Theor Appl Genet (2006) 113: 156–162
DOI 10.1007/s00122-006-0282-6

ORIGINAL PAPER

Philip M. Maher · Hui-Hsien Chou · Elizabeth Hahn
Tsui-Jung Wen · Patrick S. Schnable

GRAMA: genetic mapping analysis of temperature gradient capillary electrophoresis data

Received: 12 September 2005 / Accepted: 21 March 2006 / Published online: 20 April 2006
© Springer-Verlag 2006

Abstract Temperature gradient capillary electrophoresis (TGCE) is a high-throughput method to detect segregating single nucleotide polymorphisms and InDel polymorphisms in genetic mapping populations. Existing software that analyzes TGCE data was, however, designed for mutation analysis rather than genetic mapping. Genetic recombinant analysis and mapping assistant (GRAMA) is a new tool that automates TGCE data analysis for the purpose of genetic mapping. Data from multiple TGCE runs are analyzed, integrated, and displayed in an intuitive visual format. GRAMA includes an algorithm to detect peaks in electropherograms and can automatically compare its peak calls with those produced by another software package. Consequently, GRAMA provides highly accurate results with a low false positive rate of 5.9% and an even lower false negative rate of 1.3%. Because of its accuracy and intuitive interface, GRAMA boosts user productivity more than twofold relative to previous manual methods of scoring TGCE data. GRAMA is written in Java and is freely available at http://www.complex.iastate.edu.

Introduction

Dense genetic maps provide a means to link genes to their corresponding functions and facilitate a wide range of genetic manipulations of agricultural species. One limitation to generating dense genetic maps is the level of detectable polymorphisms in the mapping populations. Single nucleotide polymorphisms (SNPs) and small InDel polymorphisms (IDPs) typically occur at higher frequencies than do other types of polymorphisms. Unfortunately, most technologies for detecting SNPs and IDPs require prior knowledge of the sequences of the polymorphisms, data that are not readily available for many experimental organisms. Hsia et al. (2005) described how TGCE can be used to reliably and sensitively detect SNPs and IDPs within mapping populations even in the absence of prior knowledge regarding the specific sequences of these polymorphisms. The resulting segregation data can then be used to create high-density genetic maps.

Temperature gradient capillary electrophoresis (TGCE) detects polymorphisms by virtue of its ability to distinguish between homoduplex and heteroduplex DNA molecules. A genetic marker amplified via PCR of template DNA from an individual organism is denatured and then allowed to reassociate before being subjected to TGCE. If the assayed organism was homozygous for the genetic marker in question then only homoduplex molecules will be detected via TGCE. If the individual was heterozygous for the genetic marker then both homoduplex and heteroduplex molecules will be detected via TGCE.

The ability of TGCE to detect SNPs and small IDPs makes it equally useful for detecting genetic variation

Communicated by H. C. Becker

P. M. Maher · H. H. Chou (✉)
Department of Computer Science, Iowa State University,
503 Science II, Ames, IA 50011, USA
E-mail: hhchou@iastate.edu
Tel.: +1-515-2949242
Fax: +1-515-2948457

H. H. Chou · P. S. Schnable
L.H. Baker Center for Bioinformatics & Biological Statistics,
Iowa State University, Ames, IA 50011, USA

H. H. Chou · P. S. Schnable
Center for Plant Genomics, Iowa State University,
Ames, IA 50011, USA

H. H. Chou · P. S. Schnable
Department of Genetics, Development and Cell Biology,
Iowa State University, Ames, IA 50011, USA

E. Hahn · T. J. Wen · P. S. Schnable
Department of Agronomy, Iowa State University,
Ames, IA 50011, USA

between two different alleles in a mapping population. Analyzing data from a mapping population is, however, more complex than simply identifying which DNA samples contain heteroduplex molecules as in the case of mutation detection. To gather mapping information, amplified PCR products from each member of a biparental mapping population are mixed with each progenitor line separately before being denatured and subjected to TGCE. Useful types of mapping population can be produced by a collection of recombinant inbred (RI) lines (Bailey 1971; Burr and Burr 1991) or by the Double Haploid technique (Sulima et al. 2000; Bariana et al. 2006). This paper reports the results based on the intermated B73×Mo17 (IBM) maize RI population (Lee et al. 2002).

