e., amount of disease) needed to double for an individual to be able to detect a visual difference. It is now acknowledged, however, that linear increases in visual stimulus can be detected (21) and that direct estimation of disease severity is generally more accurate (13). Experience and training have been shown to improve the accuracy of disease assessment (6). Nutter et al. (22) showed that computer training programs that simulate diseased leaves can increase the accuracy and the precision of disease assessment ratings and that the use of a standard area diagram can also improve the accuracy of visual disease ratings.

Image analysis and remote sensing have been proposed and utilized as tools for accurately determining disease levels (7). Image analysis has been used to identify QTL for Gibberella stalk rot resistance, though the results were not compared with visual evaluation of the trait (26). Outside the field of plant pathology, image analysis has been used to analyze and map QTL for other traits such as kernel morphology (9) and flour-milling yield (3). With continual advancements in remote sensing, it is possible that this technology will replace visual assessment of disease severity in breeding programs and germplasm evaluation, providing both high-throughput phenotyping capacity as well as more accurate and precise disease measurements (7). However, past and current evaluation of disease resistance is conducted almost exclusively by visual assessment, warranting further investigation into the effects of rater variability.

Although many studies have analyzed the ability of individuals to give accurate and precise visual assessments of disease and have assessed differences among visual rating scales, there has been limited work to assess the impact of variability (i.e., precision) in visual disease estimation on the results and inferences made from such studies. There have been many reports of mapping disease resistance QTL in plants with the objectives of identifying chromosomal regions associated with resistance, identifying genes underlying this complex trait, or tagging QTL with molecular markers for selection in a breeding program. These studies have largely been conducted with a single rater using a single rating method (i.e., ordinal or direct percentage). The effects of rater variability and the use of different rating scales on the conclusions drawn from these studies have not, however, been determined. Therefore, the objectives of this present study were to (i) compare the precision of different raters in assessment of plots in a field nursery and examine how the variability among raters affects the results and interpretations of QTL mapping studies for disease resistance and (ii) compare the use of different rating scales on precision of disease assessment and QTL mapping.

To address these objectives, we utilized the maize–*Setosphaeria turcica* pathosystem. *S. turcica* (anamorph *Exserohilum turcicum*) is the causal agent of northern leaf blight (NLB), an economically important disease of maize (*Zea mays* L. spp. *mays*) throughout the world. The pathogen spreads locally through the plant vasculature, causing large necrotic lesions on the leaves. Through the Maize Diversity Project (www.panzea.org, www.maizegenetics.net), excellent genetic resources have been developed for the maize community, providing tools for study of quantitative traits in maize. The combination of well-designed maize populations, the economic importance of NLB, and the quantitative genetic expertise found in the maize community makes this pathosystem an excellent model system for studying QDR in plants.

MATERIALS AND METHODS

Plant materials. The maize nested association mapping population (NAM) is a set of 25 recombinant inbred line (RIL) families that were derived by crossing each of 25 diverse inbred lines with a common reference inbred line (8,19,34). From the 25 RIL families, the MS71 × B73 population was previously identified as having large variation in quantitative resistance to NLB and minimal variation for relative maturity (J. A. Poland, *unpublished data*). Minimal variation in relative maturity is beneficial because quantitative disease resistance in plants has been previously associated with maturity (10,12,31). Seed for the MS71 × B73 RIL population was generously supplied by E. Bucker (United States Department of Agriculture–Agricultural Research Service, Ithaca, NY). The population consists of 200 S_5 RILs, of which 191 were used for this study. The two inbred parents, MS71 and B73, were used as checks throughout the experiments. The population has been genotyped with 1,106 single-nucleotide polymorphism markers, of which 701 were polymorphic and used for mapping. Marker positions based on the NAM composite map were used (19).

Raters. Twenty-two individual raters volunteered to participate in the experiment in 2008, 2009, or both years. Raters included undergraduates, graduate students, and faculty in Plant Science, Plant Breeding and Genetics, and Plant Pathology and Plant–Microbe Biology at Cornell University, Ithaca, NY. Though not all had experience in plant disease rating, all had some experience in research related to plant biology. To gauge the amount of disease rating experience prior to the study, each rater was asked to assess their previous experience in two areas: (i) experience scoring plant disease in general and (ii) experience scoring NLB in maize. Experience levels were rated on a 1-to-5 scale, where 1 = no experience, 2 = little experience, 3 = some experience, 4 = experienced, and 5 = very experienced.

Field trials for northern leaf blight. Field trials were conducted during 2008 and 2009 at the Robert B. Musgrave Research Farm in Aurora, NY. Trials were planted on 14 May 2008 and 18 May 2009. Lines were planted as single-row plots 2.2 m in length, with 0.76 m between rows. Plots were overplanted and thinned to an average of 10 plants/row. Pre-emergence and post-emergence herbicide applications were applied each year. Trials were laid out in an augmented incomplete block design with one replication in 2007 and two replications in 2008. Each block consisted of 20 RILs and two checks (B73 and MS71). Artificial inoculation was conducted as described by Chung et al. (11). In brief, all plants were uniformly inoculated with *S. turcica* isolate NY001 at the six- to eight-leaf stage, which corresponded to 27 June 2008 and 16 July 2009. Spring 2009 was extremely cool and plant growth was very delayed compared with 2008. Each year, two types of inocula were simultaneously applied to every plant: (i) 2.5 to 3.0 ml of dried infected sorghum grains, previously inoculated and cultured for 2 weeks, and (ii) 1.0 ml of a spore suspension (1×10^3 conidia/ml) in H_2O with 0.02% Tween 20 cultured on lactose casein agar (LCA) plates for 2 to 3 weeks at room temperature under 12-h-light and 12-h-darkness conditions. Spores were harvested by flooding the plates with sterile H_2O, scraping with a glass rod, and filtering through cheese cloth. Spore concentrations were determined and the inoculum was diluted to 1×10^3 conidia/ml.

