Theor Appl Genet (1999) 98: 113–118

R. Van Berloo · P. Stam

Comparison between marker-assisted selection and phenotypical selection in a set of *Arabidopsis thaliana* recombinant inbred lines

Received: 29 November 1997 / Accepted: 8 June 1998

Abstract Parents were selected from a well-characterised *Arabidopsis* recombinant inbred line (RIL) population based on (1) their phenotype for flowering time or (2) marker and QTL information that had been assessed previously. The F_2 offspring obtained from pairs of selected RILs was analysed for these traits, and the results obtained with these two methods of selection were compared. Selection based on marker and QTL information gave approximately the same result as selection based on phenotype. The relative high heritability of flowering time in *Arabidopsis* facilitated successful phenotypical selection. The difference in selection result that was anticipated to be in favour of the marker-assisted approach was therefore not observed.

Key words Marker-assisted selection · Quantitative trait loci · *Arabidopsis* · Selection methods

Introduction

Marker and QTL information obtained from a segregating population can be used for the design of efficient breeding strategies. In recent years major advances in marker availability and statistical methods for assessing marker-trait correlations have been achieved (e.g. Lander and Botstein 1989; Jansen and Stam 1994; Falconer and Mackay 1996). Marker-assisted selection (MAS) has been advocated as a useful tool for rapid genetic advance in the case of quantitative traits (Lande and Thompson 1990; Knapp 1994, 1998). In a previous paper (Van Berloo and Stam 1998) we describe a procedure for the application of MAS to an autogamous population of recombinant inbred lines (RILs). In this paper we report experiments using *Arabidopsis* as a model species. *Arabidopsis* is well-suited for model selection experiments because of its small size and short generation cycle (Meyerowitz and Pruitt 1985). Over the years a vast body of genetic data on *Arabidopsis* has become available. Kuittinen et al. (1997) described a quantitative trait locus (QTL) mapping experiment for flowering time in *Arabidopsis*. Five to seven QTLs affecting flowering time were found in a BC_1 population derived from the Finnish Naantali genotype and the German strain Li-5. In a different population, consisting of 165 Ler × Cvi RILs, Alonso-Blanco et al. (1997) found four QTLs affecting flowering time. Jansen et al. (1995) used the *Arabidopsis* RIL set, obtained from a cross between the Columbia (Col) and Landsberg *erecta* (Ler) ecotypes (Lister and Dean 1993), in a QTL mapping experiment involving various environments. Day length was varied, and in some cases a vernalisation treatment was applied. In this experiment 12 QTLs for flowering time were detected, of which 8 had an effect in all environments and 4 showed an effect in only some of the environments.

In this paper we describe an experiment using the Col × Ler *Arabidopsis* RILs of Lister and Dean (1993). The objective was to compare a MAS breeding strategy, using molecular marker and QTL information, with conventional breeding methods, based on phenotype only. The focus lay on the selection of suitable parents for crossing. The F_2 offspring derived from these parents was the target generation in which the quality of selection was evaluated. In both MAS and phenotypical selection procedures the target was the production of genotypes that contain as many as possible advantageous alleles for the QTLs that affect the trait of interest (these will be referred to as 'superior' or

Communicated by H. C. Becker

R. Van Berloo (✉) · P. Stam
CT de Wit Graduate School for Production Ecology,
Laboratory of Plant Breeding Wageningen Agricultural University,
PO Box 386, Wageningen, The Netherlands
Fax: + 31-317-483457
E-mail: ralph.vanberloo@users.pv.wau.nl

Table 1 Presence of QTL alleles for earliness in RIL set assessed through graphical genotype analysis

Number of 'earliness' QTL alleles	0	1	2	3	4	5	6	7	8
Frequency within set of RILs	0	1	9	26	27	21	12	2	0

'extreme' genotypes). In this case, the trait of interest was flowering time.

Materials and methods

Plant material

The Col × Ler *Arabidopsis* RIL set, consisting of 99 lines, was obtained from the *Arabidopsis* stock centre in Nottingham, UK. The set of RILs was developed by Lister and Dean (1993) and was derived, through single-seed descent, from an F_2 population that resulted from a cross between the Landsberg *erecta* and Columbia ecotypes. In our experiment we identified the lines according to the Arabidopsis stock centre line numbers, using RIL numbers from 1900 to 1998.

