Graphical Data Presentation, with Emphasis on Genetic Data

Brian S. Yandell[1]

Horticulture and Statistics Departments, University of Wisconsin–Madison, 1300 University Avenue, MSC-room 1239, Madison, WI 53706

Well-designed graphics can guide the initial exploration of data and can show results of formal inference. However, too often we rely on "cute" graphics designed for the corporate boardroom that hide relationships and bias interpretation. Graphics can help present *all* the data in revealing ways for modest-size experiments. Many of today's experiments involve massive amounts of data and many questions that can productively be condensed into graphics that summarize detailed patterns and guide further investigation. This paper is based on an invited talk at the ASHS 2004 annual conference.

Bottom line: A picture is worth a thousand words. This overused phrase applies quite strongly to scientific data presentation. Graphics help us to spot unusual patterns in preliminary investigation and later to present key relationships in a concise and convincing manner.

Large tables of numbers may be important for documenting certain studies, but they are almost always better hidden from view. It is too easy for the eye to gravitate to spurious patterns and miss the main story. Such visual confusion can bias our perception and sidetrack useful research, providing misleading clues that can waste time and money. Why not organize material creatively into pictures that show relationships *objectively?* Graphics can even include cues to inference, such as SE or LSD bars that show the extent of variation relative to mean tendency.

Creative graphics are not always useful; on the contrary, they can be quite harmful. For numerous examples in a variety of settings, see Wainer (1984) and the books of Edward Tufte (1983, 1990, 1997). The classic *How to Lie with Statistics* by Huff (1993) contains many examples of poor data analysis, often illustrated or "enhanced" with misleading graphics.

Excellent examples of graphics can be found in Tufte (1983, 1990, 1997) and Cleveland (1993, 1994). Both Tufte (1983) and Wainer (1984) show a famous diagram of Napoleon's defeat march across Europe, including details of geography, troop size, and weather conditions. Many of the best graphic examples are decades or centuries old. Although many are from designed experiments, others are used in everyday life, such as maps of transit systems. Cleveland (1993, 1994) focuses on more experimental settings, presenting a wide array of excellent graphical devices.

We illustrate some good and bad portrayals using plant breeding data from a *Brassica napus* L. cross of Tom Osborn. My contact with this work originated with Ferreira et al. (1995), with the map fully developed in Kole et al. (2002). These data are freely available as part of my R package bim (discussed later).

SPREADSHEETS TO ORGANIZE, NOT GRAPH, DATA

Spreadsheets tend to have pull-down menus with confusing palettes presenting techniques to graph data. They are often dominated by business-style graphics such as pie charts and bar plots that can severely bias perception of relationships. (Trained data analysts basically avoid pie charts, and only use bar plots in limited situations.) Spreadsheets are great for data entry and data organization, and are quite useful for data management over the long term, but they do not model graphical presentation well. The worst examples of spreadsheet graphics involve three-dimensional (3-D) renderings. Although 3-D diagrams have their place, mostly for geographical and high-dimensional data, they should be used sparingly.

Beyond the ill-advised array of graphics choices, the chief difficulty with spreadsheets is documenting and reproducing methods of data manipulation and calculation. These tend to involve many operations by hand-using mouse devices, with calculations hidden within "cells" of a spreadsheet, available only by "mousing" over the cell. Quite often, these rely on hard-coded values or location-sensitive references to other cells. Graphics are often tied to particular subsets of rows and columns on a sheet, and may involve considerable artistic modifications before completion. Thus, it can be difficult to reproduce graphics on a given data set, or to repeat these graphics and accompanying analysis on new data.

Statistical methods provided with spreadsheets tend to have limited capabilities and are typically unable to handle unbalanced data or data with multiple factors, nesting, or blocking. Experiments of any complexity tend to need tools found in a full-featured statistical package. Although many statistical packages can embed menus in spreadsheets for easy data transfer, the full power of a statistical package is best used by importing spreadsheet-prepared data.

STATISTICAL COMPUTATION AND GRAPHICS

Statistical packages are often used in conjunction with a graphics or presentation package, to prepare graphics for talks and publication. The choice of packages can enhance or limit our ability to present data well. There are now many reliable statistics packages available, including SAS (www.sas.com), SigmaPlot (www.sigmaplot.com), STATA (www.stata.com), and Minitab (www.minitab.com). Each of these offer a myriad of graphical choices—some good, some bad.