Let the two original parental lines of the RI mapping population be line 1 and line 2. If heteroduplex molecules are discovered in the mixture of an RI and line 1, then that RI likely carries the allele inherited from line 2 at this genetic marker. If, however, heteroduplex molecules are discovered in the mixture of the RI and line 2, then the RI likely carries the allele inherited from line 1. Here, simply marking samples that carry heteroduplex molecules has no straightforward meaning unless the progenitor line mixed with is also identified. The same issue arises for samples in which homoduplex molecules are discovered. The conclusions drawn from detecting homoduplex molecules in a well are different depending on with which of the original progenitor lines the RI line is mixed.

To obtain mapping data two duplicate microtiter plates are prepared. Each well contains the amplification product for a single genetic marker from a different RI line. The first and second plates are then mixed with the amplification products from lines 1 and 2, respectively. This set up allows a single control to be selected and the results can be interpreted the same way for every well on a given plate. The expectation is that no heteroduplex molecules will be present in one mixture when present in the other. Hence, to improve the accuracy of mapping scores, it would be desirable to be able to compare the results from both mixtures for a particular RI line. This is not easily accomplished using existing software.

Genetic recombinant analysis and mapping assistant (GRAMA) is a software tool that has been developed specifically for using TGCE data to perform analyses of genetic mapping populations and to generate mapping scores that can be used as input to genetic mapping programs (Lathrop and Lalouel 1984; Curtis and Gurling 1993; Lincoln et al. 1993; Stam 1993; Mester et al. 2003). GRAMA allows the user to view all relevant data simultaneously and automates accurate decisions to be made as to the allelic content of an RI line for a particular genetic marker (or if it is undeterminable). This article focuses on the GRAMA tool; the biological findings of the GRAMA application on a maize genetic mapping project using TGCE data has been previously reported in Hsia et al. (2005).

Materials and methods

Peak calling

A major functionality required of any tool that is to analyze TGCE data is electropherogram analysis. It is important to detect in each well after renaturation has taken place whether only homoduplex molecules are present or whether both homoduplex and heteroduplex molecules are present. As described in Hsia et al. (2005), this leads to the formation of single peaks in the electropherograms generated from wells where only homoduplex molecules are present and the formation of multiple peaks in the electropherograms generated from wells where both homoduplex and heteroduplex molecules are present. Humans can easily identify the generated peak patterns when observing a graph. However, for a computer to automatically detect peaks without user intervention, generic rules must be established that work for the majority of situations. The rules that have been adopted for the peak identification algorithm in GRAMA are based on changes in the slope of a trace graph from point to point.

Each electropherogram has a baseline from which the peaks arise. This baseline is not completely flat. In fact, it can be quite uneven due to noise and other factors. In addition, the baseline may be oriented around a significantly different value on the *y*-axis in different regions of the graph though this deviation is typically a gradual change. The deviation of the baseline and the noise in the electropherograms cause inaccuracies if the simple technique of height above the baseline is employed to identify peaks. Many noise peaks observed tend to be sudden spikes in the electropherograms as opposed to the smooth curves that are typically observed and expected for legitimate peaks. The GRAMA algorithm employs a technique that only identifies smooth curves as possible peaks. The algorithm also is able to cope with the shifting baseline problem because the fact that in general a baseline exists is sufficient for the algorithm to operate properly.

Moving from left to right along the electropherogram, if the beginning of a peak is encountered, the slope of the graph between consecutive pairs of trace points will begin to increase. The region of the graph in which the slope between consecutive points is increasing is said to be concave upward. At some point while continuing from left to right, the slope between consecutive points will cease to increase and begin to decrease. The region of the graph where this occurs is said to be concave downward. The maximum point of the peak should occur in this area barring imperfections in the shape of the peak. Continuing from left to right, the slope between consecutive points will begin to increase again resulting in another concave upward region. At this point the maximum value of the peak has already been encountered. This increase in slope must occur so the curvature of the peak can eventually reconverge with the baseline.

