Theor Appl Genet (2000) 101:323–326

T. Kraft · M. Hansen · N.-O. Nilsson

Linkage disequilibrium and fingerprinting in sugar beet

Received: 25 May 1999 / Accepted: 18 Oktober 1999

Abstract It has been suggested that map information for molecular markers can be used to strengthen finterprinting analyses. The success of this strategy depends on the distribution of linkage disequilibrium over the genome. Using 451 mapped AFLP markers, we investigated the occurrence of linkage disequilibrium in nine sugar beet breeding lines. A low but significant level of linkage disequilibrium was found for unlinked markers. Only for very tigthly linked (<3 cM) markers was this level substantially higher. This implies that little is gained in utilising the map position of the markers in fingerprinting applications.

Key words AFLP · Fingerprinting · Linkage disequilibrium · Sugar beet

Introduction

Fingerprinting with molecular markers has become widely used for characterising germplasm in crops. Most fingerprinting studies are performed using markers chosen randomly without consideration of their genomic location. However, with the availability of dense marker maps in most crops, it is possible now to choose markers in a more structured way.

At least two advantages in utilising map information have been proposed. First, markers evenly distributed over the entire genome can provide more accurate estimates of genetic similarities different accessions (e.g. Song et al. 1995; Powell et al. 1996; Noli et al. 1997). Secondly, mapped markers provide an opportunity for studying the relationship between accessions separately for different genomic regions. The latter approach has been termed map-based fingerprinting (Zhu et al. 1998, 1999; Hansen et al. 1999). Both approaches are based on the idea that while for interbred materials the genealogical tree differs for different regions of the genome, closely linked markers share the same genealogy (Kaplan and Hudson 1985). Thus, linked markers are assumed to be highly correlated, i.e. to be in linkage disequilibrium. Unfortunately, very little is known about the distances spanned by linkage disequilibria in breeding materials.

Linkage disequilibrium is created in breeding materials when several lines become fixed for a given set of alleles at a number of different loci. Alleles can become fixed within the lines due to the random action of genetic drift or to hitchhiking effects during selection. One should note that any pair of loci in linkage disequilibrium may nevertheless be in equilibrium with loci located between them. When lines are crossed, different haplotypes come together in the same individuals. Recombination can then begin breaking up the linkage disequilibrium in succeeding generations. This happens very rapidly for unlinked loci but much more slowly for loci that are tightly linked. If groups of lines with common ancestries are kept separate, linkage disequilibrium for unlinked loci can remain.

We investigated the relation between linkage and linkage disequilibrium among mapped AFLP markers in a selection of sugar breet breeding lines. The implications of our fingerprinting results are exemplified and discussed.

Materials and methods

Nine inbred sugar beet accessions representing a broad variation within the breeding pool of Novartis Seeds AB were employed in the study. The lines had previously been fingerprinted using 11309 AFLP markers (Hansen et al. 1999). The present analysis of linkage disequilibrium relied on a previous mapping of AFLP markers in an F_2 population of 237 individuals (Nilsson et al. 1999). All the data from 451 mapped AFLP markers in the nine breeding lines selected were recorded twice as the presence or the absence of bands.

Linkage disequilibrium between markers *i* and *j* was measured as $D' = \left| \frac{P_{i,j} - p_i p_j}{D_{\max}} \right|$ where P_{ij} is the frequency of accessions in

Communicated by P. Langridge

T. Kraft (✉) · M. Hansen · N.-O. Nilsson
Novartis Seeds AB, P.O. Box 302, S-261 23 Landskrona, Sweden
e-mail: thomas.kraft@seeds.novartis.com
Fax: +46-418-437283

which the band is present both at locus *i* and *j*, and p_i and p_j are the frequencies of the band at locus *i* and *j*, respectively. D_{max} is the highest possible value for the numerator, given the band frequencies at locus *i* and *j*. Since one sample from each line was studied, D' measures linkage disequilibrium between the lines and not between individuals within the lines. It is also based on the assumption that the lines are completely homozygous. The banding pattern of a specific line corresponds then to a particular haplotype. Since this assumption is clearly false the variance of the estimates is elevated, but no systematic bias in the estimates is introduced. Markers for which the band is present or absent in one accession only contain virtually no information for estimating linkage disequilibrium and were therefore excluded from analysis (Lewontin 1995). All measures of D' were classified according to the map distance between the marker loci. Confidence intervals for the mean D' for each class were determined by bootstrap analyses (Weir 1996). Bootstrap datasets were created by random sampling, with replacement, of the same number of observations as in the original dataset. The mean was calculated for each bootstrap dataset, and the distribution of all means was used to estimate confidence intervals.

