Theor Appl Genet (1999) 98: 1279–1292

D. L. Remington · R. W. Whetten
B.-H. Liu · D. M. O'Malley

Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*

Received: 7 August 1998 / Accepted: 27 October 1998

Abstract *De novo* construction of complete genetic linkage maps requires large mapping populations, large numbers of genetic markers, and efficient algorithms for ordering markers and evaluating order confidence. We constructed a complete genetic map of an individual loblolly pine (*Pinus taeda* L.) using amplified fragment length polymorphism (AFLP) markers segregating in haploid megagametophytes and PGRI mapping software. We generated 521 polymorphic fragments from 21 AFLP primer pairs. A total of 508 fragments mapped to 12 linkage groups, which is equal to the *Pinus* haploid chromosome number. Bootstrap locus order matrices and recombination matrices generated by PGRI were used to select 184 framework markers that could be ordered confidently. Order support was also evaluated using log likelihood criteria in MAPMAKER. Optimal marker orders from PGRI and MAPMAKER were identical, but the implied reliability of orders differed greatly. The framework map provides nearly complete coverage of the genome, estimated at approximately 1700 cM in length using a modified estimator. This map should provide a useful framework for merging existing loblolly pine maps and adding multiallelic markers as they become available. Map coverage with dominant markers in both linkage phases will make the map useful for subsequent quantitative trait locus mapping in families derived by self-pollination.

Key words *Pinus taeda* · Linkage map · AFLP · Locus ordering · Genome length estimation

Communicated by G. H. Hart

D. L. Remington (✉) · R. W. Whetten · B.-H. Liu
D. M. O'Malley
Forest Biotechnology Group, Department of Forestry,
North Carolina State University, Raleigh, NC 27695-8008, USA
Fax: (919) 515-7801
E-mail: dlreming@unity.ncsu.edu

Introduction

Genetic maps with high levels of genome coverage and confidence in locus order are necessary for the reliable detection, mapping, and estimation of gene effects on phenotypic traits. The ability to order markers depends upon observing one or more recombination events between a pair of loci in the mapping population (Thompson 1987), and reliable ordering will usually require a number of meioses that is many times the number of loci (Edwards 1991). Genotyping errors interfere with locus ordering by indicating an excess of apparent double recombination events and may generate statistically significant support for incorrect locus order (Buetow 1991; Ehm et al. 1996). Incorrect locus orders and genotyping errors can also severely inflate map length estimates (Collins et al. 1996; Shields et al. 1991). Very large mapping populations are needed to order closely spaced markers with a high confidence level. Finding the most likely locus order may become computationally intractable because the number of possible locus orders increases multiplicatively with the number of available markers (Falk 1992). A point of diminishing returns occurs at which further resolution in genetic maps is not feasible and other approaches such as breakpoint analysis become necessary (Elsner et al. 1995). Choosing a subset of available markers that can be ordered reliably is an important but nontrivial task. For example, only 970 loci in a 5840-locus human microsatellite map could be ordered uniquely at specified support levels, given the available number of informative meioses (Murray et al. 1994).

A distinction has been made between "framework" maps consisting of only those markers whose order meets statistical support criteria, and "comprehensive" maps that attempt to place all markers in the most likely order (Keats et al. 1991). The predominant method for evaluating order support is a comparison of log likelihoods of alternate locus orders. However, the

likelihood ratio for alternate orders lacks a clear statistical interpretation, and compares the chosen order against only one alternate at a time (Buetow 1991; Keats et al. 1991). Bootstrap resampling provides another, more conservative method for evaluating confidence in locus orders (Liu 1998; Marques et al. 1997, 1998). Matrices of bootstrap location frequency for each locus provide a visually powerful evaluation of assigned locus position confidence. Sets of markers with strong order support will map to the same position in a high percentage of bootstrap replicates, which will lie in a single diagonal in the matrix. Consequently, the optimal locus order is immediately apparent from the bootstrap matrix. Error-prone markers will tend to be placed in widely varying positions in different replicates. The percentage of replicates in which a marker maps to the same position provides an empirical confidence level for marker position (Weir 1996).