Trait

Our trait of interest was flowering time. Flowering time is generally regarded as a quantitative trait which may influence other traits (Kuittinen et al. 1997). Flowering time is measured as the number of days from planting of the germinating seeds until the first petal becomes visible. Scoring of flowering time is approximated by using 1-day classes.

Marker and QTL data

Restriction fragment length polymorphism (RFLP) marker data for all 99 RILs were obtained from Jansen et al. (1995). These data were used to construct a genetic map using the JOINMAP package (Stam and Van Ooijen 1995). This map corresponded with the integrated genetic map, which is freely available on the Internet.[1] From 12 flowering-time QTL estimates, obtained from Jansen et al. (1995, and personal communication), 8 that had a significant effect under long-day conditions without seed vernalisation were selected for marker and QTL analysis. In our experiment we used the same set of RILs as Jansen et al. (1995).

Graphical genotypes

The RILs were subjected to analysis using graphical genotypes (Young and Tanksley 1989). Marker data for all RILs were displayed graphically using a different colour for each parent of the RIL population (Col/ Ler). For analysis the computer programme GGT[2] was used. When markers indicated that a chromosomal region at a QTL was of the same origin as the parent that contributed the favourable allele it was assumed that the RIL inherited this allele. In this way the number of favourable QTL alleles present could be assessed for all RILs. The distribution of the number of favourable QTL alleles for early flowering over the RILs is listed in Table 1. Columbia contained three favourable QTLs for earliness and Landberg *erecta* five. While none of the RILs contained all of the favourable alleles for early flowering, all of the RILs contained at least one favourable allele for this trait.

Selection

Arabidopsis is a self-fertilising species (Abbot and Gomez 1989). Therefore, the selection result should be a single genotype or line that contains as many favourable QTL alleles as possible. The procedure used for obtaining this 'extreme' genotype was the same as the one we applied in earlier simulation studies (Van Berloo and Stam 1998). Basically, the method identifies those pairs of RILs which, upon crossing, give rise to a high number of superior QTL-genotypes among their F_2 offspring. This is done by preselecting RIL pairs on the basis of their marker-genotype and subsequently simulating their F_2 offspring. Selection for flowering time was aimed in two directions, for late flowering and for early flowering.

Two criteria were used to select RIL combinations for crossing: (1) the predicted breeding potential of a line pair based on marker and QTL data, and (2) the observed line phenotype.

Predicted breeding potential

The available marker and QTL data were used by MS, the computer programme for MAS, which identifies line pairs that have a high probability of accumulating favourable QTL alleles in F_2-offspring genotypes (Van Berloo and Stam 1998). The programme was run with marker and QTL data from the 99 RILs. This resulted in a list of preferable crosses.

Observed phenotype

RILs were ordered according to their phenotype (calculated as an average over 24 plants). Next, a subset of RILs comprising the extreme 10% were selected. Within this subset line pairs were selected at random for crossing.

Out of a possible 4851 (1/2*99*98) pairs, 25 were selected using MAS and 25 were selected based on their phenotype. We harvested seeds from 14 'MAS crosses' and 17 'phenotypic crosses'. A subset of 11 F_1s from MAS crosses and 12 F_1s from phenotypic crosses were selfed to obtain F_2 seeds. F_2 plants from 4 MAS crosses and 4 phenotype-based crosses were evaluated in a greenhouse trial.

Experimental setup

All plants were grown in the same greenhouse under long-day conditions (18 h light, 6 h dark). Seeds were not vernalised before sowing, but the germinating seeds were allowed 48 h at 4°C to break dormancy.

[1] Nottingham Arabidopsis stock centre, Nottingham, UK; URL: http://nasc.life.nott.ac.uk/

[2] A paper describing the GGT package for display and selection using graphical genotypes will be published in the Journal of Heredity. For more information on the GGT package, contact the author via e-mail

Per line 24 plants were grown in two replications. Lines were randomised within a replication. Flowering time of the RILs was observed. Selected line combinations (see selection paragraph for criteria) were crossed, and their F_1 seeds were harvested.