Phenotypic evaluation. Trials were visually evaluated for disease severity at two time points each year (ratings 1 and 2). In both years, disease severity was assessed as diseased leaf area defined as the percentage of total leaf area in the plot that was

covered by necrotic lesions from NLB (percentage scale). There was little senescence at the time of rating. Raters were advised, however, to exclude any tissue that was senescent rather than diseased. Chlorosis was not observed in this germplasm under NLB infection and there was no significant presence of secondary diseases. To evaluate the effect of rating scale on QTL identification, a second rating scale, a 0-to-9 logarithmic-based scale (0-to-9 scale), was also used in 2009. For the percentage rating, disease severity was estimated as the percentage of total leaf area necrotic with disease using a 0 to 100 percentage scale with 1% increments. For the 0-to-9 scale, a disease severity scale based on logarithmic increases in diseased leaf area was used (13) and additional semiqualitative descriptors were incorporated (Table 1). In 2008, four raters participated in the first rating and six raters participated in both ratings. In 2009, 12 raters evaluated the population at both ratings and four raters evaluated only the second. In 2009, four raters used both scales, while the remaining raters were randomly assigned one of the two scales. The raters, rating time-points evaluated, and scale used are listed in the supplemental data.

Data analysis. The primary objective of this study was to examine the precision of different raters in disease severity assessment and how this affected QTL mapping. We considered the correlation between replications and the correlation among different raters and the mean rating as measures of precision. Correlations between raters and between replications for individual raters were evaluated in R statistical software (28). Statistical tests based on correlation values were evaluated by conducting a *t*-test or fitting a general linear model (GLM) where noted. Because an independent measure of disease severity (e.g., image analysis) was not taken, the accuracy (correlation to true value) of visual assessments was not considered here.

Best linear unbiased predictions (BLUPs) for each line were determined using PROC MIXED in SAS statistical software (v9.1; SAS Institute Inc., Cary, NC). A single model was fit for each of the rating scales, incorporating all of the assessments from all raters, year, block, and rating time-point effects. The model solution gave BLUPs for each line. For reduced models where certain raters, years, ratings, or replications were to be analyzed, the random effect terms were left out accordingly.

Two methods for QTL mapping were employed. Stepwise GLM selection was conducted in SAS v9.1.3 using PROC GLMSelect (SAS Institute Inc.). The BLUPs from each respective rater, year, and rating combination (and the BLUPs from the full model) were assigned as the response variables, and model selection was conducted for marker effects. Marker effects were fit as a continuous variable consistent with an additive effects model. A selection threshold of P value = 0.001 was used for entry and removal of selected effect in the model. Model solutions were saved for the estimated effects of each selected marker in the model.

Inclusive composite interval mapping (ICIM) (18) was conducted using QGene v.4.2.3 (15). Cofactor selection was conducted with stepwise selection with a selection threshold of F = 3.0. The default value of 2 centimorgans was used for a scan interval. As with GLM, the BLUPs from each of the raters were mapped as a separate trait. Due to computational constraints, permutation analysis to determine an experimental significance threshold for each individual and rating was not conducted. A threshold of likelihood ratio = 3 (log of odds) was used as a general threshold for significant QTL.

RESULTS

Precision of disease ratings. We examined all possible pairwise correlations between raters for each of the years. All pairwise correlations examined between raters were significant at $P <$ 0.0001. In year 1, the correlations between raters (all using a percentage scale) were 0.72 to 0.91 for the first rating and 0.75 to 0.91 for the second rating. In year 2, the correlations between raters using a percentage estimate were 0.65 to 0.88 for the first rating and 0.70 to 0.93 for the second rating. For ratings using the ordinal scale in year 2, between-rater correlations were 0.59 to 0.73 for the first rating and 0.58 to 0.82 for the second rating. Though the percentage scale used continuous values of 1 to 100, there was a tendency to score using intervals of 5 (data not shown). This was particularly evident at higher disease levels.

To examine the precision of each of the different raters, the correlations between replications for the year 2 trial were examined for each rater (Table 2). The correlations varied from 0.409 to 0.959, showing large differences between raters. The

TABLE 2. Pearson correlation coefficients for individual raters in year 2[a]

Individual	Rating 1	Rating 2
Percentage scale		
2	0.6828	0.6919
3	0.7448	0.7269
4	0.8368	0.8435
5	0.7916	0.8067
7	0.8426	0.8719
8	0.6969	0.8594
11	0.6787	0.7651
14	NA	0.7809
18	0.6817	0.7856
19	NA	0.7719
21	0.6936	0.7273
Average	0.8819	0.8964
Ordinal scale		
1	0.4377	0.7308
2	0.5821	0.5642
3	0.409	0.5554
5	0.5407	0.7433
7	0.6507	0.7357
15	NA	0.6869
17	0.578	0.6657
20	NA	0.5732
Average	0.7353	0.8193

[a] Correlation between disease estimates for the two replications was determined for each rater for the first and second ratings. NA = not available.

TABLE 1. Ordinal rating scale used for assessment of northern leaf blight severity describing the disease severity classes and additional descriptors

Category	DLA (%)[a]	Additional descriptors
0	0	No lesions visible
1	>0–1	Few small lesions on lower leaves
2	>1–2	Several lesions on lower leaves
3	>2–5	Many lesions on lower leaves
4	>5–8	Coalescent lesions on lower leaves
5	>8–12	Lower leaves mostly blighted, few small lesions on middle leaves
6	>12–20	Lower leaves almost completely blighted, some lesions on middle leaves
7	>20–33	Lower leaves completely blighted, considerable lesions on middle leaves
8	>33–66	Lower leaves completely blighted, middle and ear leaf largely blighted
9	>66–100	Most to all of green leaf tissue blighted

[a] Diseased leaf area (DLA) percentage range.

average correlation when using the percentage scale was 0.764 whereas the mean correlation for individuals using the 0-to-9 scale was only 0.603 (Fig. 1). To test the effect of rating scale, rater, experience, and rating time-point on precision of disease estimates, a linear model was fit to the between-replication correlations. There was a significant effect of rating scale (P value < 0.0001), with the percentage scale being more precise than the 0-to-9 scale. More experience with scoring NLB resulted in increased precision (P value = 0.0009), whereas there was only a trend for general disease scoring experience (P value = 0.0967). There was also an increase in precision for the second disease rating (P value = 0.001).

Disease severity ratings are conducted after flowering for this pathosystem (no further host development); therefore, the disease severity of a line should be higher (or the same) at the later of two time-points. For raters who conducted two disease assessments, the number of lines that were assessed lower for the second rating was determined. There were large differences, with some raters assessing <1% of the lines lower on the second rating while other raters assessed >50% of the lines lower at the second rating.