Genetic mapping in pines (*Pinus* spp.) is still at an early stage, and the development of markers, mapping populations, and genetic maps generally have been done concurrently. Genetic maps have been constructed for several species of pines using restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), microsatellite, protein and, recently, amplified fragment length polymorphism (AFLP) markers (Devey et al. 1994, 1996; Echt and Nelson 1997; Kubisiak et al. 1995; Nelson et al. 1993, 1994; Plomion et al. 1995a, b; Travis et al. 1998). All of the maps constructed so far have contained more than the 12 linkage groups expected for the chromosome number in *Pinus*, except for that of Plomion et al. (1995a). Pine genetic maps constructed to date are generally reported to be incomplete, but this assessment is based on widely varying estimates of genome length. The advent of anonymous polymerase chain reaction (PCR)-based marker techniques such as AFLP (Vos et al. 1995) has made rapid *de novo* generation of large numbers of genetic markers feasible. This allows the construction of much more complete genetic maps from individual trees than has been practical until now. Locus-ordering algorithms such as simulated annealing (Kirkpatrick et al. 1983) and bootstrap methods of order evaluation improve efficiency when ordering the large numbers of markers generated by these techniques.

The Pinaceae have very large genomes, approximately 2×10^{10} bp. Consequently, individual pine chromosomes have about 57 times the physical length as those of *Arabidopsis*, even though the average map lengths are similar (Plomion et al. 1995a). The large genome size and predominance of repetitive DNA in the Pinaceae make the use of RFLP- and microsatellite-based genetic markers more difficult (Kinlaw and Neale 1997; Pfeiffer et al. 1997), a situation that we have found to be true for AFLP markers as well.

In this paper, we report construction of a genetic linkage map with complete coverage and 12 linkage groups (corresponding to the haploid chromosome number) in loblolly pine (*Pinus taeda* L.) from a single parent using AFLP markers. We discuss a novel approach to developing a framework linkage map from a large set of genetic markers, using PGRI software (Liu 1998). PGRI uses a simulated-annealing algorithm to order entire sets of linked markers and bootstrap resampling to evaluate locus order confidence levels. This facilitates framework map construction by permitting an efficient interactive process of identifying and dropping markers likely to contain scoring errors and evaluating the reliability of the resulting orders. We also describe successful methods for adapting the AFLP technique to mapping in physically large genomes, using automated fluorescence-based detection. We demonstrate complete map coverage using several approaches and consequently provide a firm genome length estimate of approximately 1700 cM Kosambi. Finally, we discuss the implications of the resulting map for development of consensus maps and trait mapping in families derived from self-pollination.

Materials and methods

DNA preparation

Megagametophytes were obtained from open-pollinated seeds from loblolly pine clone 7-56 (NCSU-Industry Cooperative Tree Improvement Program). Seeds were germinated in 1% hydrogen peroxide for approximately 4 days. Genomic DNA was extracted from ground, frozen megagametophytes by incubating these for approximately 1 h in 400 µl Puregene SDS-TRIS-EDTA cell lysis solution (Gentra Systems) containing 100 µg/ml Proteinase K and 20 µg/ml RNAse A, followed by the addition of 125 µl Puregene ammonium acetate protein precipitation solution (Gentra Systems). The DNA was precipitated from the supernatant by adding an equal volume of isopropanol, rinsed in 70% ethanol, and resuspended in 50 µl TE buffer. The DNA preparations were quantitated by electrophoresing of 2 µl of each suspension on 0.8% agarose gels containing 0.2 µg/ml ethidium bromide and then comparing band intensities with known quantities of lambda phage DNA.

AFLP template preparation and reactions

Templates for AFLP reactions were prepared following Vos et al. (1995) using 500 ng megagametophyte DNA for restriction digests with *Eco*RI and *Mse*I and ligation of adapters. The restriction-ligation (RL) mixture was diluted 1 : 10 in deionized water prior to preamplification.

Preamplification was carried out using standard AFLP *Eco*RI (E) and *Mse*I (M) primers (Vos et al. 1995) containing selective nucleotides E + AC and M + CC. Reaction mixture volumes were 20 µl, with 5 µl diluted RL mixture as template, 1.2 U *Taq* polymerase (Boehringer), 30 ng E primer, 30 ng *M* primer, 10 mM TRIS-HCl pH 8.3, 1.5 m*M* $MgCl_2$, 50 m*M* KCl, and 0.2 m*M* each of all four dNTPs. PCR amplifications were carried out with 28 cycles of a 30-s denaturation at 94°C, a 30-s annealing at 60°C, and a 60-s extension at 72°C.

