SIMULATION STUDY OF LINKAGE MAP CONS 1 RUCTION WITH MISSING AND MIS-SCORED RAPD DATA

T.L. Kubisiak[1/], C.D. Nelson[2/], and M. Stine[1/]

Abstract. --Due to the recent interest in the random amplified polymorphic DNA (RAPD) technique for use in genetic linkage mapping, a series of computer simulations was conducted to investigate the effects of missing and mis-scored data on recombination estimates and linkage group construction. Ten maps (100 randomly distributed markers scored on 80 individuals) constructed with the software GREGOR_ were modified to produce varying levels of missing (5%, 10%, 15%, and 20%) and mis-scored (1%, 2%, 4%, and 8%) entries. The distribution and levels of missing data were modeled after that found in actual RAPD data sets and mis-scored data was targeted to occur within specific markers. The resulting data sets were entered into MAPMAKER II and two-point recombination estimates and linkage group information were obtained. Analysis of variance was used to determine if there were significant differences among treatment means for the standard deviation of two point recombination estimates, number of framework markers, number of linkage groups, number of markers per linkage group, and number of marker order changes. Given no mis-scored data, si[g]nificant differences among treatment means were not detected until a level of 20% missing data was reached. At this level, on average, 17.3% fewer markers could be placed into framework groupings. Given no missing data, significant differences among treatment means were not detected until a level of 4% mis-scoring was reached. At this level, on average, 28.6% fewer markers could be placed into framework groupings. The interaction between missing and mis-scored data was also investigated and not found to be significant. Based on these simulations, it is suggested that levels as high as 15% missing data and 2% mis-scored data can be tolerated during primary genetic map preparation.

Keywords: Genetic mapping, linkage, polymerase chain reaction, PCR, random amplified polymorphic DNA, RAPD

INTRODUCTION

The relative ease and speed with which large numbers of RAPD markers can be generated makes them extremely appealing for use in constructing primary genetic linkage maps. RAPD markers are generated by the use of single, randomly sequenced oligonucleotide primers and the polymerase chain reaction (Williams et al., 1990). A segment of DNA is amplified whenever two nucleotide sequences with high degrees of similarity to that of the primer occur

[1]/ Louisiana State University Agricultural Center, Louisiana Agricultural Experiment Station, School of Forestry, Wildlife, and Fisheries, Baton Rouge, LA 70803.

[2]/ U.S.D.A. Forest Service, Southern Forest Experiment Station, Institute of Forest Genetics, Gulfport, MS 39505.

within 2-3 Kb of one another on opposite strands of the template DNA. Repeated cycles of denaturation and extension result in the exponential amplification of the segment. Despite its conceptual simplicity the kinetics of the RAPD reaction are quite complex. Annealing temperature, degree of sequence similarity at priming sites, and primer competition all can affect the amplification of RAPD markers. In addition, when large numbers of RAPD reactions are being run on a daily basis, a small percentage of reactions fail. As a result, amplification inconsistencies could produce spurious data in the form of mis-scored individuals, and unless re-amplified, failed reactions would have to be recorded as missing data. The goal of this research was to investigate what effects various levels of missing and mis-scored RAPD data have on recombination estimates and linkage group construction.

MA LERIALS AND METHODS

Ten known marker maps and corresponding data sets were constructed using the software GREGOR version 1.3 (Nick Tinker, McGill University). A configuration to model 10 pairs of chromosomes, 160 possible loci per chromosome and 1% recombination between adjacent loci, was chosen. A marker list consisting of 100 randomly distributed loci was defined for each map. The parents used for generating the mapping population were defined as follows: parent 1 was heterozygous (complete coupling) for all 100 marker loci, parent 2 was defined as being homozygous recessive for all 100 marker loci (tester). This coding arrangement would be similar to that used when constructing maps from haploid megagametophyte data. The mapping population consisted of 80 individuals.

In order to investigate the effects of missing data, five MAPMAKER II-compatible data sets were produced from each GREGOR data set. One represented the true data set and four represented various levels of missing data (5%, 10%, 15%, and 20%), for a total of 50 data sets. In order to determine how missing data should be targeted, we looked at the distribution of missing entries in actual RAPD data sets. Based on data generated for 2 different slash pines (Nelson et al., 1992; vanBuijtenen et al., 1992) and a longleaf pine (Kubisiak et al., 1992), missing data appear to be exponentially distributed. Most markers have no, or a few missing entries, with considerably fewer markers being found as levels of missing data increase. By randomly sampling from the function describing this distribution, the study targeted each marker to receive a specified number of missing entries, so that when averaged over all markers, the data set-wide levels were equal to 5%, 10%, 15%, or 20% missing.