To compare the direct percentage scale and the ordinal scale (with underlying percentage ranges), the percentage and 0-to-9 scores for the four individuals who evaluated with both scales were compared. Each class of the 0-to-9 scale was defined by an underlying percentage range. This range was used to determine whether the percentage assessment assigned to a given plot fell within the defined ranges of the 0-to-9 scale. The fraction of percent ratings outside the defined class ranges was 36 to 97%. For the first rating, three of the four raters had >70% of the percentage observations outside of their respective 0-to-9 class ratings. This general lack of agreement between the two scales indicates that comparison of results from different scales is particularly uncertain, even if an ordinal scale has a defined underlying percentage basis. Although the correlation between percentage and 0-to-9 ratings was high, there was a nonlinear relationship between the two scales (Fig. 2). Based on residual values from a linear fit of genotype (line), we observed that estimates of NLB disease severity using the direct percentage scale were heteroscedastic; that is, the variability of estimates increased with increasing disease severity. This same trend was not apparent in the 0-to-9 ratings.

Agreement of QTL identification among raters and between rating scales. To identify molecular markers associated with NLB resistance, stepwise GLM selection was conducted for each of the respective rater, year, rating, and scale combinations. The results show general agreement for QTL identification. Several QTL were identified across all raters, though there was discrepancy between individuals for the identification of QTL with small effect (Fig. 3). To best compare the difference between rating scales, the line BLUPs calculated using disease estimates from all raters were used for mapping with the percentage and 0-to-9 ratings scales and also a square root transformation of the percentage scale (Fig. 4). Because the percentage and 0-to-9 rating scales had different means and variances, standardized allele effects were used to compare the results from these two scales. Eleven QTL were identified for the 0-to-9 scale and the square-root percentage scale, while nine QTL were identified for the percentage scale. Eight of the QTL were identified across all three scales and the standardized allele effect estimates were roughly equivalent. Two QTL were identified using the 0-to-9 and square-root percentage scale, indicating that the detection of smaller-effect QTL might be sensitive to the type of scale and the resulting distribution. Three additional QTL were found by mapping using only one rating scale; these might be sensitive to the trait distribution or false positives. The same trend was seen when ratings from different raters were used for mapping; the large-effect QTL were identified across all raters, whereas small-effect QTL were identified only by some.

The GLM solutions gave estimates of selected QTL marker effects as well as standard errors for those estimates. These effect estimates are analogous to the allele effect at that locus. The estimated effects of different raters for the percentage scale in year 2 were compared and showed significant differences for each of the loci that were identified by multiple individuals (Fig. 5). To determine whether this was an effect of different variances for the

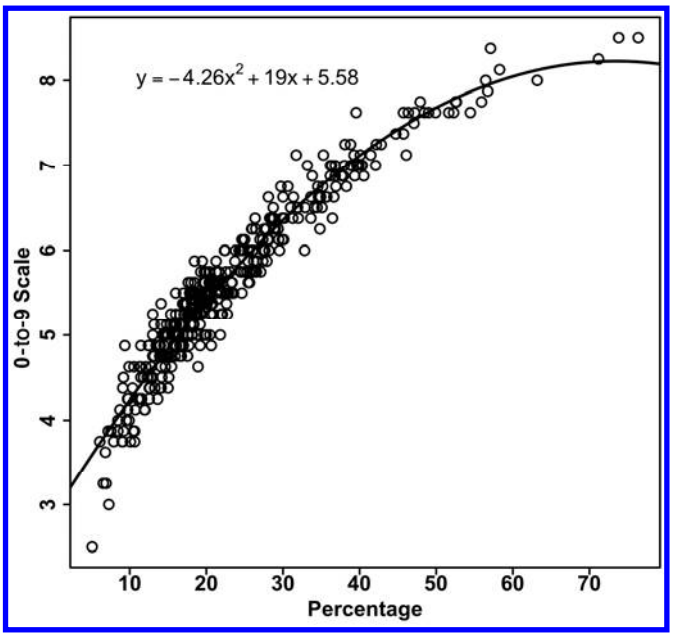

Fig. 1. Plot of disease severity estimates using either direct percentage rating or 0-to-9 ordinal scale shows a nonlinear relationship between the two scales. Data are plotted for each recombinant inbred line using phenotypic scores from the average of all individuals using either a 0-to-100 percentage scale or a 0-to-9 ordinal scale. A second-degree polynomial was fit (R^2 = 0.938). The log-based classes in the 0-to-9 scale are evidenced in the nonlinear relationship.

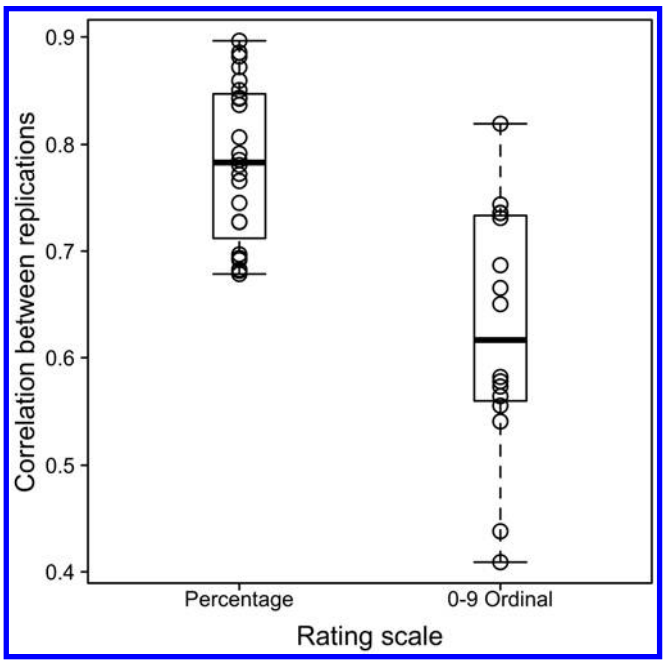

Fig. 2. Box and whisker plot showing higher precision results from using percentage scale than ordinal scale. Data are based on within-rater correlation between replications in year 2. Correlation values for individual raters are shown as circles with the box plot showing the 25th and 75th percentiles. The difference between the means of the two rating scales is significant (P value < 0.0001).

trait distributions among raters, standardized estimates were compared. These values were less variable among raters though significant differences remained (Fig. 6).

ICIM produced very consistent results across different rating scales. The BLUPs of all individuals were used to compare the different rating scales for ICIM, which produced almost identical results in terms of QTL position and significance of identified QTL effects (Fig. 7). This is despite a nonlinear trend between the percentage and the 0-to-9 scale (Fig. 2). For ICIM across all individuals, the positions of significant QTL were largely consistent, though the height of the peaks varied, indicating differences in the power of QTL detection among individuals (Fig. 8).