I illustrate good graphics using the R language/environment for statistical computing and graphics. The R system (Ihaka and Gentleman 1996) (www.r-project.org) is freely available worldwide for today's standard computing platforms. Free, open-source software is now quite respectable, and is arguably better maintained than many commercial packages. It is being used more and more for data analysis by ecologists (Kangas 2004), sociologists, breeders and geneticists, and bioinformatic scientists (www.bioconductor.org). There are many introductory resources, notably that of Dalgaard (2002). Several statistical methods courses aimed at biologists now use R for teaching, including our "Stat 571" at the University of Wisconsin–Madison (www.stat.wisc.edu/~yandell/st571). Because R code is open source, you can examine the details of computations after a bit of digging.

The primary drawback of R for biologists has been the lack of a good graphical user interface. R has traditionally relied on command line interface, which requires us to type or cut-and-paste commands. The advantage of this is that we can document analysis and graphics in ways that are reusable and can be readily shared with other scientists. However, it does require a modest effort to get started, which is best done by learning from examples that create analysis and graphics similar to your needs.

R is a structured language. Thus, it is possible to include "meta-data" to document the

This work was supported in part by U.S. Department of Agriculture/Hatch grants through the University of Wisconsin–Madison.

I thank ASHS for inviting me to speak at the 2004 meeting in Austin, TX. In addition, I thank the laboratories of Irwin Goldman, Jiming Jiang, Shawn Kaeppler, Tom Osborn, and Phil Simon for collaborations, use of data, and stimulating discussions.

[1] To whom correspondence should be addressed; e-mail yandell@stat.wisc.edu

characteristics of an experiment as part of a data "object" for direct use in analysis. Furthermore, it is possible to create interactive packages that allow others to reproduce the analysis and graphics. A companion package for this paper, including access to the data and commands for nonspreadsheet plots, is available at www.stat.wisc.edu/~yandell/software.

DESCRIPTIVE STATISTICS AND DATA EXPLORATION

Descriptive statistics are useful initial summaries of data to get a feel for patterns. In addition, they can be the basis for formal tests. However, overly adorned graphs merely take up space. Consider the following table:

	ec2d1a		
E33M59.59	M	S	Total
M	24	16	40
S	32	32	64
Total	56	48	104

These are the counts of how many of 104 plants had a particular combination of Major (M) or Stellar (S) genotype at two genetic markers: ec2d1a and E33M59.59. (Missing data have been filled in with the most probable genotype for convenience.) Both these markers are located on chromosome 2 of *Brassica napus*. Fig. 1 shows pie charts and bar plots for these data. Pie charts and bar plots are highly biased by choice of color, shading, and the order of entries. For instance, we perceive dark shades as bigger, and red as closer than blue (see books by Cleveland and Tufte for details). Pie charts should be avoided at all costs because they are so strongly subject to bias. Bar plots use substantial space for little content, with bars grounded at zero. Zero usually means a plant is dead! Although zero is meaningful for germination, it is probably not for days to flower or many physiological measurements. Bar plots should be used in limited situations only. Adding error bars to bar plots throws good money after bad.

What if we produced pie charts or bar plots for every pair of markers?! The primary interest in the table lies in the recombination rate between the markers, which is crudely estimated by the number of recombinant (off-diagonal) plants over the total: $r = 48/104 = 0.46$. This can be improved by using information along the linkage map where data are missing to get $r = 0.49$. These markers show no evidence of linkage, summarized by a log of odds (LOD) score of 0.0027. The recombination rates among pairs of markers, along with the LOD score to assess the strength of linkage, can be displayed compactly in a single plot (Fig. 2). This allows quick visual inspection of patterns, to locate obvious anomalies. For instance, chr N7 has high linkage among many markers, reflecting a dense marker map. There appear to be no major marker order problems in these data, which is reassuring because this linkage map has been developed over several years of careful work (Kole et al., 2002).