The points where the change in slope switches from increasing to decreasing or from decreasing to increasing are called inflection points. These features can be seen in Fig. 1.

The initial step of the GRAMA algorithm is to locate all of the inflection points. The algorithm searches the graph from left to right checking the slopes between consecutive pairs of points. When the graph switches from being concave upward to being concave downward, GRAMA marks this as the potential beginning of a peak. Then when the graph switches from being concave downward to being concave upward, GRAMA marks this as the potential end of a peak.

The idea is that for every peak the maximum point should occur between the beginning of a peak and the following end of a peak. Because of this, GRAMA stores the maximum point in each of these intervals. However, bumps or ridges on the surface of the peak will cause changes in concavity, yielding inflection points that are not true beginnings and endings of peaks. An example of this is shown in Fig. 2a. To combat this problem GRAMA "slides" the beginning and ending markers along the graph to the left and right, respectively. The slope on the left side of the peak will be positive so the left marker is allowed to slide left until it encounters a sufficiently large negative slope ($<$ −maximum height of tallest peak × 2%). By allowing the marker to continue sliding left even after encountering a small negative slope, minor bumps on the peak are ignored. After this process, the beginning marker should be in a location that is the very beginning of the peak. The ending marker is adjusted via a similar process. When the inflection points are flanking a bump or ridge on the peak, either the beginning or ending marker will not be able to slide far as it will be sliding toward the maximum point of the peak.

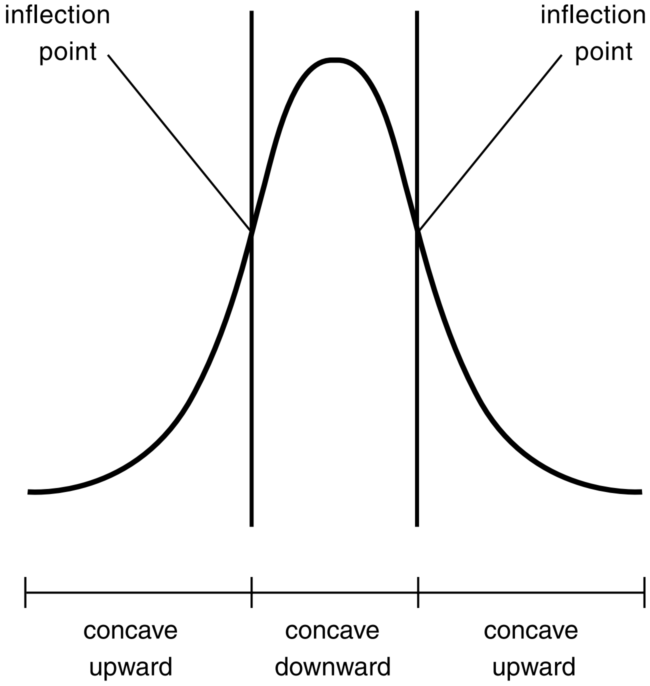

Fig. 1 A peak consists of a concave upward region followed by a concave downward region then followed by another concave upward region. The points at which the peak switches from concave upward to concave downward or concave downward to concave upward are called inflection points

Once the sliding has completed, the GRAMA peak detection algorithm checks whether the peak flanked by each pair of markers is tall enough to be considered relevant. This is controlled by a sensitivity value set by the user. Irrelevant markers are then removed. The final step is to remove beginning and ending marker pairs that are flanked by the other marker pairs. As mentioned, marker pairs flanking bumps and ridges are not able to slide along both sides of a peak, but marker pairs that actually flank the maximum point of the peak can slide along both sides of the peak. The end result is that markers flanking bumps or ridges are well contained within the interval defined by the beginning and ending markers that flank the maximum of the peak. Therefore these internal markers are removed. The electropherogram from Fig. 2a is shown again in Fig. 2b after GRAMA processing. The effectiveness of the GRAMA peak calling algorithm is quite obvious.