Selective amplifications were done using various combinations of E primers with three selective nucleotides and M primers with four selective nucleotides (E + 3/M + 4). Reaction mixtures were as

described above for preamplification, except that 5 µl of 1:100 dilutions of the preamplification products was used as template, and only 5 ng of infrared dye (IRD)-labeled E primer (Li-Cor) was used. PCR amplifications consisted of 36 cycles of a 30-s denaturation at 94°C, a 30-s annealing (see below), and a 60-s extension at 72°C. The annealing temperature was 65°C for the first cycle, was reduced by 0.7°C for each of the next 12 cycles, and was 56°C for the remaining 23 cycles.

Detection and scoring of AFLP fragments

AFLP reaction products were resolved on denaturing gels containing 6% or 7% Long Ranger polyacrylamide (FMC), 7.5 *M* urea, and 1 × TBE (89 m*M* TRIS, 89 m*M* boric acid, 2 m*M* EDTA). Loading buffer (10 µl) consisting of 95% deionized formamide, 20 m*M* EDTA pH 8.0, and 1 mg/ml bromophenol blue (USB) was added to each selective amplification product prior to gel loading. This mixture was heated at 94°C for 3 min, then quickly cooled on ice before loading 1.5 µl of each sample on the gel. IRD-labeled molecular-weight markers (Li-Cor) were loaded in two lanes as a standard.

Electrophoresis was carried out on Li-Cor 4000L automated sequencers using 1 × TBE running buffer, with run parameters of 2000 V, 35 mA, 70 W, signal channel 3, motor speed 3 or 4, 50°C plate temperature, and 16-bit pixel depth for collection of TIFF image files.

Polymorphic fragments were scored by eye in the TIFF image files using RFLPscan Version 3.0 (Scanalytics). Automatic detection thresholds were set at the maximum level to minimize the number of automatically scored fragments, and polymorphic fragments were scored electronically by the user. The software automatically assigned molecular weights to fragments, binned the corresponding fragments from different samples representing single polymorphisms, and generated reports of fragment presence/absence strings for each sample. These reports were converted into mapping software formats using a spreadsheet program.

Linkage map construction

Map construction using PGRI version 1.0 (Liu 1998) consisted of assigning polymorphisms to linkage groups, ordering markers, and choosing a set of framework markers that could be ordered confidently. Linkage between pairs of markers was evaluated with a likelihood ratio test. The threshold p-value (α) for linkage evaluation was chosen so that the likelihood of obtaining any false linkages would be less than a target level a. The appropriate α is based not only on the number of two-point tests (m), but also on the prior probability of linkage ($\approx 1/C$) and the power to detect true linkage ($1 - \beta$), where C is the haploid chromosome number and β is the probability of type-II error (Morton 1955; Ott 1991). We can estimate m in terms of the number of markers n and make an approximation for β in terms of C, with the threshold map distance d corresponding to detectable linkage, and the genome length L, and solve for α (see Appendix):

$$\alpha \approx \frac{4dC^2a}{n^2L(C-1)^2}.$$

For each declared linkage group, the "manually interactive" option of PGRI was used to order candidate markers and select a set of framework markers with strong order support. Preliminary marker orders were generated using simulated annealing (Kirkpatrick et al. 1983) with minimum sum of adjacent recombination fractions (SARF) as the criterion. In simulated annealing, an initial marker order is chosen randomly and the SARF (E_i) is calculated. Then, two randomly selected loci are permuted. If the new SARF (E_j) is smaller than E_i, the new order is selected. If $E_j > E_i$, the new order will be accepted with probability

$$\frac{1}{k_bT}\exp(E_i - E_j),$$

where k_b is the Boltzman constant, and T is typically chosen to be greater than the largest likely values of $E_i - E_j$. This process is repeated iteratively with gradual reduction in the value of T, until a lower value of E is not obtained in a specified number of iterations. This algorithm allows orders with longer SARFs to be chosen occasionally, which reduces the likelihood of reaching a local rather than global minimum SARF (Liu 1998). The ordered markers were then evaluated for consistency of bootstrap placement and excesses of apparent double recombination events with adjacent markers. Less reliable markers were dropped in an iterative process, and the ordering repeated, until a reliable set of framework markers was obtained.

Linkage grouping and marker ordering were also evaluated in MAPMAKER version 2.0 for Macintosh (Lander et al. 1987). The entire marker data set was duplicated, and marker presence/absence scores were recoded in the duplicate set to allow MAPMAKER to detect repulsion-phase linkages. Marker distribution by linkage phase was evaluated in MAPMAKER, and alternate markers that improved linkage phase distribution were identified. All three-locus permutations of marker order within each linkage group were compared in MAPMAKER using the "ripple" command to evaluate LOD support for order.