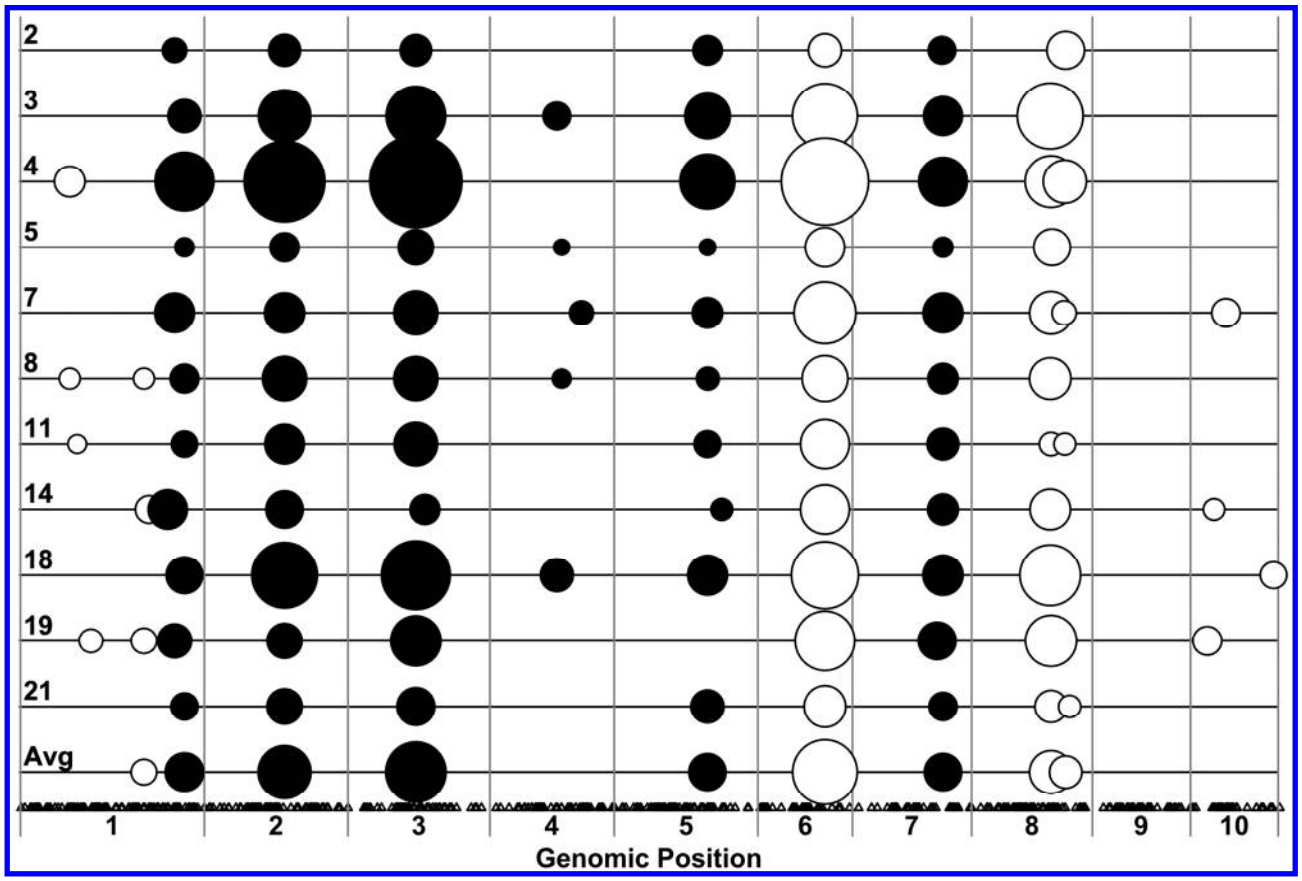

Fig. 3. Genomic position and relative effect of quantitative trait loci (QTL) identified from stepwise general linear model selection for each rater using the percentage scale. Direct percentage estimates of disease severity were used to detect resistance QTL for each rater. Genomic positions of QTL identified using stepwise model selection are shown along the horizontal lines, with vertical lines separating chromosomes. Each horizontal line corresponds to a single rater (as noted by a number) or the results using average estimates from all raters (Avg). Relative effect size of identified QTL is represented by the size of the circle. Solid circles represent resistance from MS71 while open circles represent resistance from B73. The position of molecular markers is shown by black triangles along the bottom.

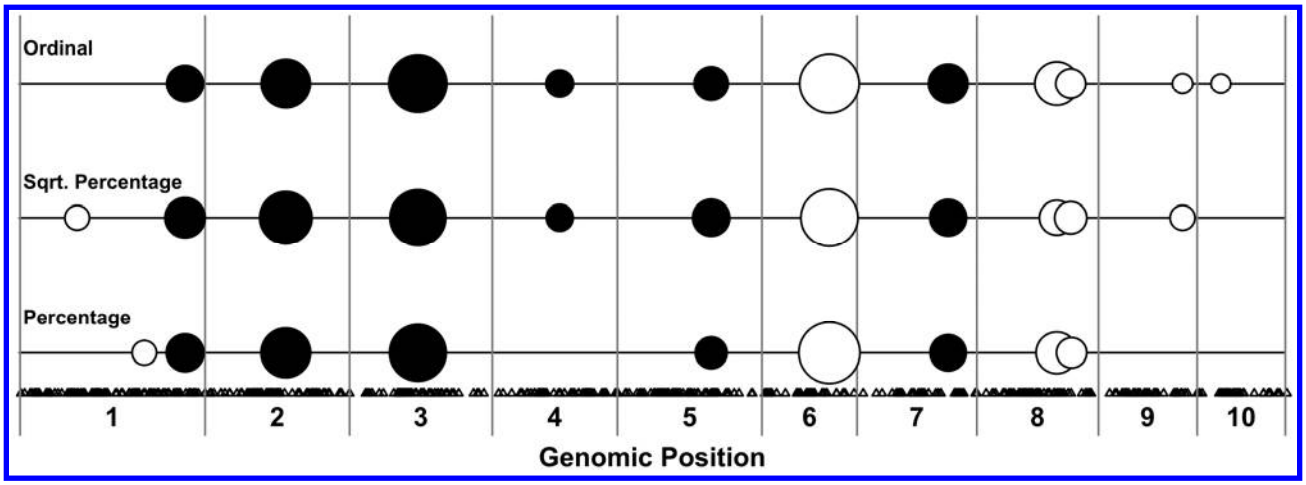

Fig. 4. Genomic position of quantitative trait loci (QTL) identified from stepwise general linear model selection using the percentage and 0-to-9 ordinal rating scales. Average disease assessments from all raters using either a direct percentage estimated of disease severity or an ordinal 0-to-9 rating scale were used to identify resistance QTL using stepwise linear model selection. A square root transformation of the percentage data is also included. Genomic positions of QTL identified are shown along the horizontal lines, with vertical lines separating chromosomes. The relative effect size of identified QTL is represented by the size of the circle. Solid circles represent resistance from MS71 while open circles represent resistance from B73. The position of molecular markers is shown by black triangles along the bottom.