Data conversion

Genetic recombinant analysis and mapping assistant obtains the electropherogram data from the TGCE system made by Spectrumedix LLC (State College, PA, USA; http://www.spectrumedix.com). The vendor software Revelation™ produces a report including all of the unedited calls along with the data for the electropherograms from all of the capillaries. By utilizing this information GRAMA can accurately reproduce the electropherograms for each run. For a particular genetic marker there will typically be two or three runs in order to capture all necessary data: one run where the RI DNA is mixed with parent 1 DNA and another run where the RI DNA is mixed with parent 2 DNA. An optional third run containing only unmixed RI DNA can be used to detect heterozygous markers, but it is optional when the mapping population is expected to be homozygous at a marker location.

Genetic recombinant analysis and mapping assistant requires two Revelation report files, also called score files, for both of the DNA mixtures. GRAMA automatically transforms the Revelation mutation scores to "1" or "2" to represent the original progenitor a particular RI resembles at a particular genetic marker. If the Revelation score file for a third unmixed plate is also loaded, its scores are also transformed. In this case, it is simply noted whether heteroduplex molecules are formed in each well, indicating that the RI line DNA may not be homogeneous and may invalid the results from the two mixture plates. In addition to reading and translating Revelation score files, the electropherograms for each capillary and plate combination are analyzed at the same