Five MAPMAKER II-compatible data sets containing various levels of mis-scoring (0%, 1%, 2%, 4% and 8%) were also produced from each GREGOR data set, for a total of 50 data sets. When RAPD loci were scored, markers were categorized based on a confidence score (Kubisiak, et al., 1992). A putative polymorphic locus was given a lower confidence rating if the locus of interest was only faintly amplified or bands of similar molecular weight as the locus of interest were present. If mis-scoring were to result from one of these two sources, most errors should be occurring within specific loci, versus being random over the entire data set. Therefore, in order to produce the overall levels of 1%, 2%, 4%, and 8% mis-scoring, 20% of the markers were randomly chosen to receive 5%, 10%, 20%, or 40% mis-scoring.

The data sets were entered into the computer package MAPMAKER II (version 1.9), and recombination estimates and linkage group information were obtained. The mapping strategy was similar to that suggested by Lander et al. (1987). To determine all two-point groupings,

a log of the odds ratio (LOD) of 5.0 and a recombination frequency of 0.25 were chosen. To determine marker order within a particular linkage group, a LOD score of 3.0 was chosen. These markers, and their respective orders, were designated as framework groupings.

In order to evaluate the effects of missing or mis-scored data, an analysis of variance was used to determine if there were significant differences among treatment means for various descriptive measures. These included the standard deviation of the departure of two-point recombination estimates from their "true" or known values [std(/r-θ/)], number of framework markers mapped, number of linkage groups obtained, number of markers per linkage group, and number of marker order changes. The std(/r-θ/) had the following form:

$$\sqrt{\frac{\sum_{i}^{p}\left[(/r_i-\theta_i/)-\frac{\sum_{i}^{p}(/r_i-\theta_i/)}{p}\right]^2}{p}}$$

r = pairwise recombination estimate
θ = "true" or known pairwise distance
p = number of pairwise comparisons
(For 100 loci p = 4950)

RESULTS

Given no mis-scored data, std(/r-θ/) was found to increase with the level of missing data (Table 1). A significant difference among treatment means occurred at 15%. With no missing data, std(/r-θ/) did not appear to increase until a level of 4% mis-scoring was attained (Table 1). However, it was not until 8% that a significant difference among

Table 1. Effect of missing and mis-scored data on the standard deviation of two-point recombination estimates [std(/r-θ/)].

Missing Data Treatment (% missing)	Mean std(/r-θ/)	Mis-scored Data Treatment (% mis-scored)	Mean std(/r-θ/)
0	0.056045 A*	2	0.055173 A
5	0.056627 A	1	0.055358 A
10	0.057566 A	0	0.056045 A
15	0.059567 B	4	0.056638 A
20	0.061033 B	8	0.060773 B

*Those means with the same letter are not significantly different at alpha=0.05 using Tukey's Studentized Range Test. Means based on 10 replicate data sets.

treatment means was detected. The number of markers placed into framework groupings was found to decrease as the level of missing or mis-scored data increased (Table 2). A significant difference among treatment means was not detected until a level of 20% missing

data was attained. For the mis-scored data sets, a significant difference among treatment means was not detected until a level of 4% was attained (Table 2). Missing data did not appear to affect the number of linkage groups obtained in any sort of a

Table 2. Effect of missing and mis-scored data on the number of framework markers mapped (F.M.).

Missing Data		Ms-scored Data	
Treatment (% missing)	Mean # F.M.	Treatment % mis-scored)	Mean # F.M.
0	69.8 A*	0	69.8 A
5	68.6 A	1	66.5 A,B
10	65.0 A	2	60.7 B
15	64.6 A	4	49.8 C
20	57.7 B	8	48.3 C

*Those means with the same letter are not significantly different at alpha=0.05 using Tukey's Studentized Range Test. Means based on 10 replicate data sets.

predictable manner (Table 3). No significant differences among treatment means was detected. However, as the levels of mis-scoring increased the number of linkage groups obtained was found to decrease (Table 3). A significant difference among treatment means occurred at 4%. The average number of markers per linkage group was, generally,

Table 3. Effect of missing and mis-scored data on the number of linkage groups (L.G.).

Missing Data		Ms-scored Data	
Treatment (% missing)	Mean # L.G.	Treatment (% mis-scored)	Mean # L.G.
15	14.8 A*	0	14.6 A
0	14.6 A	1	14.6 A
10	14.3 A	2	13.9 A
5	14.0 A	4	11.9 B
20	13.8 A	8	11.5 B

*Those means with the same letter are not significantly different at alpha.05 using Tukey's Studentized Range Test. Means based on 10 replicate data sets.

found to decrease as the levels of missing or mis-scored data increased (Table 4). However, the differences among treatment means was more problematic as groupings overlapped. Finally, the number of marker order changes was not found to be affected by missing data in any sort of a predictable manner, and no significant differences among treatment means was detected (Table 5). The number of marker order changes was found to increase up to the level of 2% mis-scoring, after which numbers decreased again (Table 5). Treatment means were not determined to be statistically different.