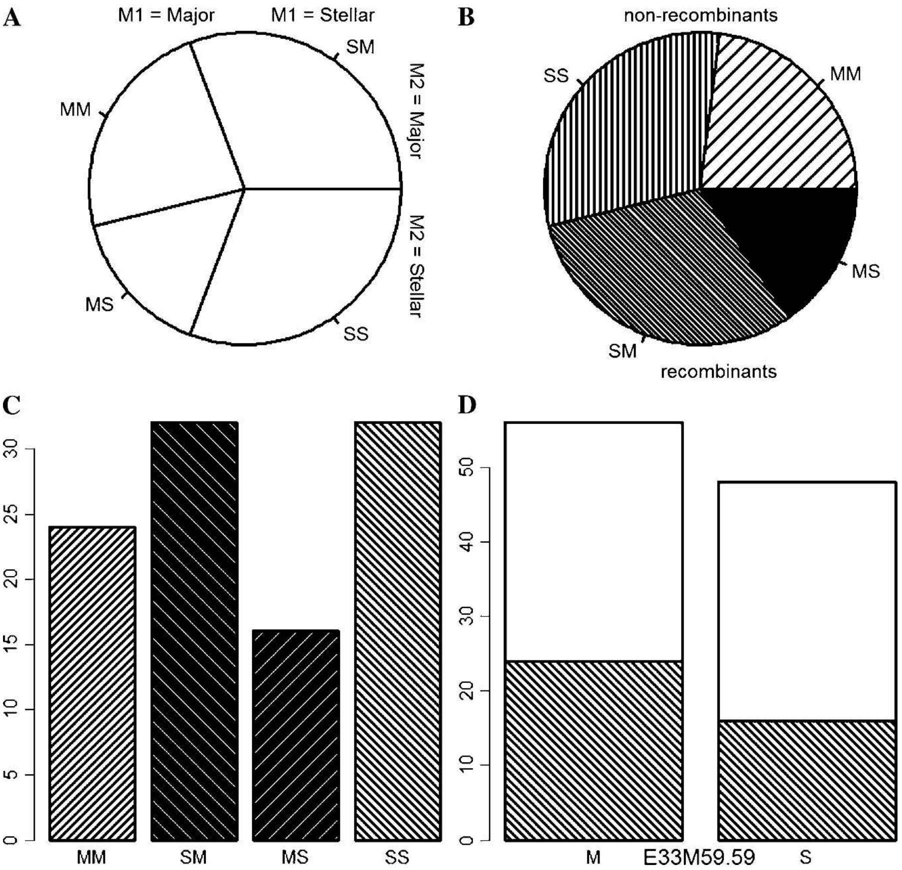

Fig. 1. (**A–D**) Pie charts and bar plots bias interpretation by choice of colors, order and perspective. It is difficult in view **A**, organized by markers, to assess that roughly 50% of individuals are recombinants, indicating no linkage. This could be seen in view **B**, except for the distraction of shading and angles. Bar plots in views **C** and **D** show two less-than-helpful ways to organize the same data. Three-dimensional effects from spreadsheets can further obscure results.

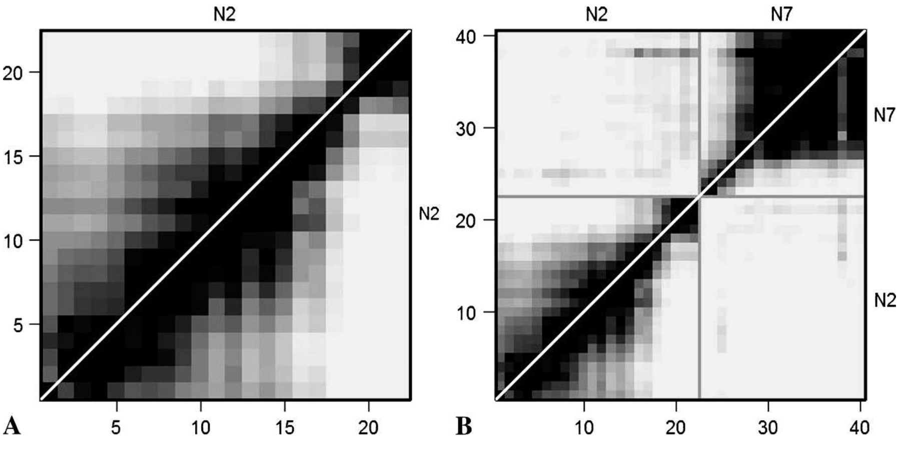

Fig. 2. (**A, B**) Recombination frequency plot, showing recombination rates between two markers above the white diagonal and LOD scores for those recombination rates below the white diagonal. Black indicates low recombination, high log odds (LOD). Darker gray areas indicate low recombination/high LOD. View **A** shows only chromosome N2, suggesting the 22 markers are in proper order. View **B** shows chromosomes N2 and N7. Note in particular that N7 has many markers that are highly linked, which is reflected in the dense linkage map for this chromosome. There is one suspicious marker on N7, showing linkage with one end of N2.