The genetic distance between lines *i* and *j* was calculated as $1-\frac{2x_{i,j}}{x_i + x_j}$, where x_{ij} is the number of loci with the band present in both lines i and j, and xi and xj are the number of loci with the band present in lines i and j, respectively (Nei 1987). This estimate takes into account only the presence of a band in both lines as a contribution to similarity, not its absence in both lines. This is an advantage since for AFLP markers the presence of a band in two individuals is more likely to represent a true homology than is the absence of a band. Dendrograms based on estimates of genetic distances were obtained using the UPGMA method contained in the PHYLIP program package (Felsenbein 1993). Support values for the nodes in the dendrograms were estimated by bootstrap analysis. Bootstrap datasets were created for each dendrogram by random sampling, with replacement, of the same number of markers as in the original dataset. The number of bootstrap datasets resulting in a certain node is a measure of how well supported that node is by the data.

Results and discussion

The level of linkage disequilibrium for all possible pairs of 451 mapped AFLP markers was measured in nine inbred sugar beet breeding accessions. For unlinked AFLP markers, the average D' was found on the basis of 54 10739 observations to be 0.414. Since it is the absolute value of D' which we study, the expectation when no linkage disequilibrium is present is not zero. The expectation was estimated by permutations through calculating D' for all pairs of loci while randomising the individuals for one of the markers in each pair, thus breaking up all linkage disequilibria in the data. The expectation of D', estimated as the mean of 100 replications of the permutations, was 0.372. Since the maximum value for the 100 replicates was 0.378, there is significant linkage disequilibrium for the unlinked markers in the material investigated. If no linkage disequilibrium for the unlinked markers were present, this would imply that the genetic distances, average over the entire genome, would be the same for all pairs of lines.

Estimates of D' for the linked markers were studied for different intervals of map distance between the markers (Table 1). Pairs of markers with a map distance of less than 3 cM had a significantly higher average D' than unlinked markers. At still shorter distances, the average level of linkage disequilibrium increased, being especially high for pairs of markers with a map distance of less than 0.2 cM. Since the map distances are based on a mapping population of only 237 individuals, a good resolution could not be achieved for the short distances involved. It is likely that, of the pairs of markers that mapped to the same position, it was the most tightly linked pairs that contributed most to the increase in linkage disequilibrium. In a similar analysis (unpublished) of six sugar beet lines, using 251 mapped restriction fragment length polymorphism (RFLP) markers, we found the same basic pattern, an increase in the level of linkage disequilibrium only being present for distances of less than 3 cM.

Table 1 Mean linkage disequilibrium (D') and 95% confidence interval for different intervals of map distances between the markers. *N* is the number of observations obtained for each interval. Confidence intervals were estimated by bootstrapping the data 2000 times

Map interval (cM)	*N*	D'	Confidence interval
0–0.01	1026	0.584	0.563–0.605
0.2–1.0	1155	0.509	0.490–0.528
1.1–3.0	1309	0.449	0.431–0.468
3.1–5.0	785	0.385	0.364–0.405
5.1–10.0	950	0.396	0.377–0.416
10.1	1915	0.414	0.400–0.428
Unlinked	54739	0.414	0.412–0.417

The increase in linkage disequilibrium for very short map intervals could have been created during the breeding history of sugar beet. All lines with a common ancestry are fixed for one of the alleles in the ancestral line or population. Since the alleles in the ancestral population are only a subset of the alleles found in the total material, the probability of fixation of the same multilocus haplotype in several of the descendent lines is increased. Consider an ancestral line, for example, for which two loci are fixed for alleles that are otherwise rare in the breeding germplasm. All lines descending from this ancestral line are necessarily fixed for this particular two-locus genotype. The genotype's frequency in the total breeding germplasm is thus increased as compared with what would be expected if there was no association between the loci. This is similar to the creation of linkage disequilibrium in a natural population when this is divided up into several subpopulations between which only limited migration occurs (Hedrick et al. 1978; Ohta 1982). It is likely that most polymorphisms in sugar beet are very old, having originated far back in the history of wild beets. Thus, the increase for short distances could likewise be a remnant of the linkage disequilibrium created already in the wild beet progenitors of sugar beet.