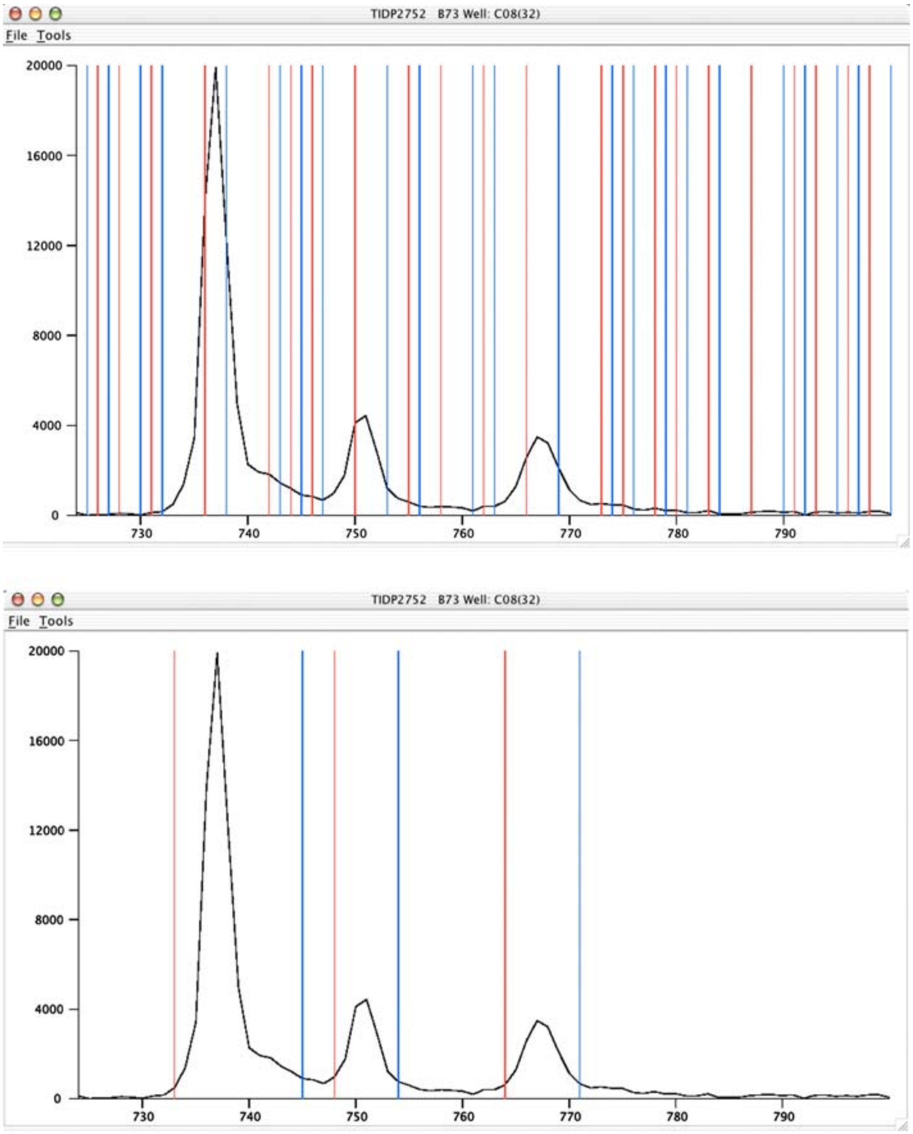

Fig. 2 **a** The initial step of the GRAMA peak determination algorithm is to locate all of the inflection points. The *red lines* indicate possible peak beginnings and the *blue lines* indicate possible peak endings. As can be seen from this figure, this step alone is not enough to correctly determine where peaks are actually located. **b** After the GRAMA peak sliding algorithm, only the beginning and ending *peak indicator lines* flanking areas where the algorithm has determined that a relevant peak exists are shown

time. As mentioned GRAMA uses its own peak determination algorithm to analyze each electropherogram and count the number of peaks. The criteria for the scoring are the same as when transforming the Revelation calls.

Genetic analysis

After data processing, GRAMA displays a main window, which summarizes all of the data collected and calculated about each of the RIs for a particular genetic marker. There is one column group for each of the two plates containing DNA mixtures and another column group for the unmixed data if present. Each of these column groups contains four subcolumns. The first subcolumn is the Revelation score column. This reports the Revelation score as translated from the score file provided. The second subcolumn is the GRAMA score column. The GRAMA score column reports the score that GRAMA calculated by using its own peak determination algorithm. A third subcolumn reports the score automatically determined for an RI on a particular plate. The user can edit this column to correct a score if necessary. The fourth and final subcolumn is the Well column. Clicking on any cell in this column pops up the electropherogram for the RI on the corresponding plate.

Immediately following all column groups are two more columns that contain consensus calls. Consensus calls are calls automatically generated from the scores of each plate. The Consensus 1 column contains detail consensus scores that provide insight into how the joint plate scores determined them. The Consensus 2 column contains simplified scores that are intended to be used as the input for a genetic mapping program. They simply indicate from which inbred line the genetic content for the RI was most likely originated.

More information can be gained from this main window then is initially displayed. As mentioned, when a user clicks on any cell in the Well columns, an electropherogram will appear. An example is shown in Fig. 2b. Red and blue lines indicate the detected beginning and ending of a peak. These indicators allow the user to see clearly how GRAMA arrived at its scoring decision. By clicking on any cell in the Capillary label column, the user can obtain a "horizontal" view allowing the simultaneously comparison of all electropherograms of a particular capillary for all of the plates. Similarly, by clicking on the Well column heading for any of the plates, the user can see a "vertical" view showing all electropherograms for that plate. These views allow the user to identify potential abnormal scores and correct them accordingly. The details of these functionalities are described in the user manual, which comes with GRAMA. Once the user is satisfied with all plate scores, the consensus scores are finalized, and all data for this particular genetic marker can be output to a spreadsheet program or a database. Results from multiple genetic markers can then be collected and formatted for input to a genetic mapping program.

Experiment design

Several experiments were performed to specifically investigate the accuracy of the GRAMA and Revelation peak determination algorithms. These experiments were performed using several hundred IDPs that exist between two inbred lines of maize (B73 and Mo17) and that can be detected via TGCE. These polymorphisms (genetic markers) were discovered using the procedure described in Hsia et al. (2005). Each of these genetic markers was used to amplify each of the RI lines from the intermated B73×Mo17 (IBM) population (Lee et al. 2002). The inbred lines B73 and Mo17 were designated as lines 1 and 2, respectively. The TGCE data were analyzed using both the Revelation and GRAMA software packages. Each mapping score was also manually evaluated by experienced users of the TGCE system.

Statistics were gathered from 529 different genetic markers. Since each microtiter plate contains 96 wells, 529 markers provide 50,784 mapping scores. For 37 of the 529 genetic markers, an unmixed plate (RI amplification products only) was also analyzed. So for 37 genetic markers, three plates were run and for the remaining 492 markers only two plates were run. Hence, a total of 105,120 electropherograms were evaluated by both programs.