Much horticultural data involves quantitative measurements of some kind. Consider the flowering time of plants with Stellar (annual) alleles at marker E33M59.59 after

4- or 8-week vernalization in the following stem-and-leaf diagrams.

```
 4-week
 0 | 999
 1 | 111
 1 | 667799999999
 2 | 0000111112233344444
 2 | 555566677899
 3 | 000233
 3 | 588
 4 | 14
 4 |
 5 |
 5 |
 6 |
 6 | 6
\\
10 | 0
```

```
8-week
0 | 9
1 | 0112233344
1 | 5566688888888999999
2 | 00001111122222233333344
2 | 555999
3 | 3
3 | 7
4 |
4 | 5
```

For instance, there are two 4-week plants and three 8-week plants that flowered at 16 d (bold underline in the diagram). The one 4-week plant recorded as 100 d actually never flowered. Stem and leaf plots are quite useful for modest data sets, but they are cumbersome when there are more than 50 or so observations.

Histograms are nice alternatives, provided they are used properly. Histograms are special types of bar plots in which the bars are ordered to show the count or proportion of observations in consecutive interval of values. Fig. 3 shows four histograms, with the Stellar allele plants shown below the Major allele plants. Histograms give some idea about distribution shape, but they are sensitive to the number of bars chosen. The Q-Q, or quantile, plots in Fig. 4 allow us to examine how close our data are to normality, with upswings on the right indicating a slight skew toward larger values. Most statistical methods, including analysis of variance (ANOVA), implicitly assume normality and equal variance. Miller (1997) points out that normality is not very important for ANOVA tests, although having a symmetrical distribution without "heavy tails" leads to estimates of mean and variance that are more reliable. Although there are formal goodness-of-fit tests for lack of normality, they do not have much power and are often no better than an experienced eye. For instance, in Figs. 3 and 4, we notice a slight skew, which is mostly corrected by a log transform (used later). Plots such as Fig. 4 that show distribution shape are typically for internal assessment and do not make it to publication.