DISCUSSION

Quantitative disease resistance remains an important area of research in plant breeding and phytopathology (17,27). To understand the underlying genetics of quantitative disease resistance, segregating populations have been used to map QTL. In virtually all of these studies mapping QTL, visual assessment has been used to determine disease severity. Although it remains unknown how differences among raters and rating scales affect the QTL mapping results, various resource-intensive endeavors are undertaken based on the conclusions of these studies, including research and product-development projects such as map-based cloning and marker-assisted selection of QTL in breeding programs. Therefore, it is pertinent to have a better understanding of how variability among raters and the use of different rating scales affect QTL mapping results.

We used the correlation between replications and the correlation among raters as indicators of precision (22). The correlations observed between raters were consistent with or slightly lower than has been seen in previous studies (13,20,23). The lower level of precision could be attributable to diseased field plots being more difficult to assess than images of single leaves or difficulty in precisely scoring disease severity in this pathosystem. There were large differences among raters as well as between rating scales used. We used a simple 1-to-5 scale to measure rater experience for general plant disease assessment and also for NLB, the specific pathosystem under study. There was a significant increase in precision with higher NLB rating experience and a slight trend for general experience. Several previous studies have also observed the trend of improved precision (and accuracy) for raters with more experience (7,13,22).

It was also observed that over half of the plots were estimated to have a lower disease severity at the second rating, which was interpreted as an indication of imprecision in disease ratings. Although in principle, this could have resulted from host growth or leaf senescence, it is more likely to be the result of imprecise rating because the experiment was evaluated after flowering (no further development of host tissue) and there was minimal leaf senesce between rating time-points. A lower score for the second rating suggests that either the first rating was overestimated, the second rating was underestimated, or both.

There was a nonlinear trend between the direct percentage scale and the 0-to-9 ordinal scale, consistent with the log-based classes defined for the ordinal scale. Though there is limited power in this

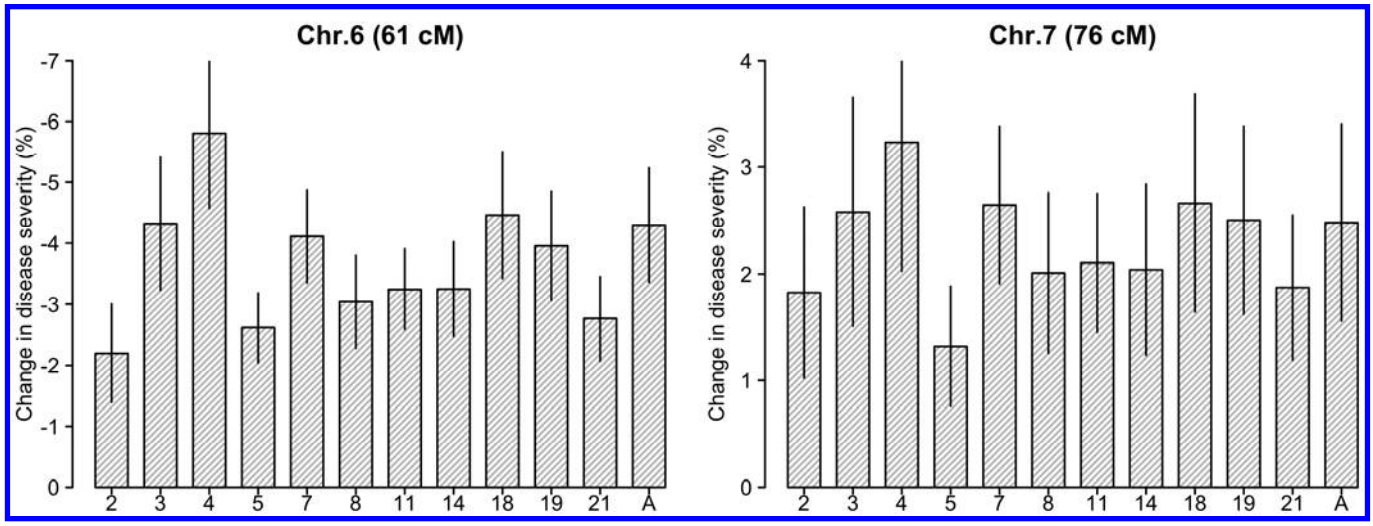

Fig. 5. Bar graph showing large variability in the estimated effect size of quantitative trait loci (QTL). Estimated allele effects at two QTL identified by all raters using the percentage scale are shown by the height of the bars. Allele effects are shown as the percentage increase or decrease in disease conditioned by the MS71 allele based on the ratings for each rater (numbered) and the average of all raters (A). Confidence intervals (95%) are shown.