Table 4. Effect of missing and mis-scored data on the average number of framework markers per linkage group (M./L.G.).

Missing Data Treatment (% missing)	Mean # M./L.G.	Mis-scored Data Treatment (% mis-scored)	Mean # M./L.G.
5	4.948 A*	0	4.811 A
0	4.811 A,B	1	4.568 A,B
10	4.591 A,B	2	4.388 A,B
15	4.410 A,B	8	4.216 B
20	4.216 B	4	4.194 B

*Those means with the same letter are not significantly different at alpha .05 using Tukey's Studentized Range Test. Means based on 10 replicate data sets.

The interaction between missing and mis-scored data was also investigated and not found to be statistically significant for any of the variables investigated (Data not shown).

Table 5. Effect of missing and mis-scored data on the number of marker order changes (O.C.).

Missing Data Treatment (% missing)	Mean # O.C.	Mis-scored Data Treatment (% mis-scored)	Mean # O.C.
20	0.50 A*	2	1.20 A
10	0.30 A	4	0.80 A,B
0	0.20 A	1	0.50 A,B
15	0.10 A	0	0.20 A,B
5	0.00 A	8	0.10 B

*Those means with the same letter are not significantly different at alpha=0.05 using Tukey's Studentized Range Test. Means based on 10 replicate data sets.

DISCUSSION

Prior to the analysis, we hypothesized that as levels of missing or mis-scored data increased within a data set the standard deviation of the departure of pairwise recombination estimates from their "true" or known values, std(/r-θ/) would increase; likewise, the accuracy with which the genetic distance between two markers can be estimated decreases. The results seem to support our hypothesis. The std(/r-θ/) was found to increase as the levels of missing or mis-scored data increased (significance at 15% and 8%, respectively).

We hypothesized that levels of mis-scored data would have a more pronounced effect on std(/r-θ/) than would comparable levels of missing data. Missing data only indirectly affects linkage estimation by reducing the effective mapping population within particular markers.

However, mis-scored data would tend to confound the linkage relationship between markers and hence directly effect recombination estimates. Interestingly, these simulations indicate that, in terms of std(/r-θ/), the overall effects of a 5% level of missing data are comparable to a 4% level of mis-scoring.

The differences among mean values for the various levels of missing data and mis-scored data appear to be quite small in terms of genetic distance, however, this is due to the fact that a majority of the markers which are unlinked are still estimated to be unlinked even when harboring missing or mis-scored data. Therefore, a large number of the comparisons are not contributing to the sum of squared deviations in the calculation. The levels at which significant differences among means were detected does appear to be indicative of a problem threshold. In other words, the levels of missing or mis-scored data at which significant differences among treatment means were detected for std(/r-θ/) are similar to the levels found to cause significant group discrimination in other measures such as the number of framework markers, number of linkage groups, and number of markers per linkage group.

It makes sense that the number of markers placed into framework groupings would decrease as levels of missing and mis-scored data increase. These results indicate just such an inverse relationship (Table 2). Given no mis-scored data, at levels of 20% missing data, 17.3% fewer markers could be placed into framework groupings. Given no missing data, at levels of 4% mis-scoring, 28.6% fewer marker could be placed into framework groupings. In terms of the number of framework markers placed, lower levels of mis-scoring appear to have a more profound effect than do comparable levels of missing data.

Developing *a priori* hypotheses regarding how missing and mis-scored data might affect the number of linkage groups and average number of markers per linkage group was a more problematic situation. It could be hypothesized that missing or mis-scored data might cause whole linkage groups to fall apart, resulting in fewer mapped groups. Alternatively, mis-scored or missing data might cause larger groups to be broken into two or more smaller groups, resulting in a larger number of mapped groups, each having fewer markers. In terms of missing data, no apparent trends were found in the number of linkage groups obtained. However, as levels of mis-scoring increased, the number of linkage groups obtained decreased. This would seem to indicate that mis-scoring is primarily causing entire linkage groups to be lost. Although the number of framework markers per linkage group appears to decrease with increased missing or mis-scored data, no significant difference among treatment mean groupings were found.