Recently, Zhu et al. (1999) reported a map-based fingerprinting study of rixe varieties. They divided the genome into "DNA fingerprinting linkage blocks" of varying size (approx. 10–50 cM) and analysed the relation-

ships among the varieties separately for each block. The relationships among the varieties varied among the blocks, which was interpreted as evidence for the effectiveness of map-based fingerprinting. However, it was not shown whether the variation among the blocks was greater than expected if each "block" comprised a similar number of markers chosen at random from the whole genome. We find that linkage disequilibrium only spans very small distances, which demonstrates that numerous, more or less independent genealogies exist in the genome. This has various implications for fingerprinting applications. First, map-based fingerprinting is impractical, since such a large number of different genetic distance matrices would need to be calculated. It would also be impossible in most cases to find sufficient markers for the small map intervals in which an increase in the level of linkage disequilibrium is evident. Secondly, in order to adequately estimate genetic distances by averaging these over the entire genome, a large number of markers would need to be sampled, since the sampling variances would be high. Finally, the gain achieved in using markers that are evenly distributed over the genome would be small because it is unlikely that randomly chosen markers would be more closely correlated to one another than markers chosen as a scaffold. Each marker appears, therefore, to represent only one genealogy of the many that are present in the genome, regardless of whether markers are chosen at random or as a scaffold.

The first step in fingerprinting is often to calculate the genetic distances between all pairs of lines and to display the resulting matrix by drawing dendrograms using the UPGMA method. It has repeatedly been found to be difficult to reliably resolve the relationships between closely related lines, even when a large number of markers are employed (e.g. Powell et al. 1996; Milbourne et al. 1997; Schut et al. 1997; Barrett and Kidwell 1998). The question of the number of loci needed to obtain high support values for the nodes in a dendrogram of ten sugar beet breeding lines was addressed by Hansen et al. (1999). Their study was based on more than 11 10000 AFLP loci. When subsets of 2000 loci were employed, most nodes showed support values of higher than 85%. However, one node had a support value of as low as 69%, even when all loci were included in the analysis. Such difficulties are scarcely surprising in view of the present results, which indicate that, except for very short distances, only low levels of linkage disequilibrium are found. This, in turn, shows that there are numerous, more or less independent genealogies in the genome.

The drawing of trees based on genetic distances involves two assumptions: (1) all the lines that are joined at a given node are more similar to each other than to any other line and (2) all of them have the same genetic distance to any given line outside the cluster. These assumptions could be true if the accessions were kept separate for a long period of time, such as for species in a phylogenetic tree. For various populations within a species or for lines within a breeding material, however, there is no reason why the genetic distance matrix should

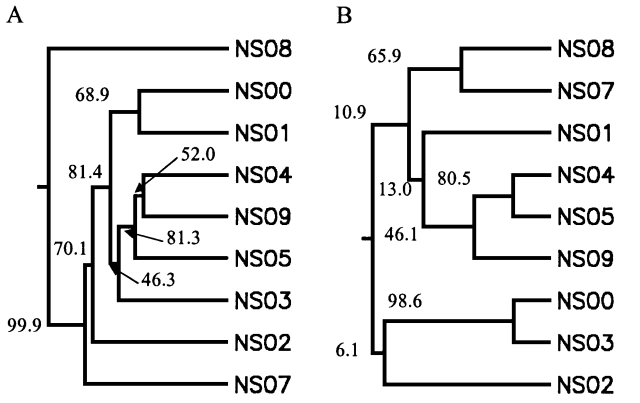

Fig. 1A, B Dendrogram based on UPGMA analysis of nine sugar beed lines analysed either using all the 381 AFLP markers (**A**) or using the 33 markers that map within a particular 1-cM region (**B**). Support values estimated by bootstrapping the data 10 0000 times are given for each node

fit such a tree structure. Nevertheless, this method is widely employed for materials of this sort and can be useful for discovering distinct groups of lines. Various examples of the implications that our results have for the construction of dendrograms can be shown. A dendrogram based on all 451 mapped markers is shown in Fig. 1A. The same set of lines was included earlier in an analysis involving more than 11 000 unmapped loci, revealing similar topologies (Hansen et al. 1999). This can be compared with an example in our material of map-based fingerprinting for a region containing a sufficient number of tightly linked markers. For one linkage group, there were 33 polymorphic markers that all mapped within a 1-cM interval; the dendrogram based on these markers is shown in Fig. 1B. Some nodes in this dendrogram have high support values but are inconsistent with a topology of the dendrogram based on all the markers as a whole. This demonstrates clearly that while the genealogies for different genomic regions differ, very few regions that are sufficiently small have enough markers for accurate estimates of genetic distances to be obtained.