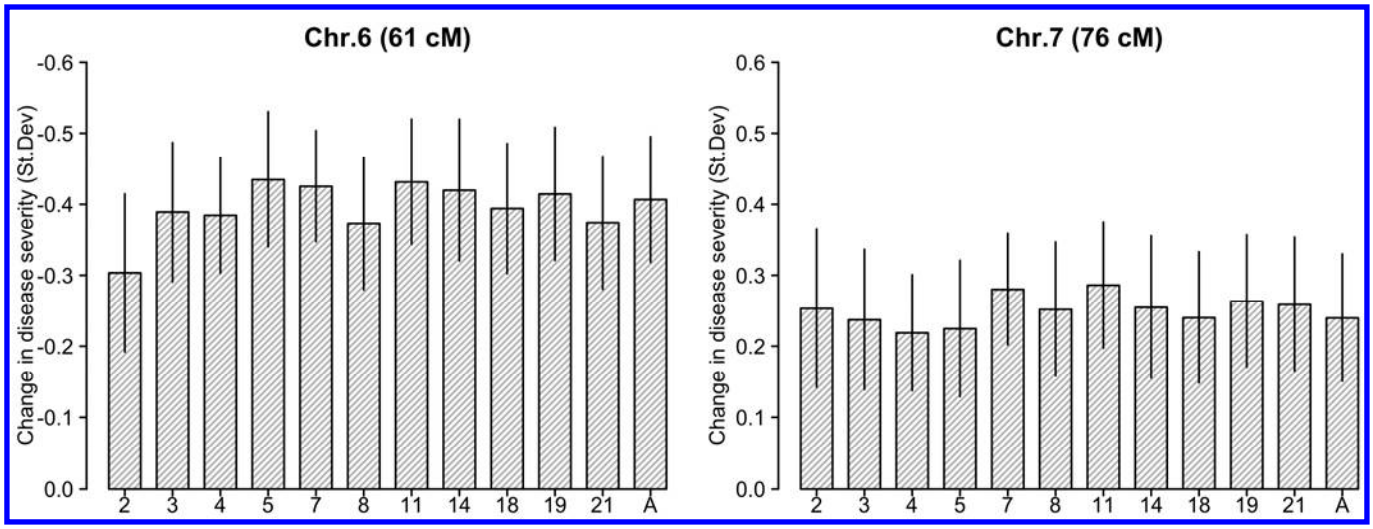

Fig. 6. Bar graph showing standardized estimated allele effects are more consistent for raters using the percentage scale at the same quantitative trait loci as in Figure 5. Estimated allele effects are shown as the increase or decrease in disease conditioned by the MS71 allele for each rater (numbered) and the average of all raters (A). Units represent standard deviations on total variance for each individual. Confidence intervals (95%) are shown.

study to compare rating scales (due to a limited number of raters using each scale), we were able to detect a significant increase in precision with the direct percentage estimates. Comparison of disease assessments using the 0-to-9 scale versus the percentage scale showed that scoring with the direct percentage scale was significantly more precise than using the 0-to-9 scale after accounting for rater experience and rating time-point. Although only measuring precision here, this is in agreement with previous studies that found higher accuracy when using a direct percentage scale than an ordinal scale (4,13). The proposed reasons for higher accuracy with direct percentage estimates are several, and include difficulty in transferring from a percentage scale (observed) to an ordinal scale, estimation error compounded with rounding error imposed by the classes, and difficulty in maintaining consistency in classes (13).

The percentage and 0-to-9 scales were correlated but there was little direct agreement between them, even though the 0-to-9 scale

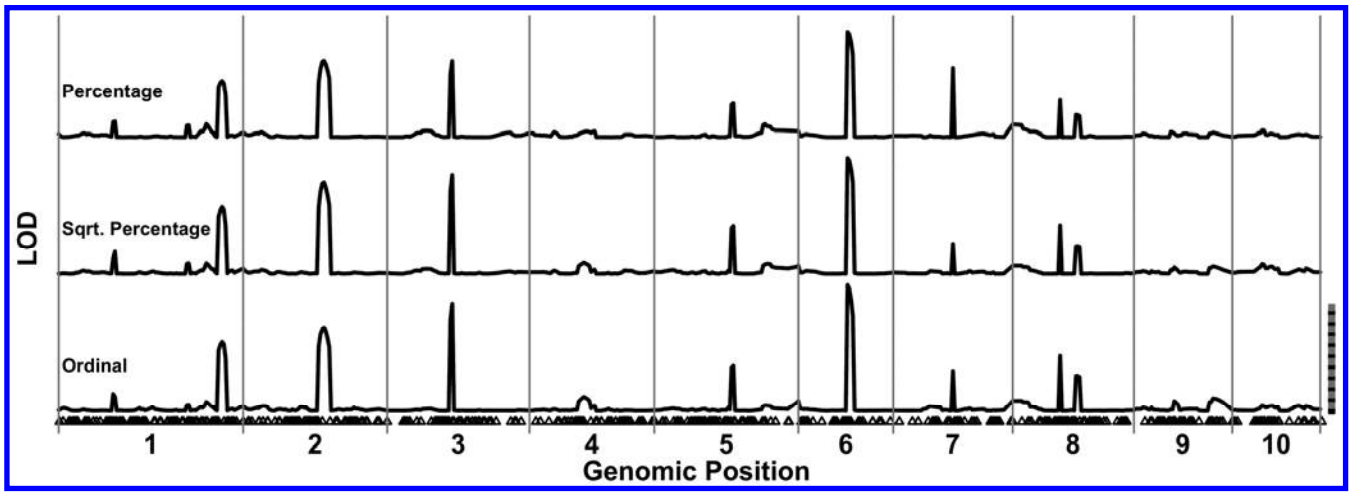

Fig. 7. Quantitative trait loci (QTL) profile for inclusive composite interval mapping using different rating scales. The best linear unbiased prediction from the mixed model using data from all raters for the two respective scales (direct percentage and 0-to-9 ordinal) as well as a square root transformation of the percentage values were used for mapping. Shown on the x-axis is the genomic position in centimorgans (cM) of the identified QTL. Chromosomes are marked by vertical gray lines and numbered along the x-axis. Positions of markers used in mapping are shown as black triangles. The QTL profile is shown as the likelihood ratio (LOD). The black and gray bar at right gives the scale of LOD = 20 with increments of 1.