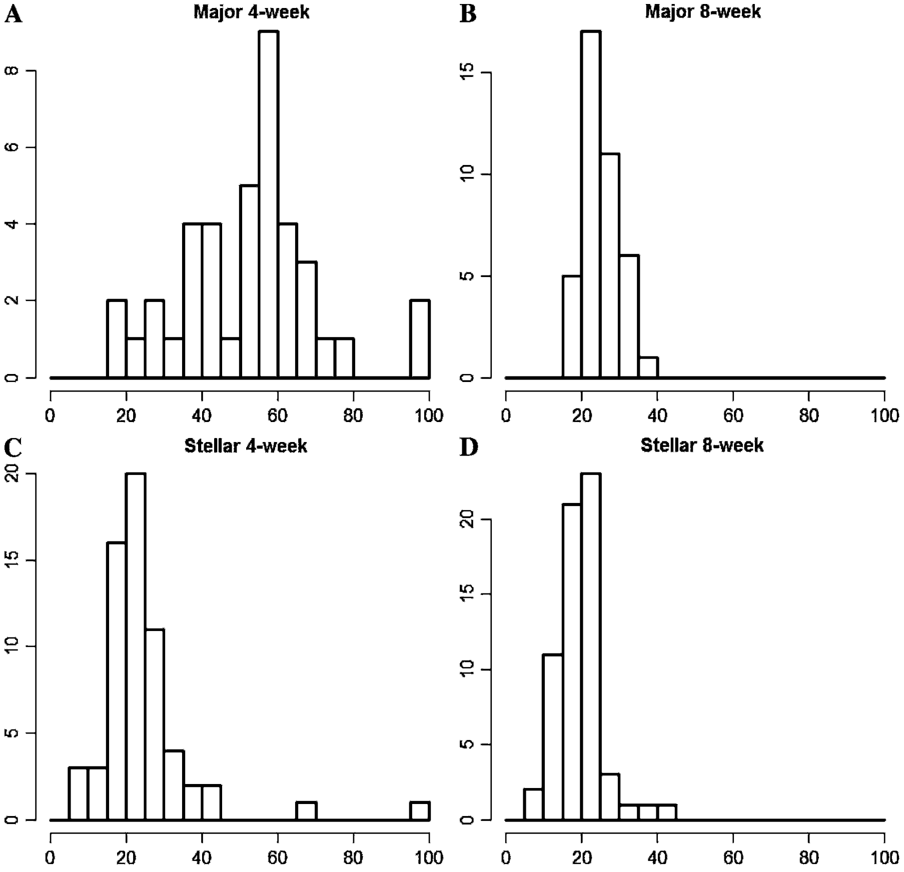

Fig. 3. (**A–D**) Histogram summaries for flowering time after 4-week (**A, C**) and 8-week (**B, D**) vernalization. Plants are partitioned by marker E33M59.59 into Major (**A, B**) or Stellar (**C, D**) type. Horizontal axes are identical; vertical axes show number of plants per 5-d interval. Plants recorded as 100 d never flowered.

I prefer showing all the data in jittered plots, adding a small amount of noise to offset values slightly, to show all the data side by side, as in Fig. 5. Such plots can be augmented with means and SEs, which allow some graphical inference about the strength of evidence for group differences. Miller (1997) notes that equal variance, or homoscedasticity, is much more important than normality in ANOVA tests, although estimated variances can differ by a factor of four without causing much harm. The recommended formal test for heteroscedasticity (unequal variance) is Levene's test, which does not require normality, but it is not very powerful; again, a practiced eye can be just as effective. A visual inspection of Fig. 5 suggests the variances are "close enough."

Pages of histograms or jittered plots would be as daunting as tables. However, for the same reasons given earlier for bar plots and pie charts, it is unwise to simply jump to 3-D histograms or pie charts. Three-dimensional ribbon plots, displaying lines as ribbons with fancy shading, merely distract the viewer from the content. These types of graphics are subject to strong bias of perspective, color choice, and so on.

With hundreds or thousands of observations, box plots or density plots may be more useful. However, these graphical summaries tend to hide subtle patterns in data. They can be quite useful with many groups, but less revealing with modest sample sizes. A dozen box plots can be displayed efficiently side by side. Box plots basically show the "middle" of the data, including the median and delineated by the upper and lower quartiles, and highlight outliers. A density plot is a smooth line that replaces the bars of a histogram. Its chief drawback is smearing over spikes in histograms (e.g., flowering times of 100 d could be spread from 90 to 110). In the interest of space, these are not included.

Measurements are not taken in isolation. In fact, the flowering times for no vernalization, and 4- and 8-week vernalization experimental conditions were measured on all 104 individuals in this genetic cross. Thus it makes sense to view the relationship among measurements in the context of a designed experiment. Scatter plots with annotation are effective tools for this, as shown in Fig. 6. Alternatively, dividing one plot into two or more paired plots, as in Fig. 7, can allow other types of annotation, such as regression lines.

FORMAL INFERENCE WITH GRAPHS

Plots can and should provide formal inference cues. Fig. 2 includes LOD test statistics as half the plot. Fig. 5 shows means and SEs for thumbnail assessment of genotypic differences, analogous to the formal *t* test or ANOVA-based F test. Fig. 7 shows regression lines, with approximate 95% confidence regions (2 SEs of the difference between the regression lines), providing a graphical test of one slope against the other. Slopes are significantly different from each other, as the confidence regions do not completely cover both lines. Values of SD are included to show near-equal variance.

The interaction plot in Fig. 8 is specifically designed for inference, showing means ± SEs for the four combinations possible with two markers. A test for interaction is enhanced by such a plot, which shows a strong difference at marker E33M59.59 only when the ec2d1a region has the Major genotype.

Diagnostic plots are useful in regression to identify observations objectively with large residuals or a large influence on analysis. Usually these do not make it into a published paper, although their use should be documented in the Materials and Methods section.

It is not wise, necessary, or possible to include all measures of inference on a graph. This can lead to a cluttered appearance. Including details in the text and/or figure legend can achieve the same purpose. Remember that the aim is to convince your skeptical audience that you have found a "real" relationship and that it has biological relevance. Tell the story using your graphs.

MASSIVE DATA: GENE MAPPING AND MICROARRAYS

Many results from experiments with massive data sets can be condensed into one or a few images, as shown in Fig. 2. The package R/qtl (www.rqtl.org, Broman et al., 2003) has many tools of this nature for gene mapping. Recently there has been considerable interest in microarrays, leading to an explosion of creative, and sometimes bizarre, graphs. Color is used extensively, although caution is in order. The most popular clustering graphs contrast green and red, which cannot be seen by readers with the most common form of color blindness.