Marker distribution

Marker distribution among linkage groups was evaluated by comparing marker density with expectations under the Poisson distribution. This test was conducted using all markers, both framewok and accessory. Each linkage group i was estimated to have a length $G_i = M_i + 2s$, where M_i is the map distance between terminal markers of linkage group i, and s is the average framework marker spacing. Under a uniform probability distribution for marker location, s is also the expected distance from a terminal marker to the chromosome end. If the underlying marker density were the same for all chromosomes, the number of markers m_i in linkage group i would be a sample from a Poisson distribution with parameter $\lambda_i = mG_i/\Sigma_i G_i$, where m is the total number of markers. The probabilities $P(X \leq m_i)$ and $P(X \geq m_i)$ were evaluated under the cumulative Poisson distribution. As this is a two-tailed test, probabilities less than $\alpha/2$ correspond to deviations from Poisson expectations of level α. Clustering of markers within linkage groups was tested by grouping each non-framework (accessory) marker with the closest framework marker. The number of accessory markers b_{ij} grouped with framework marker j in linkage group i was compared with Poisson expectations for a window of width W_{ij} cM. W_{ij} is half the combined distance to the adjacent framework markers, and for terminal framework markers it includes the expected distance of 8.9 cM to the chromosome end. If accessory markers are randomly distributed, the expected number λ_{ij} in a given window is equal to b_iW_{ij}/G_i, where b_i is the number of accessory markers in linkage group i, and the distribution of b_{ij} should be Poisson. The probabilities $P(X \leq b_{ij})$ and $P(X \geq b_{ij})$ were evaluated for each framework marker window under the cumulative Poisson distribution. Clustering of accessory markers can occur due to the procedure for selecting framework markers as well as inherent clustering of markers. Consequently, the number of b_{ij} values that deviate significantly from expectations may overestimate the degree of clustering.

Map length and genome coverage

Average framework marker spacing s was calculated by dividing the summed length of all linkage groups by the number of framework

marker intervals, which is the number of framework markers minus the number of linkage groups. The proportion c of the genome within d cM of a marker, assuming random marker distribution, was estimated using the relationship

$$c = 1 - e^{-2dn/L},$$

where L is the estimated genome length and n is the number of markers (Lange and Boehnke 1982). As a further check on genome coverage, all unlinked polymorphisms segregating in a 1:1 ratio were evaluated in MAPMAKER for linkage to each other and to the terminal framework markers of all linkage groups, using a low LOD threshold. Genome length L was estimated using the method of Hulbert et al. (1988), as modified in method 3 of Chakravarti et al. (1991), in which $\hat{L} = n(n-1)d/k$, where n is the total number of markers, d is the map distance corresponding to the LOD threshold Z for declaring linkage, and k is the number of markers linked at LOD Z or greater. We also used a modified estimator $\hat{L}_a$ that corrects the Hulbert estimate for an upward bias related to chromosome ends (see Appendix):

$$\hat{L}_a = \frac{n(n-1)d}{2k}\left(1 + \left[1 - \frac{2Ck}{n(n-1)}\right]^{1/2}\right).$$

Results

Generation and inheritance of AFLP polymorphisms

The genomes of conifers are very large (approx. 2×10^{10} bp). Consequently, the usual AFLP selective amplifications using E + 3/M + 3 primer combinations resulted in too many faint and overlapping fragments (results not shown). To address this problem, we added a fourth selective nucleotide to the M primer and did preamplifications with E + 2/M + 2 primer combinations in place of the typical E + 1/M + 1 combinations. The modified preamplification is important because some primer-template mismatch appears to be tolerated at sites other than the two bases at the 3′ end of the primer (Vos et al. 1995).

The base composition of the primer selective extensions also had a significant effect on the number of segregating AFLP fragments (Table 1). In particular, CpG dinucleotides in either the E or M primer selective extension substantially reduced the number of fragments detected and gave the most suitable results in most cases. However, CpG dinucleotides in the selective regions of both primers tended to result in too few fragments. This effect was not surprising, as CpG is known to be under-represented in vertebrate genomes (Cooper and Krawczak 1990).