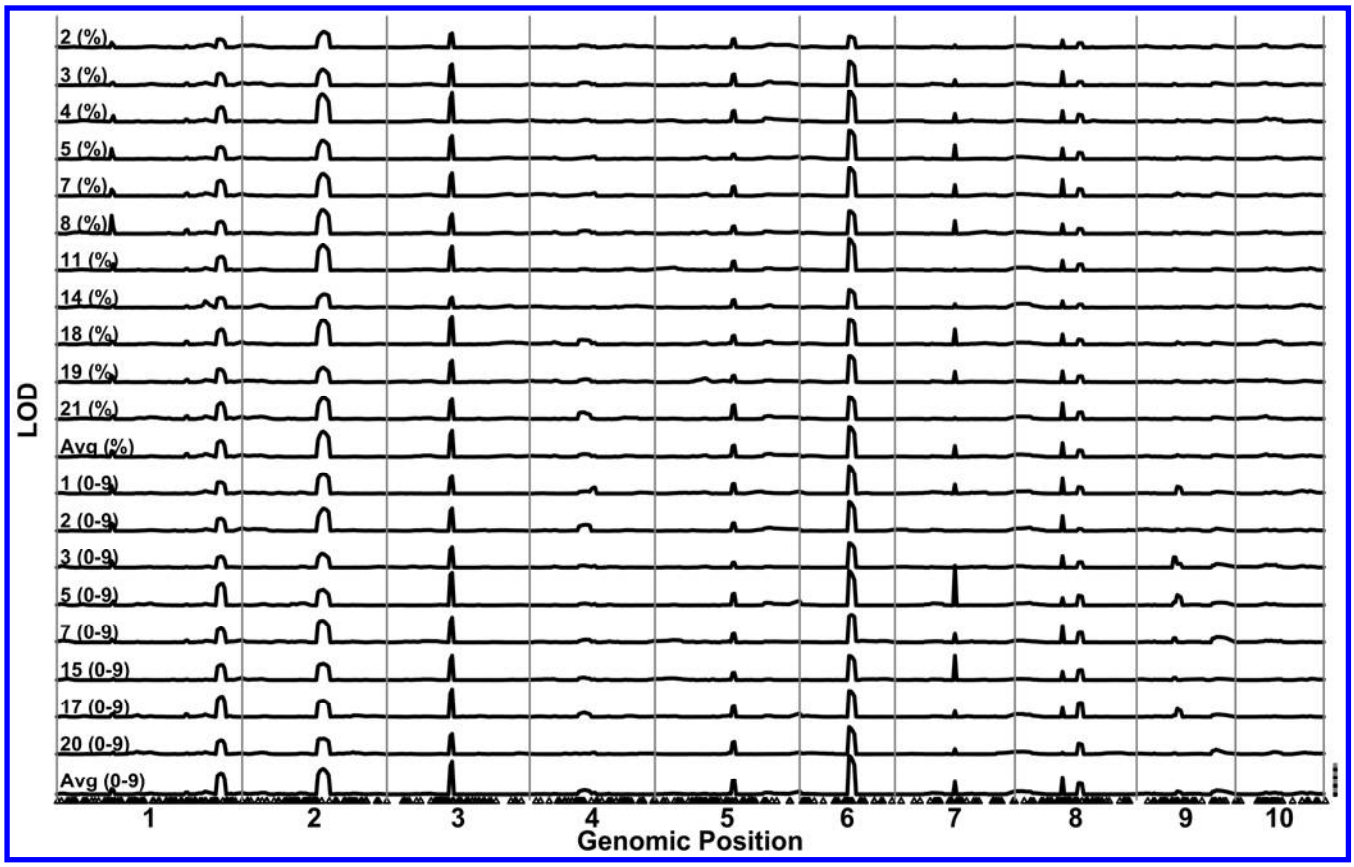

Fig. 8. Quantitative trait loci (QTL) profile for inclusive composite interval mapping using disease estimates from different raters. Raters are numbered to the left of each QTL profile and the scale used is shown in parentheses. The x-axis is the genomic position in centimorgans (cM) of the identified QTL. Chromosomes are marked by vertical gray lines and numbered along the x-axis. Positions of markers used in mapping are shown as black triangles below the plot. The QTL profile is shown as the likelihood ratio (LOD). The black and gray bar at right gives the scale of LOD = 20 with increments of 2.5.

was specified with an underlying percentage scale. For three of the four raters who evaluated with both scales, fewer than half of the direct percentage assessments were within the respective range of the assigned class from the 0-to-9 scale. This agrees with findings that transferring ratings from an ordinal scale to a percentage scale can be unreliable, even if the ordinal scale was developed with a defined underlying percentage scale (16).

For QTL mapping, ICIM was robust to differences among individuals and rating scales in identifying the position and significance of QTL. Using the line BLUPs from all individuals, the ICIM mapping results from percentage and 0-to-9 ratings were almost identical. As measured by correlation between replications, the line BLUPs using estimates from all individuals were the most precise assessments. This indicates that, for precise phenotypic values, the type of scale used should not have an effect on the position and significance of identified QTL. The consistency of the ICIM results from the percentage and 0-to-9 scales was in contrast to the nonlinear relationship of these two ratings. The consistency of ICIM was further observed in the individual ratings because most individuals, regardless of the rating scale used, identified the same QTL.

Although there was agreement among raters for the positions of identified QTL, the estimated additive effects of those QTL was highly variable. At some QTL that were identified by all raters, there were threefold differences in the estimated effects. This variation in allele effect estimates is largely based on the population variance from the individual raters. Raters who tended to score using a larger range of phenotypic values (higher population variance) had estimated QTL effects that were larger. When allele effects were standardized, the estimated allele effects were more consistent, though significant differences remained among raters.

It is our observation that, for QTL mapping using visual observations of disease severity, precision of the disease estimates has the greatest effect on power to detect QTL. Because all raters were fairly precise, there was general agreement among raters for the detection of QTL. The magnitude of the estimated allele effects, however, was highly variable among raters. Most previous QTL mapping studies should be considered to have high precision because multiple seasons and replications were generally used in the disease resistance evaluation, leading to accurate assessments of the genotype means. In this regard, the identified QTL can be considered reliable. However, the accuracy of these studies is not known (and cannot be known) and, hence, the estimated allele effects of identified QTL are likely specific to the individual study and the rater who conducted the disease assessments.

ACKNOWLEDGMENTS

This work was supported by The McKnight Foundation and The Generation Challenge Program. We thank each individual who conducted disease evaluations and made this study possible: P. Balint-Kurti, C.-L. Chung, S. Collier, A. Fialko, E. Heffner, L.-S. Hsieh, T. Jamann, K. Kennedy, J. Kolkman, K. Lyons, P. Manosalva, A. Mello, S. Mideros, O. Ott, C. Schweighofer, S. Sen, F. Tian, N. Zhang, Z. Zhang, and P. Zuluaga; E. Buckler for providing the seed for the B73 × MS71 RIL population, which was developed as part of NSF Maize Diversity Project; N. Lepak for help with the seed stocks; and two anonymous reviewers for the helpful comments that greatly improved the quality of this manuscript.