Many of the best graphics examples are incorporated into R packages that can be found at Bioconductor (www.bioconductor.org). One example is the use of false discovery rate q values to assess thousands

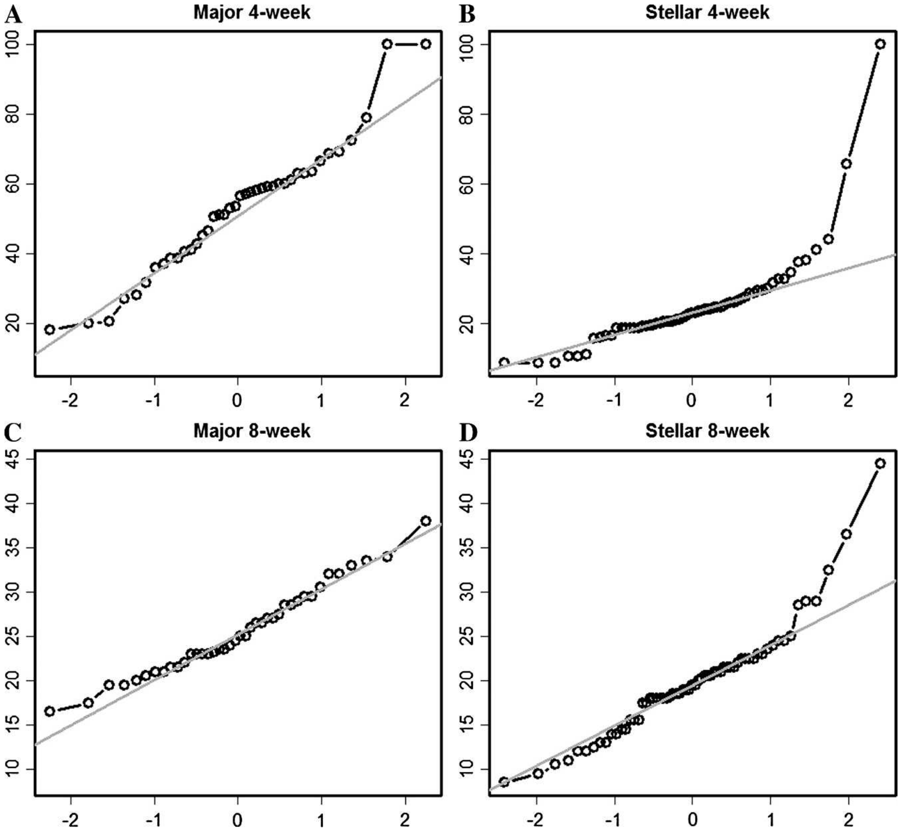

Fig. 4. (**A–D**) Q-Q plots for flowering time after 4-week (**A, B**) and 8-week (**C, D**) vernalization. Plants are partitioned by marker E33M59.59 into Major (**A, C**) or Stellar (**B, D**) type. Flowering times (vertical axes) are ordered from smallest to largest and are plotted against corresponding expected values, or quantiles, from a standard normal. Solid gray lines run through lower and upper quartiles; deviation from this line indicates lack of normality.