Infrared dye-labeled E primers were substituted for the conventional 5′ end labeling with [^{33}P] for detection with the Li-Cor automated sequencer system. Overall sensitivity of band detection using the autoradiogram-like TIFF images appeared equal to or better than that obtained with autoradiography (Fig. 1).

We screened 36 primer combinations compatible with the E + AC/M + CC preamplification by doing selective amplifications from six samples. Most of the screened primer combinations contained at least one CpG dinucleotide in the selective extension. Each primer combination was scored for number of segregating polymorphic fragments detected and overall sharpness and intensity of polymorphic fragments. Based on this screening, we selected 21 primer combinations for use in mapping. AFLP reactions were carried out on DNA samples from 93 megagametophytes from open-pollinated seeds of loblolly pine clone 7-56. Diploid DNA samples from clone 7-56, an unrelated individual (7-51), and a progeny of these selections (7-1037) were also included to verify normal inheritance of fragments in megagametophytes from 7-56, and to identify which fragments were transmitted to 7-1037.

Table 1 Number of scored AFLP fragments by primer combination

EcoRI primer[a]	*Mse*I primer[a]	Number of CpG in selective extensions	Number of scored fragments
ACA	CCAG	0	38
	CCCG	1	13
	CCGC	1	21
	CCGG	1	27
	CCTG	0	43
ACC	CCAG	0	47
ACG	CCAA	1	47
	CCAC	1	20
	CCAG	1	20
	CCCA	1	25
	CCGA	2	15
	CCGC	2	10
	CCTA	1	28
	CCTC	1	32
	CCTG	1	14
	CCTT	1	20
ACT	CCAG	0	32
	CCCG	1	19
	CCGC	1	12
	CCGG	1	19
	CCTG	0	19
	Total:		521

Summary:

Number of CpG	Mean number of scored fragments	Number of primer combinations
0	35.80	5
1	22.64	14
2	12.50	2

Regression of scored fragment number on number of CpG:

Slope:	−12.11
Intercept:	35.19
Adjusted R^2:	0.348
F:	11.692
p-value:	0.0029

[a] Selective extensions only. See Vos et al. (1995) for core primer sequences

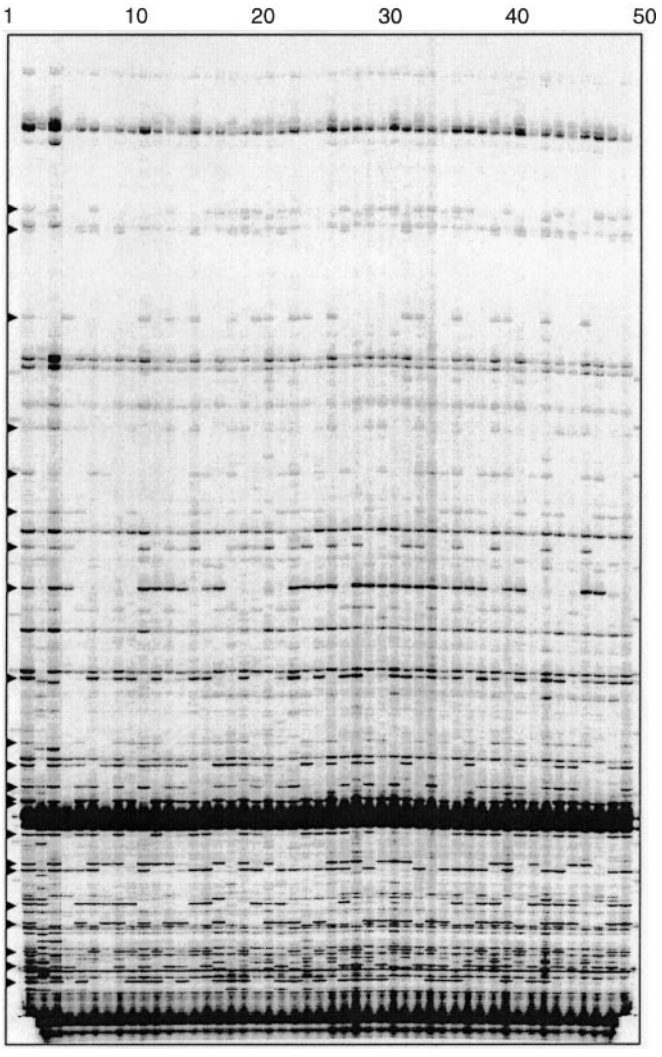

Fig. 1 A portion of a TIFF image for AFLP primer combination E + ACA/M + CCGG. *Lanes 2–4* contain diploid DNA from clones 7-56, 7-51, and 7-1037, respectively. *Lanes 5-49* contain haploid megagametophyte DNA from 45 seeds collected from clone 7-56. *Lanes 1* and 50 contain molecular-weight markers with a range of 50–350 bases. Fragments present in some samples and absent in others (arrowheads) were scored as polymorphisms