Prior to the analysis, we hypothesized that marker ordering would be adversely influenced by increased levels of missing and mis-scored data. We also felt that, at comparable levels, mis-scored data would have more of an influence on marker ordering than would missing data. Over all simulated data sets, surprisingly few marker order changes occurred (38 or 0.38 per data set). Consistent with our *a priori* expectations, 27 (71%) of the changes were found to occur in mis-scored data sets. There does not appear to be any apparent trend in the number of marker order changes for the various levels of missing data. However, for the mis-scored data, the number of marker order changes increased up to the 2% level, beyond which they decreased. Low levels of mis-scoring tend to cause marker order changes, whereas higher levels confound linkages, causing markers to be dropped from the map.

Interestingly, there does not appear to be an additive effect when both missing and mis-scored data are included in the same data set. For example, with 10% missing data, the mean number of markers placed into framework groupings was 65, that is 4.8 fewer than with no missing, and with 2% mis-scored, the mean number placed was 60.7, 9.1 fewer than with no mis-scored. When both were included, the number of markers placed was 64.2. We would have expected this mean to be somewhat lower if the effect of both missing and mis-scored data was additive.

In addition to gaining a better understanding of how missing and mis-scored data effect primary map construction, we were also interested in how trends in the simulated data sets compared with, those seen in actual RAPD data sets. Due to the fact that we do not know the true distances between markers or the level and distribution of mis-scoring in actual RAPD data sets, some of the measures investigated in this paper are not directly comparable. However, we do know the level and distribution of missing values. In a RAPD data set generated for longleaf pine, levels of missing data were found to approach 5%. When compared with simulated data sets with 5% missing data, some general trends appear. For the simulated data sets, on average, 68.6% of the markers could be placed into framework groupings. For longleaf pine, 64.3% of the markers (121 out of 188) were placed into framework groupings. In the simulated data sets, the majority of the markers that are lost as a result of 5% missing data are those harboring the highest levels of missing entries. Only 52.6% of the markers with greater than 15% missing entries mapped, whereas 71.4% of those markers with less than 15% missing entries mapped. For the longleaf data set, only 41.2% of the markers with greater than 15% missing entries mapped, whereas 66.1% of the markers with less than 15% missing entries mapped.

Although the amount of mis-scoring that occurs within an actual RAPD data set is not known, we felt that if mis-scoring were occurring it would be primarily concentrated in the markers with lower confidence scores. For the longleaf pine data set, 70.6% of the markers classified as good were mapped, whereas only 52.2% of the markers classified as fair were mapped. Ms-scoring might possibly be responsible for the difference between these percentages.

CONCLUSION

Based on the variables analyzed in this simulation study, it appears as if levels of missing and mis-scored data as high as 15% and 2%, respectively, can be tolerated during primary genetic map preparation, since they do not significantly effect recombination estimates or linkage group construction. The genetic system simulated in this analysis, 100 markers distributed randomly over a 1600 cM genome, is fairly representative of the situation encountered during the early phases of linkage mapping. We caution, however, that the results from this study are only applicable to low density mapping situations (i.e. early map construction). When more saturated linkage conditions exist the effects of missing and mis-scored data will have more of an impact, particularly on marker ordering.

ACKNOWLEDGMENTS

We appreciate the efforts of M.S. Bowen, T.J. Dean, V.W. Wright, and J. Chambers for reviewing this manuscript. This work was supported in part by funds from McIntire-Stennis

project 2895 and Louisiana Education Quality Support Fund Research and Development LEQSF (1991-1994) RD-A-01. Approved for publication by the Director of the Louisiana Agricultural Experiment Station as Manuscript Number 93-22-7177.

LITERATURE CUED

Kubisiak, T.L., Stine, M., Nelson, C.D., and W.L. Nance. 1992. Single tree RAPD linkage mapping of longleaf pine. Proceedings of Plant Genome I, The International Conference on the Plant Genome, Nov. 9-11, San Diego, CA. p 49.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E., and L. Newberg. 1987. An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174-181.

Nelson, C.D., Nance, W.L., and RL. Doudrick. 1992. A partial genetic linkage map of slash pine (Pines elliottii var. elliottii) based on randomly amplified polymorphic DNAs. Proceedings of Plant Genome I, The International Conference on the Plant Genome, Nov. 9-11, San Diego, CA. p 39.

vanBuijtenen, J.P., Kong, X, Funkhouser, E., Nance, W.L., Nelson, C.D., Nelson, L.S., and G.N. Johnson. 1992. Linkage map of slash pine based on megagametophytic DNA. Proceedings of Plant Genome I, The International Conference on the Plant Genome, Nov. 9-11, San Diego, CA. p 51.

Williams, J.G.K., Kubelik, A.R., Livak, Rafalski, J.A., and S.V. Tingey. 1991. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18:6531-6535.