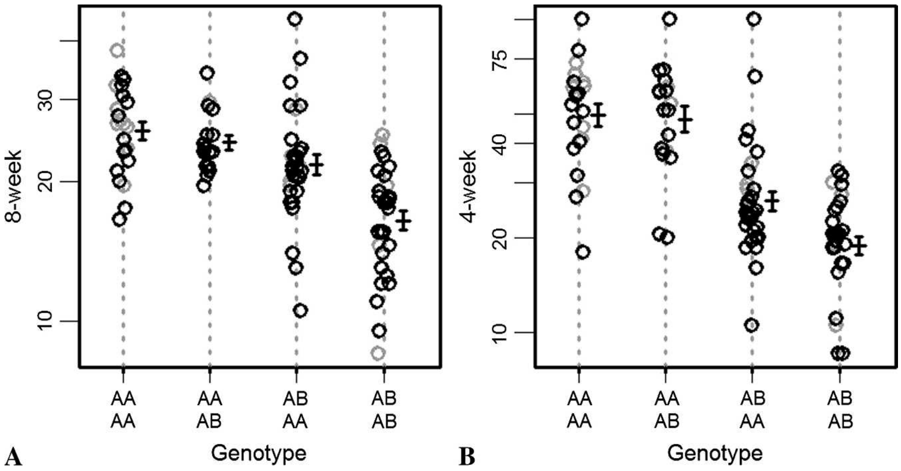

Fig. 5. (**A, B**) Jittered phenotype by genotype plots of flowering time in days from 8-week (**A**) and 4-week (**B**) vernalization treatments. Missing genotypes are imputed (gray circles) based on other map information. The three 4-week values of 100 d did not flower. Stellar (AA) and Major (AB) are identified by markers E33M59.59 (top row of horizontal axis) and ec2d1a (bottom row). Thus the second column of each plot (AA/AB) is for plants that are Stellar at E33M59.59 and Major at marker ec2d1a. Values are jittered horizontally to highlight the vertical spread in data better. Vertical scale is logarithmic to addresses increasing variance with mean. Bars to right of jittered points indicate means ±1 SE. There appears to be a strong phenotype difference at E33M59.59 for the 4-week vernalization, although other group differences appear significant based on SEs. This is confirmed by formal analysis of variance.

of *t* tests (or other statistical summaries) for differential expression at once (Storey and Tibshirani, 2003). In the interest of space, I do not include any other massive data figures.

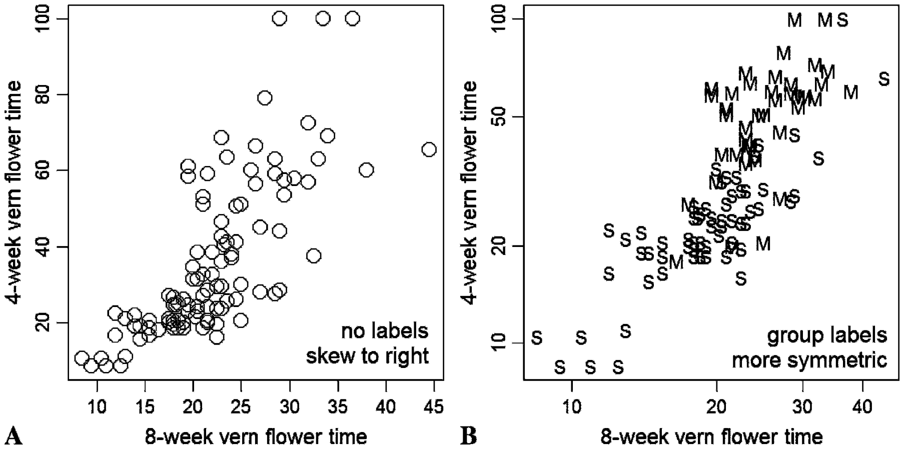

Fig. 6. (**A, B**) Scatter plots of 4-week versus 8-week vernalization. View **A** is a raw plot that has data bunched in the lower left and no annotation. View **B** identifies data by marker E33M59.59 with both color and symbol, and is plotted on a log-log basis.

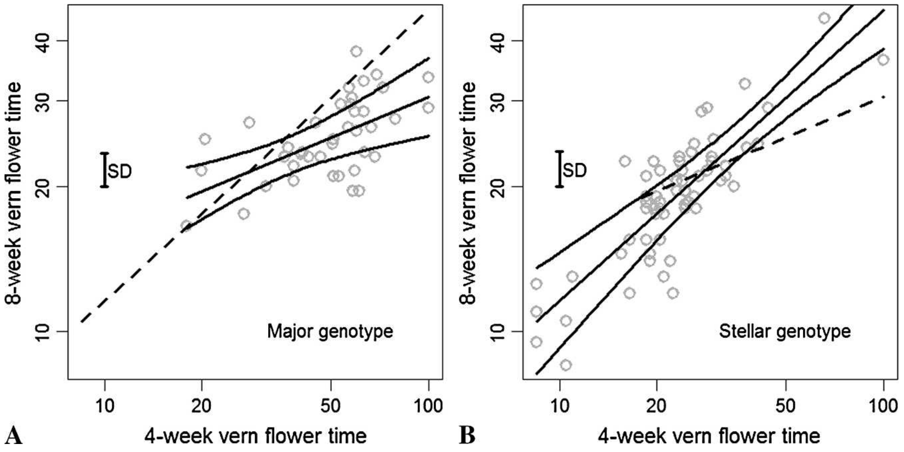

Fig. 7. (**A, B**) Regression curves on separate plots by marker E33M59.59 genotype. Plot regions are kept the same for comparison. Solid lines are for the featured genotype—Major (**A**) or Stellar (**B**)—with a dashed line confidence envelope of 2 × SE of the difference between regression lines; dotted lines are regression lines from the other genotype. Lines are significantly different using analysis of covariance. Note the similar spread from the regression line for both genotypes when plotted on log-log scale (SD = 0.068 and 0.072, respectively, in log10 units).