Next, F_1 seeds from 23 selected crosses were grown without replications. On average 12 plants per cross were grown. Plants were allowed to self-fertilise, and F_2 seeds were harvested.

For each of the four categories two crosses were selected (see Table 2). Each selected cross was represented by 200 F_2 plants that were grown in a greenhouse trial. As a control 800 plants from the RIL set were grown. Four RILs were selected to represent the RIL set, 1 early-flowering and 1 late-flowering RIL, and 2 RILs of moderate flowering time. The experimental setup was a block design with 17 blocks. Plant rows were randomised within blocks, and blocks were randomised over the greenhouse. For each of the 2400 plants the flowering time and the number of leaves at the time of flowering were recorded.

Data analysis

The observations on the 2400 plants were used to obtain estimates for population average and variance. This was done using the statistical computer package ASREML,[3] provided by Gilmour et al. (1995). ASREML allows the estimation of population variances and their standard errors. A square-root transformation was applied to the discrete data in order to obtain a normal distribution of residuals. The model fitted to the data was:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \underline{h}_{jk} + \underline{e}_{ijk}$$

with α_i = contribution of blocks, β_j = contribution of population mean, $\underline{h}_{jk}$ = contribution of specific plant genotype and $\underline{e}_{ijk}$ = remaining error term.

Since the controls (RILs) are genetically homogeneous (within lines), the variance within these controls (averaged over RILs) was used to assess the environmental variance. The genetical component of the F_2 population variances was obtained by subtracting the environmental variance from the experimental variance. Heritability was estimated as the ratio of the genetic and phenotypic variance. Since we were interested in plants within the populations that possess 'extreme' or superior genotypes we considered the 95% percentiles of the distribution of the F_2 populations. From statistical theory (e.g. Levert 1959) it is known that the 95% confidence interval for the 95% percentile of a normal distribution (x_p) can be found by:

$$\hat{\mu} + 1.45\,\hat{\sigma} < x_p < \hat{\mu} + 1.88\,\hat{\sigma}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation, respectively.

Confidence intervals for the 95% percentile of the F_2 phenotypic distribution were estimated for each cross.

Results

Figure 1 shows a scatter plot of RIL flowering time (phenotypic value) versus the number of favourable QTL alleles present in the RILs. RILs that are part of pairs that were selected by MAS or phenotypical selection are highlighted. Phenotypical selection was less successful than MAS in selecting RILs with the highest number of favourable QTL alleles. RILs selected by MAS showed a less extreme trait value. This was expected because these RILs were selected for their ability to complement each other genetically, not because they showed a high trait value themselves. For reasons of simplicity, no effect of the sizes of the QTL alleles was taken into account in Fig. 1. Therefore, caution should be taken in making comparisons between data points. A large difference in the number of QTL alleles does not necessarily result in an equally large difference in genetic potential.

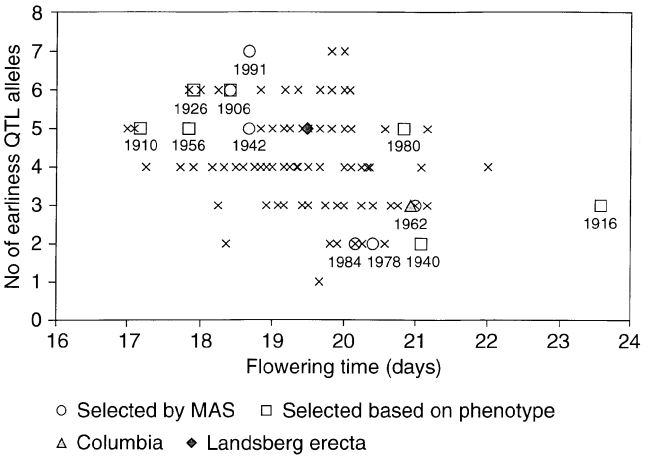

Fig. 1 Scatter plot of the number of QTL alleles favourable for early flowering versus the realised flowering time of the RILs. RILs that were selected by MAS or phenotypical selection are indicated separately

The RILs showed a continuous, unimodal phenotypic flowering-time distribution. Extreme flowering times were 13 and 27 days; RIL means ranged between 17 and 24 days.