We also compared the alternatives of choosing markers as a scaffold and of choosing them at random. The markers were ordered according to their map position and were divided into 92 groups of 5 markers each. A scaffold of markers was created by randomly choosing one marker from each group. Parallel to this, a dataset was created by randomly choosing 92 markers without regard to their map positions. Dendrograms were constructed for both sets of data. This was repeated 10000 times, and the number of times that a particular node was identical to that of a consensus tree based on all the markers was noted (Fig. 2). Both selection methods resulted in the same consensus tree, there being only a very slight increase in the probability of finding of "correct" tree when a scaffold of markers were selected. Hence, very little is gained in using a scaffold of markers as compared with using random markers.

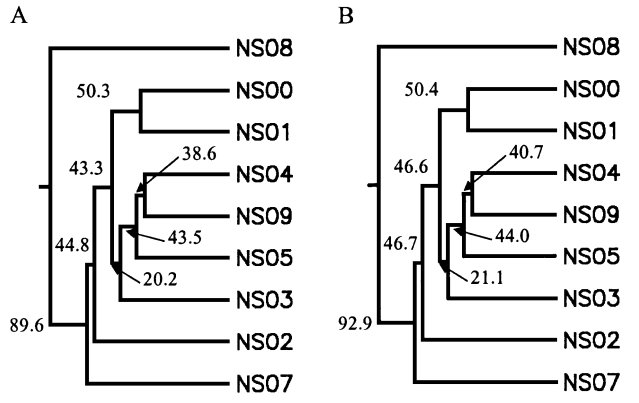

Fig. 2A, B Dendrogram based on UPGMA analysis of nine sugarbeet lines analysed using 92 AFLP markers either chosen at random (**A**) or evenly distributed over the genome (**B**) (see the main text). The probability of getting the same topology as for a dendrogram based on all the markers is given at each node

Conclusions

Considerable precaution should be exercised in interpreting fingerprinting results for interbred materials such as breeding lines. Well-defined heterotic groups of lines can usually be detected by the use of markers, since many of the markers are in linkage disequilibrium. Within such groups, however, the levels of linkage disequilibrium are much lower, making it difficult to assess accurately the genetic distances between the lines. In cases of this sort little is gained by taking account of the map position of the markers, since only for very short distances is there an increase in the level of linkage disequilibrium. The generality of our results also for other crop species needs to be investigated.

Acknowledgements We would like to thank Magnus Nordborg for his helpful comments on the manuscript. The research was supported by the Swedish Strategic Network for Plant Biotechnology.

References

Barrett BA, Kidwell KK (1998) AFLP-based genetic diversity assessment among wheat cultivars from the Pacific northwest. Crop Sci 38:1261–1271

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, Wash

Hansen M, Kraft T, Christiansson M, Nilsson N-O (1999) Evaluation of AFLP in *Beta*. Theor Appl Genet (in press)

Hedrick P, Jain S, Holden L (1978) Multilocus systems in evolution. Evol Biol 11:101–184

Kaplan N, Hudson RR (1985) The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. Theor Pop Biol 28:382–396

Lewontin RC (1995) The detection of linkage disequilibrium in molecular sequence data. Genetics 140:377–388

Milbourne D, Meyer R, Bradshaw JE, Baird E, Bonar N, Provan J, Powell W, Waugh R (1997) Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. Mol Breed 3:127–136

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nilsson N-O, Hansen M, Panagopoulos AH, Tuvesson S, Ehlde M, Christiansson M, Rading IM, Rissler M, Kraft T (1999) QTL analysis of *Cercospora* leaf spot resistance in sugar beet. Plant Breed 118:327–334

Noli E, Salvi S, Tuberosa R (1997) Comparative analysis of genetic relationships in barley based on RFLP and RAPD markers. Genome 40:607–616

Ohta T (1982) Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc Natl Acad Sci 79: 1940–1944

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey A, Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238

Schut JW, Qi X, Stam P (1997) Association between relationship measures based on AFLP markers, pedigree data and morphological traits in barley. Theor Appl Genet 95:1161–1168

Song K, Slocum MK, Osborn TC (1995) Molecular marker analysis of genes controlling morphological variation in *Brassica rapa* (syn. *campestris*). Theor Appl Genet 90:1–10

Weir BS (1996) Genetic data analysis. II. Sinauer Assoc, Sunderland, Mass

Zhu J, Gale MD, Quairrie S, Jackson MT, Bryan GJ (1998) AFLP markers for the study of rice biodiversity. Theor Appl Genet 96:602–611

Zhu JH, Stephenson P, Laurie DA, Li W, Tang D, Gale MD (1999) Towards rice genome scanning by map based AFLP fingerprinting. Mol Gen Genet 261:184–195