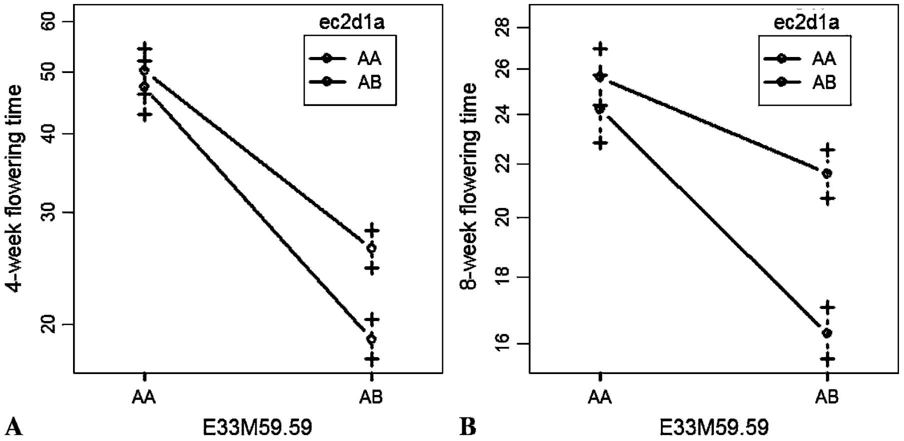

Fig. 8. (**A, B**) Interaction plot for days to flower after 4-week (**A**) or 8-week (**B**) vernalization for the two markers. The vertical axis is on the log scale again. Bars around means correspond to ±1 SE. Stellar (AA) and Major (AB) are identified by markers E33M59.59 (horizontal axis) and ec2d1a (colors and legend). Thus, there are no significant flowering time differences at ec2d1a if E33M59.59 is Stellar, whereas the 7-d (A) or 5-d (B) differences when E33M59.59 are Major are significant.

CONCLUSION

Great pictures are worth a thousand words. Constructing a great picture for publication requires some time and artistic skill, in addition to scientific acumen. It is worth previewing graphics with colleagues outside your field to ascertain that key points are made cogently. Take a few moments to examine the excellent graphics found in references cited herein. Your research has involved great efforts to get to publication; present it in a meaningful way with beautiful, informative graphics.

Literature Cited

Broman, K.W., H. Wu, S. Sen, and G.A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, NJ.

Cleveland, W.S. 1994. The elements of graphing data. Chapman and Hall/CRC Press, Boca Raton, FL.

Dalgaard, P. 2002. Introductory statistics with R. Springer-Verlag New York.

Ferreira, M.E., J. Satagopan, B.S. Yandell, P.H. Williams, and T.C. Osborn. 1995. Mapping loci controlling vernalization requirement and flowering time in *Brassica napus.*. Theor. Appl. Genet. 90:727–732.

Huff, D. 1993. How to lie with statistics. Norton, New York.

Ihaka, R. and R. Gentleman. 1996. R: A language for data analysis and graphics. J. Comp. Graph. Stat. 5:299–314.

Kangas, M. 2004. R: A computational and graphical resource for ecologists. Frontiers in Ecol. and Environ. 2:277.

Kole, C., C.E. Thorman, B.H. Karlsson, J.P. Palta, P. Gaffney, B.S. Yandell, and T.C. Osborn. 2002. Comparative mapping of loci controlling winter survival and related traits in oilseed *Brassica rapa* and *B. napus*. Mol. Breed. 9:201–210.

Miller, R.G., Jr. 1997. Beyond ANOVA 2nd ed. Chapman and Hall/CRC Press, Boca Raton, FL.

Storey, J.D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100:9440–9445.

Tufte, E.R. 1983. The visual display of quantitative information. Graphics Press, Cheshire, CT.

Tufte, E.R. 1990. Envisioning information. Graphics Press, Cheshire, CT.

Tufte, E.R. 1997. Visual explanation. Graphics Press, Cheshire, CT.

Wainer, H. 1984. How to display data badly. Am. Stat. 38:137–147.