Table 2 shows the RIL pairs that were selected, and the associated prediction value that resulted from the model prediction. The F_1 plants showed a clear distinction between the group selected for late-flowering and the group selected for early flowering, as was expected (data not shown). In the F_2 populations we observed plant flowering times ranging from 26 to 52 days. The estimates of population means, standard deviations and heritabilities are shown in Table 3. The average heritability for flowering time over all populations was 0.34. Two distinct groups of crosses emerged: an early-flowering group and a late-flowering group. When the results within these groups were compared, the differences were less clear, and no significant differences between phenotypically selected crosses and MAS crosses were observed.

The 95% percentile was used as a parameter for comparison between the tails of normal distributions. Confidence intervals ($\alpha = 0.05$) were calculated for the 95% percentile of each population. The right percentile was used for the 'late' crosses and the left percentile for

[3] The ASREML package is freely available through the Internet via anonymous FTP: ftp.res.bbsrc.ac.uk in pubs/uploads/aar

Table 2 RIL pairs that were selected for crossing by the different selection methods and the prediction for ability to produce extreme F_2 offspring

Selection type[a]	RIL pair	Allele types[b]	Prediction[c]
ME	1991 × 1906	EEEEEELE × LEEEEELE	100
	1942 × 1991	LLEEEELE × EEEEEELE	98
PE	1926 × 1906	EEEEEELL × LEEEEELE	94
	1956 × 1910	ELEEEELL × LEEELEEL	70
ML	1962 × 1984	LLLLELEE × LLLLELEL	73
	1978 × 1984	LLLLLELE × LLLLELEL	75
PL	1916 × 1940	LLELEELL × LLLLELLE	44
	1916 × 1980	LLELEELL × LLLEEEEE	29

[a] ME = MAS, early flowering; PE = phenotypic selection, early flowering; ML = MAS, late flowering; PL = phenotypic selection, late flowering
[b] Allele types indicate the QTL alleles for the eight QTLs listed as Parent-1 × Parent-2 = $Q_1Q_2Q_3Q_4Q_5Q_6Q_7Q_8 \times Q_1Q_2Q_3Q_4Q_5Q_6Q_7Q_8$. E, Early allele; L, late allele
[c] Prediction based on the average of ten replicates of extremes found by computer simulations of 100 F_2 progeny. Predictions, indicating RIL pair potential for obtaining extremes, range between 0 and 100, 100 being the highest possible value, according to the direction of selection

Table 3 Flowering-time means, standard deviations and heritabilities for F_2 populations obtained after marker-assisted selection or phenotypical selection for either late or early flowering

		√(Flowering time)[a]				
Population	Type[b]	S.Q.[c]	$\hat{\mu}$	$\hat{\sigma}_g$	$\hat{\sigma}_e$	h^2
1991 × 1906	ME	1	5.70 a	0.042	0.15	0.07
1942 × 1991	ME	2	5.75 a	0.085	0.15	0.24
1926 × 1906	PE	2	5.68 a	0.108	0.15	0.34
1956 × 1910	PE	5	5.53 a	0.060	0.15	0.14
1962 × 1984	ML	1	6.12 b	0.156	0.15	0.52
1978 × 1984	ML	4	6.04 b	0.159	0.15	0.53
1916 × 1940	PL	3	6.08 b	0.115	0.15	0.37
1916 × 1980	PL	4	6.04 b	0.143	0.15	0.48

[a] $\hat{\mu}$: Mean of F_2 population; a and b indicate groups that show a significant difference at $\alpha = 0.05$; $\hat{\sigma}_g$, Estimated genetic standard deviation; $\hat{\sigma}_e$, estimated environmental standard deviation; h^2, observed heritability of the transformed trait (F_2)
[b] See legend of Table 2
[c] S.Q. = The number of segregating QTLs, derived from graphical genotype analysis, see Table 2