A total of 521 polymorphisms were scored from AFLP reactions using the 21 selected primer combinations. Preamplifications using the primer combination E + AC/M + CC were used as template for all selective amplifications. On average, 25 polymorphisms were scored per primer pair, with a range of 10–47 scored polymorphisms (Table 1). The TIFF images produced by the automated sequencer provided sufficient resolution to distinguish fragment mobilities at single-base resolution over the entire fragment size range (42–600 + nucleotides), although polymorphisms were difficult to score in regions in which 3 or more polymorphic fragments were separated in size by a single base each.

Repeatability of fragment scoring was evaluated by scoring 48 of the megagametophyte samples independently, on two separate occasions, from 2 separate selective amplifications with a representative primer pair (E + ACG/M + CCTG). The observed proportion of scoring discrepancies (w) was 0.021, which corresponds to an error rate ε of 1.1% using the relationship $w = 2\varepsilon(1 - \varepsilon)$ (Shields et al. 1991).

Linkage map construction

An initial p-value (α_{2p}) of 1×10^{-8} was chosen for declaring two-point linkages so as to achieve a likelihood of less than 5% of obtaining any false linkages. Using an initial estimate of 2000 cM for L and 32 cM for d (corresponding to a recombination fraction of approximately 0.28), $n = 521$ polymorphisms, $C = 12$, and a target a of 0.05, we obtained a value of 1.41×10^{-8} for α. However, the smallest p-value treated as nonzero in PGRI was 5.97×10^{-8}, so this value was used for initial linkage grouping. This p-value and a maximum recombination fraction r of 0.22 resulted in the grouping of 508 markers into 12 linkage groups (designated LG1-LG12), leaving 13 polymorphisms unlinked. We also grouped polymorphisms in MAPMAKER version 2.0 for Macintosh (generously provided by S. Tingey, DuPont) using a LOD threshold of 7.0, which corresponds to a p-value of 1.37×10^{-8}. This produced 13 rather than 12 linkage groups, with LG12 separated into 2 groups. The two sets of markers comprising LG12 could be joined at a p-value of 2.47×10^{-8} (LOD 6.74).

Polymorphic fragments inherited in 1 : 1 ratios from the maternal parent (7-56) that could be mapped to a linkage group were considered candidate genetic markers. Fragments that deviated from a 1 : 1 segregation at probability levels between 0.01 and 0.05 were not automatically dropped, as some deviations at this level are expected to occur by chance alone in a large data set. A band amplified from 7-56 genomic DNA corresponded with nearly every candidate marker. The few exceptions could be attributed to weak or failed 7-56 amplifications that prevented the scoring of some fragments. Final acceptance as useful markers also required that fragments could be scored reliably, which was evaluated during the subsequent ordering process.

Framework maps were constructed for each linkage group. To simplify ordering, we initially used marker subgroups generated by restricting the recombination fraction r to a maximum of 0.15. These were numbered 1a-18a, 21a-24a, and 28a-30a. Preliminary marker orders were generated using the simulated annealing/sum of adjacent recombination fractions (SA-SAR) algorithm, and the program produced a recombination matrix of the ordered loci, a map table, and a bootstrap confidence matrix for locus order. We first checked the

Genome position	Individual locus map position														
	CAG 86	AGG 124	GTG 163	TGC 134	ATG 486	CAG 489	GGC 178	AAG 196	GTA 151	TGG 221	TGG 227	CAG 92	AGG 434	GAG 152	GGC 262
1	89	0	2	0	1	3	0	5	0	0	0	0	0	0	0
2	2	89	0	1	3	0	5	0	0	0	0	0	0	0	0
3	0	2	87	6	0	5	0	0	0	0	0	0	0	0	0
4	0	1	6	88	5	0	0	0	0	0	0	0	0	0	0
5	1	3	1	4	91	0	0	0	0	0	0	0	0	0	0
6	3	1	3	1	0	91	0	0	0	0	0	0	0	1	0
7	1	4	0	0	0	0	88	3	0	0	0	0	1	1	2
8	4	0	1	0	0	0	3	88	0	0	0	0	1	1	2
9	0	0	0	0	0	0	0	0	89	2	0	1	2	4	2
10	0	0	0	0	0	0	0	0	2	86	1	6	0	2	3
11	0	0	0	0	0	0	0	0	0	3	85	7	5	0	0
12	0	0	0	0	0	0	0	0	1	4	9	86	0	0	0
13	0	0	0	0	0	0	0	1	3	0	5	0	91	0	0
14	0	0	0	0	0	0	1	3	0	5	0	0	0	83	8
15	0	0	0	0	0	1	3	0	5	0	0	0	0	8	83