Results and discussion

Single call correctness

The first property analyzed was the abilities of Revelation and GRAMA to distinguish between wells containing homoduplex and heteroduplex molecules. The scores produced by both software packages were compared to those assigned by experienced TGCE data curators that were assumed to be correct. In this analysis, if the data curator was unable to determine a well containing homoduplex or heteroduplex molecules, then the results are not included because a computer program is not expected to score correctly under the circumstance.

Revelation incorrectly scored a well containing homoduplex molecules as containing heteroduplex molecules 2,162 times. A total of 51,739 wells were scored as containing homoduplex molecules by data curators for those wells that Revelation scored. Thus, Revelation had a 4.2% false positive error rate. GRAMA incorrectly scored a well containing homoduplex molecules as containing heteroduplex molecules 3,180 times. For those wells where GRAMA attempted to make a call, data curators scored the well as containing homoduplex molecules 53,989 times. Thus, GRAMA has a 5.9% false positive error rate. The difference in the total number of wells containing homoduplex molecules as scored by the two programs is due to the fact that Revelation can categorize a well as undeterminable. Thus, there were 2,250 times where GRAMA attempted to make a call but Revelation did not. The accuracy of GRAMA's peak determination algorithm on wells uncalled by Revelation will be discussed later.

Revelation has a 1.3% false negative rate on the sample set. False negatives occur when the peak determination algorithm fails to identify heteroduplex molecules and scores the well as containing only homoduplex molecules. There were 621 wells out of 48,698 that were scored as containing heteroduplex molecules by data curators but Revelation scored incorrectly. GRAMA's false negative rate was very similar at 1.3%. For a total of 48,877 wells that GRAMA attempted to score, GRAMA incorrectly classified 630 wells as containing only homoduplex molecules while data curators classified them as containing heteroduplex molecules. The difference in the total number of wells that were scored between the programs again results from the fact that Revelation scored some wells as undeterminable while GRAMA attempted to score all wells. Out of the 2,458 instances where Revelation did not make a call and GRAMA did, GRAMA made the correct call 2,376 times. Therefore, GRAMA has a 96.7% accuracy rate on wells that Revelation opted not to score.

Combined call correctness

Wells uncalled by either Revelation or GRAMA are not included in the combined call analysis since they always trigger involvement by a data curator. For the 100,408 wells in the sample set that were automatically processed, 94,561 were scored correctly by both GRAMA and Revelation, or about 94.2%; 3,077 were scored correctly by Revelation but not by GRAMA; and 2,119 were scored correctly by GRAMA but not by Revelation. Thus, the two algorithms disagreed on about 5.2% of the wells. Among those, Revelation made correct calls 59.2% of the time, while GRAMA was correct 40.8% of the time. The number of wells where both algorithms made the same mistake is 651. This is only 0.6% of the total number of wells included in this study.

From the above results, it can be seen that both algorithms are very accurate in distinguishing between wells containing homoduplex molecules and wells containing heteroduplex molecules and, in turn, alerting the user of possible mistakes made by one of the algorithms. By comparing the results from multiple plates together, the ability to detect mistakes and flag them for the data curator are further improved. For a pair of mixture plates, the expectation is to find homoduplex molecules in one plate and heteroduplex molecules in the other for a particular RI. Because a particular RI typically carries only one allele of a given marker (i.e., it is homozygous), the scores for both mixtures should match; if that is not the case the data curator should be alerted. Since one of the mixture plates should only contain homoduplex molecules and both algorithms have very low false positive error rates on this, it is highly unlikely that mistakes will not be discovered and flagged for manual inspection by GRAMA.

Combined call with plate coupling correctness

The sample contains 50,784 mapping scores. Of these Revelation made a mistake for both mixtures 23 times. GRAMA made a mistake for both mixtures 28 times. What is interesting, however, is that these mistakes did not occur for the same RI and genetic marker combination. For each of the 23 cases where Revelation made a mistake for both mixtures, GRAMA was able to correctly score at least one of the mixtures, so this inconsistency was brought to the attention of a data curator. In the same way by correctly scoring at least one of the mixtures Revelation was able to bring to the attention of a data curator 28 mistakes by GRAMA. There were no cases in which both Revelation and GRAMA scored both mixtures incorrectly for the same RI and genetic marker based on all the data analyzed.