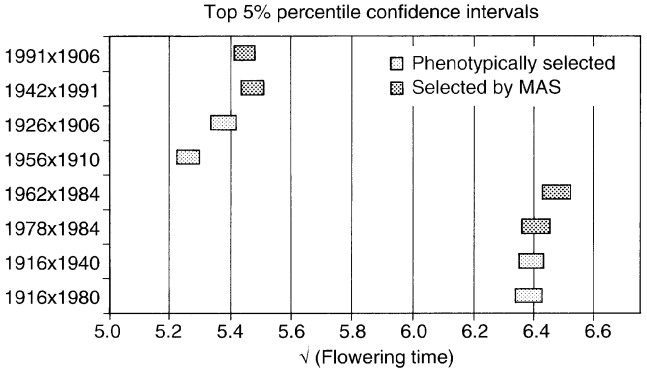

Fig. 2 Confidence intervals (95%) for the *right* ('late' selections) and *left* ('early' selections) 95% percentile of the F_2 flowering time distributions

the 'early' crosses. Confidence intervals are drawn in Fig. 2. Figure 2 again shows that selection has led to two distinct groups: a late- and an early-flowering group. However, within such a group no large differences between selection methods could be seen. Within the 'late' group, the MAS confidence intervals lay more in the direction the selection was aiming for than the other confidence intervals, while in the 'early' group the reverse situation was true. Most confidence intervals of the different selection methods overlapped.

Discussion

This experiment was aimed at a comparison of two different selection methods. The source of information, on which selection was based, was different for each method. Marker-assisted selection used only marker data and information on QTL locations obtained from previous experiments to predict useful crosses.

Phenotypical selection used plant phenotypic data that were collected in an additional experiment. The final results did not favour one selection method over the other.

Although we expected the marker-assisted selection procedure to be more efficient in obtaining extreme phenotypes in an F_2 progeny resulting from crossing selected parents, the results of this experiment did not confirm this expectation. This may be due to the nature of the trait we investigated. In our experiment, we found an average heritability for F_2s of 0.34 for flowering time. Assuming absence of dominance, conversion into a heritability for RILs would yield about 0.7, and this heritability may well be too high to take full advantage of marker-assisted selection. Benefits of the MAS procedure are to be expected only when the trait heritability (calculated for RILs) lies approximately within the range of 0.1–0.3 (Van Berloo and Stam 1998). When the heritability is too high, the cost involved in genotyping many plants may not outweigh the expected benefits of more direct gene selection. On the other hand, when the heritability drops below 0.1, the QTLs cannot be identified with the accuracy required to rely on flanking markers for selection.

One of the main theoretical reasons why MAS outperforms phenotypical selection (PS) is that RIL pairs selected by MAS will generate, on average, more genetic variance in the offspring because such RIL pairs will tend to be complementary with respect to QTL alleles. In our experiment, however, this advantage of MAS over PS was, in hindsight, not realised. From Table 3 it can be seen that there is no clear relationship between the estimated genetic variance and the number of segregating QTLs in a cross. There are possible explanations for the absence of such a relationship. First, the size of the effects may vary among QTLs; since different sets of QTLs are segregating in the crosses, this does not necessarily result in a larger genetic variance as the number of segregating QTLs increases. Second, apart from the identified QTLs, other genes affecting flowering time may be segregating in each cross, inducing additional genetic variance. Although the true cause is unknown, it is obvious that these disturbing factors may have influenced the performance of MAS.

The RILs selected by MAS showed, on average, a lower phenotypic value and a higher genotypic value than the RILs selected on the basis of their phenotype, but the differences were small. We conclude that both methods of selection have succeeded in obtaining RIL pairs that are roughly equal with respect to their breeding potential. In fact, the prediction scores, presented in Table 2, seem to corroborate this for the early-flowering selection.

This experiment showed that we were able to successfully obtain transgression in offspring populations from selected crosses. Maximum observed flowering time in the F_2 populations was twice the maximum value observed in the RIL population. Since these populations were not grown in the same experiment we should be cautious when comparing them. Nevertheless, it is clear that the MAS procedure that we used can be applied successfully in other cases as well.