Fig. 2 Matrix of 100 bootstrap replicates for marker position of LG4 framework markers, as generated by PGRI. Matrix values are the percentage of replicates in which each marker fell in the indicated position. Values on the *diagonal* represent the percentage confidence for correct locus position. *Off-diagonal values* are the frequency with which loci were placed in different positions due to sampling error (Liu 1998)

bootstrap matrix to ensure that the order generated was reasonable, as evidenced by a plurality of bootstrap orders for each locus falling close to a single diagonal (Fig. 2), and generated a new order if necessary. Errors in scoring generally show up as an excess of apparent double crossovers. These are easily detected in the recombination matrix because the sum of recombination fractions to nearby flanking pairs of markers will substantially exceed the recombination fraction between the flanking markers (Fig. 3). Error-prone markers also tended to be placed in widely varying locations in different bootstrap replicates, especially at the linkage group ends, in the bootstrap matrix. Markers initially ordered at the ends of linkage groups were closely scrutinized, and those with lower recombination fractions to interior markers were dropped. Polymorphisms were dropped a few at a time, a new order was generated, and the process was repeated. If the recombination matrix properties were not improved in the vicinity of the dropped markers, or if dropping the markers did not substantially shorten the map, they were added back in and other markers were dropped. In the final iterations, additional markers were dropped where spacing was too close to obtain reliable ordering. This iterative process was continued until all remaining loci were consistently placed at

Fig. 3 Recombination matrix from LG4, as generated by PGRI. *Boldface numbers* show recombination fractions between marker ACA/CCAG-710 and other markers. The sum of the recombination fractions between ACA/CCAG-710 and adjacent markers ACT/CCGC-134 and ACA/CCTG-486 (shown in *bold italics*) is substantially greater than the recombination fractions between the two adjacent markers (*underlined*). Dropping marker ACA/CCAG-710 reduces the length of LG4 by 6.8 cM Kosambi

Locus	CAG 86	AGG 124	GTG 163	TGC 134	**AAG 710**	ATG 486	CAG 489	GGC 178	AAG 196	GTA 151	TGG 221	TGG 227	CAG 92	AGG 434	GAG 512
AGG124	0.05														
GTG163	0.11	0.05													
TGC134	0.16	0.10	0.07												
AAG710	**0.22**	**0.18**	**0.13**	***0.08***											
ATG486	0.24	0.20	0.15	0.09	***0.12***										
CAG489	0.29	0.28	0.23	0.18	**0.18**	0.11									
GGC178	0.59	0.58	0.63	0.33	**0.32**	0.24	0.16								
AAG196	0.55	0.54	0.57	0.38	**0.62**	0.28	0.22	0.10							
GTA151	0.52	0.51	0.48	0.57	**0.46**	0.41	0.62	0.25	0.21						
TGG221	0.52	0.51	0.47	0.57	**0.48**	0.42	0.58	0.29	0.24	0.05					
TGG227	0.52	0.51	0.47	0.57	**0.48**	0.42	0.58	0.31	0.29	0.10	0.04				
CAG92	0.49	0.50	0.53	0.43	**0.53**	0.59	0.41	0.32	0.28	0.10	0.07	0.02			
AGG434	0.51	0.52	0.53	0.47	**0.49**	0.50	0.50	0.47	0.43	0.25	0.21	0.17	0.15		
GAG152	0.48	0.49	0.48	0.52	**0.48**	0.52	0.48	0.47	0.53	0.33	0.29	0.24	0.23	0.12	
GGC262	0.51	0.50	0.49	0.48	**0.49**	0.45	0.54	0.55	0.49	0.34	0.32	0.28	0.27	0.14	0.04

a single position in at least 70% of the bootstrap replicates. By this point in the process, bootstrap support for most loci was typically about 90%. After this process was completed for all linkage subgroups, we recombined the retained markers from each subgroup into the initial 12 linkage groups. Additional markers were dropped as needed until bootstrap placement was again greater than 70% for all positions.