LITERATURE CITED

1. Balint-Kurti, P. J., Zwonitzer, J. C., Pe, M. E., Pea, G., Lee, M., and Cardinal, A. J. 2008. Identification of quantitative trait loci for resistance to southern leaf blight and days to anthesis in two maize recombinant inbred line populations. Phytopathology 98:315-320.
2. Baumgarten, A., Suresh, J., May, G., and Phillips, R. 2007. Mapping QTLs contributing to *Ustilago maydis* resistance in specific plant tissues of maize. Theor. Appl. Genet. 114:1229-1238.
3. Berman, M., Bason, M. L., Ellison, F., Peden, G., and Wrigley, C. W. 1996. Image analysis of whole grains to screen for flour-milling yield in wheat breeding. Cereal Chem. 73:323-327.
4. Bock, C., Gottwald, T., Parker, P., Cook, A., Ferrandino, F., Parnell, S., and van den Bosch, F. 2009. The Horsfall-Barratt scale and severity estimates of citrus canker. Eur. J. Plant Pathol. 125:23-38.
5. Bock, C. H., Parker, P. E., Cook, A. Z., and Gottwald, T. R. 2008. Characteristics of the perception of different severity measures of citrus canker and the relationships between the various symptom types. Plant Dis. 92:927-939.
6. Bock, C. H., Parker, P. E., Cook, A. Z., Riley, T., and Gottwald, T. R. 2009. Comparison of assessment of citrus canker foliar symptoms by experienced and inexperienced raters. Plant Dis. 93:412-424.
7. Bock, C. H., Poole, G. H., Parker, P. E., and Gottwald, T. R. 2010. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. Crit. Rev. Plant Sci. 29:59-107.
8. Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., Peiffer, J. A., Rosas, M. O., Rocheford, T. R., Romay, M. C., Romero, S., Salvo, S., Villeda, H. S., Sofia da Silva, H., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M. D. 2009. The genetic architecture of maize flowering time. Science 325:714-718.
9. Campbell, K. G., Bergman, C. J., Gualberto, D. G., Anderson, J. A., Giroux, M. J., Hareland, G., Fulcher, R. G., Sorrells, M. E., and Finney, P. L. 1999. Quantitative trait loci associated with kernel traits in a soft × hard wheat cross. Crop Sci. 39:1184-1195.
10. Ceballos, H., Deutsch, J. A., and Gutierrez, H. 1991. Recurrent selection for resistance to *Exserohilum turcicum* in eight subtropical maize populations. Crop Sci. 31:964-971.
11. Chung, C., Jamann, T., Longfellow, J., and Nelson, R. 2010. Characterization and fine-mapping of a resistance locus for northern leaf blight in maize bin 8.06. Theor. Appl. Genet. 121:205-227.
12. Collins, A., Milbourne, D., Ramsay, L., Meyer, R., Chatot-Balandras, C., Oberhagemann, P., De Jong, W., Gebhardt, C., Bonnel, E., and Waugh, R. 1999. QTL for field resistance to late blight in potato are strongly correlated with maturity and vigour. Mol. Breed. 5:387-398.
13. Hartung, K., and Piepho, H. P. 2007. Are ordinal rating scales better than percent ratings? A statistical and "psychological" view. Euphytica 155:15-26.
14. Horsfall, J. G., and Barratt, R. W. 1945. An improved grading system for measuring plant disease. (Abstr.) Phytopathology 35:655.
15. Joehanes, R., and Nelson, J. C. 2008. QGene 4.0, an extensible Java QTL-analysis platform. Bioinformatics 24:2788-2789.
16. Koch, H., and Hau, B. 1980. Ein psycholo-gischer aspect beim schatzen von pflanzenk-rankheiten. Z. Pflanzenkrankh. Pflanzenschutz 87:587-593.
17. Kou, Y., and Wang, S. 2010. Broad-spectrum and durability: Understanding of quantitative disease resistance. Curr. Opin. Plant Biol. 13:181-185.
18. Li, H., Ye, G., and Wang, J. 2007. A modified algorithm for the improvement of composite interval mapping. Genetics 175:361-374.
19. McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., Pressoir, G., Romero, S., Rosas, M. O., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J. C., Goodman, M., Ware, D., Holland, J. B., and Buckler, E. S. 2009. Genetic properties of the maize nested association mapping population. Science 325:737-740.
20. Nita, M., Ellis, M. A., and Madden, L. V. 2003. Reliability and accuracy of visual estimation of Phomopsis leaf blight of strawberry. Phytopathology 93:995-1005.
21. Nutter, F., and Esker, P. 2006. The role of psychophysics in phytopathology: The Weber–Fechner law revisited. Eur. J. Plant Pathol. 114:199-213.
22. Nutter, F., Esker, P., and Netto, R. 2006. Disease assessment concepts and the advancements made in improving the accuracy and precision of plant disease data. Eur. J. Plant Pathol. 115:95-103.
23. Nutter, F. W. J., Gleason, M. L., Jenco, J. H., and Christians, N. C. 1993. Assessing the accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment systems. Phytopathology 83:806-812.
24. O'Brien, R. D., and van Bruggen, A. H. C. 1992. Accuracy, precision, and correlation to yield loss of disease severity scales for corky root of lettuce. Phytopathology 82:91-96.
25. Parker, S. R., Shaw, M. W., and Royle, D. J. 1995. The reliability of visual estimates of disease severity on cereal leaves. Plant Pathol. 44:856-864.
26. Pè, M. E., Gianfranceschi, L., Taramino, G., Tarchini, R., Angelini, P., Dani, M., and Binelli, G. 1993. Mapping quantitative trait loci (QTLs) for resistance to *Gibberella zeae* infection in maize. Mol. Gen. Genet.

241:11-16.
27. Poland, J. A., Balint-Kurti, P. J., Wisser, R. J., Pratt, R. C., and Nelson, R. J. 2009. Shades of gray: the world of quantitative disease resistance. Trends Plant Sci. 14:21-29.
28. R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
29. St. Clair, D. A. 2010. Quantitative disease resistance and quantitative resistance loci in breeding. Annu. Rev. Phytopathol. 48:247-268.
30. Wang, G. L., Mackill, D. J., Bonman, J. M., McCouch, S. R., Champoux, M. C., and Nelson, R. J. 1994. RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. Genetics 136:1421-1434.
31. Welz, H.-G., and Geiger, H.-H. 2000. Genes for resistance to northern corn leaf blight in diverse maize populations. Plant Breed. 119:1-14.
32. Wisser, R. J., Sun, Q., Hulbert, S. H., Kresovich, S., and Nelson, R. J. 2005. Identification and characterization of regions of the rice genome associated with broad-spectrum, quantitative disease resistance. Genetics 169:2277-2293.
33. Wisser, R. J., Balint-Kurti, P. J., and Nelson, R. J. 2006. The genetic architecture of disease resistance in maize: a synthesis of published studies. Phytopathology 96:120-129.
34. Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. 2008. Genetic design and statistical power of nested association mapping in maize. Genetics 178:539-551.