More efficient user operation

By flagging all potential mistakes, it may seem as if this approach would create more work for the user. Nevertheless, it was found that for 76% of the mapping scores both algorithms agreed for all mixtures (and the unmixed plate if included), and the final scores agreed for both plates. Thus, more than three-fourth of the time a data curator can simply accept the results of GRAMA's combined analysis and quickly move on to mapping scores that have been flagged. We have observed a greater than twofold increase in productivity by using GRAMA to conduct genetic recombinant analysis as opposed to using a spreadsheet-based manual method.

Conclusion

To generate a high-density genetic map it is essential that the needed mapping statistics be gathered as efficiently and accurately as possible. TGCE is beneficial as polymorphisms between two different alleles can be detected with no prior knowledge of the sequence (Hsia et al. 2005). The GRAMA software package described in this paper was developed specifically for the high-throughput analysis of TGCE mapping data. GRAMA has its own peak determination algorithm and also combines scoring results produced by Revelation, the TGCE vendor supplied software. Because each program uses different methods to make scoring decisions, when they agree, it is likely that both are correct. In addition, results from both parental mixtures are included, and each algorithm makes one call for each mixture. Thus, there is a four-way check on whether a given mapping score is correct. If any of the four calls disagree with the rest, a data curator is alerted to perform an inspection of the data. This observation concurs with the experimental study, which demonstrates that GRAMA indeed has a very low-error rate. Because of the many checks and balances built into GRAMA's automatic analysis process, nearly all potential mistakes are brought to the attention of a data curator. Through usage analysis we also found that GRAMA significantly improved user productivity over previous spreadsheet methods by more than twofold. We thus conclude that the combination of the TGCE method and the GRAMA software package is an important advance toward the goal of producing high-quality dense genetic maps efficiently for genome sequencing and functional genomics.

Acknowledgments We thank Dr. An-Ping Hsia for providing assistance in gathering materials and resources and editing the manuscript, Hsin "Debbie" Chen for early descriptions of the functionalities needed in a software program that was to become GRAMA, and Yi-Yin "Rita" Chen for entering much of the data from which our statistics were derived. We would also like to thank Sunyoung Park for her assistance in the creation of some of the figures and the other colleagues in the ISU Complex Computation Lab for discussion. This research was funded by National Science Foundation award DBI 0321711 to P.S.S. and others.

References

Bailey DW (1971) Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. Transplantation 11:325–327

Bariana H, Parry N, Barclay I, Loughman R, McLean R, Shankar M, Wilson R, Willey N, Francki M (2006) Identification and characterization of stripe rust resistance gene *Yr34* in common wheat. Theor Appl Genet 112(6):1143–1148

Burr B, Burr FA (1991) Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations. Trends Genet 7:55–60

Curtis D, Gurling H (1993) A procedure for combining two-point lod scores into a summary multipoint map. Hum Hered 43:173–185

Hsia A-P, Wen T-J, Chen H, Liu Z, Yandeau M, Wei Y, Guo L, Schnable P (2005) Temperature gradient capillary electrophoresis (TGCE): a tool for the high throughput discovery and mapping of SNPs and IDPs. Theor Appl Genet 111(2):218–225

Lathrop GM, Lalouel JM (1984) Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 36:460–465

Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol Biol 48:453–461

Lincoln S, Daly MJ, Lander ES (1993) Constructing genetic linkage maps with MAPMAKER/EXP version 3.0: a tutorial and reference manual. http://www.broad.mit.edu/genome_software/other/mapmaker.html

Mester DI, Ronin YI, Hu Y, Peng J, Nevo E, Korol AB (2003) Efficient multipoint mapping: making use of dominant repulsion-phase markers. Theor Appl Genet 107:1102–1112

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Sulima I, Kalendar RN, Sivolap IM (2000) The mapping of the barley genome by RAPD analysis using double haploid strains. Tsitol Genet 34(4):41–49