Our MAS procedure (Van Berloo and Stam 1998) can be seen as aiming at the efficient pyramiding of favourable QTL alleles that are present in a choice of sources, i.e. the RIL set. In both our simulation study and the experimental verification described in this paper, we have dealt with a single trait supposedly controlled by non-epistatic QTLs. Since QTLs were mapped in a set of RILs, no dominance effects could be detected. Had we been able to detect and use dominance at QTLs this would most likely have influenced the selection of RIL pairs in MAS. It is quite conceivable that, in the case of non-additivity of QTL effects, pyramiding QTLs based on the phenotype of the parents will be less efficient than pyramiding based on QTL flanking markers. In our previous paper this was demonstrated using simulated data. Although not the subject of this study, another example in which the MAS approach will outperform phenotypic selection is the accumulation of disease resistance (R) genes, when beyond a given number of R-genes the addition of more of them does not lead to an observable increase in phenotypic resistance. In that case pyramiding R genes beyond a phenotypically observable threshold may nevertheless be useful to enhance the durability of the resistance.

Although in our experiment the results of MAS fell a little short of our expectations, our experiment clearly demonstrates an important, more general, point – the potential usefulness of publicly available data on linkage maps and putative QTL positions for breeding purposes. Today this type of data is accumulating at a high rate. Applied plant breeders as well as the scientific community can, and should, take advantage of it. In the present paper we have considered a single, simple trait, controlled by only a few QTLs. It needs little imagination to realise that in a more realistic setting of plant breeding, where many traits are to be considered simultaneously, knowledge about QTLs and their map positions will be of great help in designing and optimising scenarios for the accumulation of favourable QTL alleles by crossing and marker-assisted selection.

Acknowledgements The authors want to thank the Nottingham Arabidopsis stock centre for providing the RIL seeds, Corrie Hanhart and Dr. Carlos Alonso-Blanco for sharing their knowledge on the growing of *Arabidopsis* and Dr. Ritsert Jansen from CPRO-DLO for providing marker and QTL data.

References

Abbott RJ, Gomes MF (1989) Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity 62:411–418

Alonso-Blanco C, El-Assal S, Peeters AJM, Koornneef M (1997) Genetic analysis of developmental traits based on the Ler/Cvi natural variation. In: 8th Int Conf *Arabidopsis* Res. Stanford University Press, Madison, Wis., p 12–1

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. 4th edn. Longman Group, Essex, UK

Gilmour AR, Thompson R, Cullis BR (1995) Average Information REML, an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51 : 1440–1450

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136 : 1447–1455

Jansen R, Ooijen JW van, Stam P, Lister C, Dean C (1995) Genotype by environment interaction in genetic mapping of multiple quantitative trait loci. Theor Appl Genet 91 : 33–37

Knapp SJ (1994) Selection using molecular marker indexes. In: Proc 2nd Plant Breed Symp Crop Sci Soc Am Am Soc Hortic Sci. Am Soc Hortic Sci, Alexandria, Va., pp 1–11

Knapp SJ (1998) Marker assisted selection as a strategy for increasing the probability of selecting superior genotypes. In: Abstracts Plant Anim Gen Conf VI. Scherago International, San Diego, Calif., p PA5 (www.intl.pag.org/pag/6/abstracts/knapp.html)

Kuittinen H, Sillanpää MJ, Savolainen O (1997) Genetic basis of adaptation: flowering time in *Arabidopsis thaliana*. Theor Appl Genet 95 : 573–583

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124 : 743–756

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121 : 185–199

Levert C (1959) Een nomogram voor betrouwbaarheidsgrenzen van quantielen en overschrijdingskansen bij normaal verdeelde grootheden. Stat Neerl 13 : 3–14

Lister C, Dean C (1993) Recombinant Inbred Lines for mapping of RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J 4 : 745–750

Meyerowitz EM, Pruitt RE (1985) *Arabidopsis thaliana* and plant molecular genetics. Science 229 : 1214–1218

Stam P, Van Ooijen JW (1995) JOINMAP (tm) version 2.0. software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen, The Netherlands

Van Berloo R, Stam P (1998) Marker assisted selection in autogamous RIL populations: a simulation study. Theor Appl Genet 96 : 147–154

Young ND, Tanksley SD (1989) Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor Appl Genet 77 : 95–101