The final map (Fig. 4) contained 12 linkage groups, as did the initial grouping in PGRI, but the initial grouping was not entirely correct. The 3 subgroups (2a, 8a and 10a) comprising LG2 did not behave as a single linkage group when the subgroups were combined. The loci could not be ordered so that a bootstrap matrix with a single prominent diagonal was generated. By dropping 1 entire subgroup at a time, we found that subgroups 8a and 10a behaved as a single group when subgroup 2a was left out. We also found that subgroup 2a and marker ACA/CCTG-380 (which grouped with 8a but could not be ordered with the other markers) were linked to LG1 at a p-value of approximately 5×10^{-7}. A bootstrap matrix with a single prominent diagonal was generated when this group of markers was combined with LG1, which then increased in length from 80.4 cM to 137.1 cM. We concluded that subgroup 2a and marker ACA/CCTG-380 belong to LG1.

In the case of LG12, which was split into 2 linkage groups at LOD 7.0 using MAPMAKER, the combined subgroups behaved as a single linkage group in the bootstrap process. Consequently, we accepted the treatment of LG12 as a single linkage group, as suggested by the slightly less restrictive criteria used in PGRI.

Variations in locus orders between bootstraps can result from failure of the simulated annealing algorithm to generate the optimal order as well as from actual changes in the optimal order due to resampling. To evaluate the effect of non-optimal initial orders on the bootstrap confidence level, we replicated the generation of initial orders for the framework markers of LG1 and LG4 without resampling. No changes in locus order were found in 80 replications with LG1, but 9 out of 100 replications with LG4 generated different orders.

Order support of the map was also evaluated in MAPMAKER V.2.0 for Macintosh using the "Ripple" command to compare all three-locus permutations of the framework order. A few additional markers were dropped from the framework map or substituted with other markers in situations where the log likelihood order support was less than 3.0. We did retain some marker combinations with order support less than 3.0 where they contributed to the distribution of marker linkage phases on the framework map. The weakest order support by this criterion is a log likelihood difference of 1.68 associated with permuting markers ACG/CCAG-152 and ACG/CCGC-262 at the tip of LG4. The optimal locus orders indicated by the Ripple procedure in MAPMAKER agreed in every case with the locus orders determined by the bootstrap procedure in PGRI.

We also used MAPMAKER to evaluate the overall distribution of framework markers by linkage phase. To ensure thorough map coverage with both linkage phases, we also identified on the map markers perfectly linked in repulsion to framework markers. In most cases these do not appear to be true codominant markers, as they were obtained using different primer pairs. In long regions with only a single linkage phase represented on the framework map, we also located additional repulsion-phase markers closely linked to framework markers ($r < 0.04$) on the map relative to the nearest framework marker. All other nonframework (accessory) markers are not shown on the map but are located in a reference spreadsheet (available from the corresponding author upon request) with respect to the nearest framework marker.

Map length and coverage

The final linkage map (Fig. 4) consists of 184 framework markers. Eight additional markers perfectly linked in repulsion to the framework markers and 11 other alternate markers for improved linkage phase distribution are also located on the map. The combined length of the 12 linkage groups is 1528 cM Kosambi. The average framework marker spacing, calculated by dividing the summed length of the linkage groups by the number of framework marker intervals, is 8.9 cM. If framework markers are not clustered and each linkage group corresponds to a single chromosome, then the estimated average distance between the terminal markers of each linkage group and the actual chromosome ends is equal to the 8.9 cM average framework marker spacing. With these assumptions, the estimated map length is 1742 cM.

Tests for marker distribution among linkage groups compared the total number of markers m_i for each linkage group with its expected value $\lambda_i = 508G_i/1742$. Poisson probabilities for deviations of m_i from λ_i in either direction were greater than 0.025 for all linkage groups (Table 2). Thus, we did not detect significant differences in marker density among linkage groups at a 0.05 level. When we tested for clustering of accessory markers within linkage groups, 18 out of 184 intervals showed deviations from the Poisson expectation at the 0.05 level [i.e., $P(X \leq b_{ij}) < 0.025$ or $P(X \geq b_{ij}) < 0.025$], and 13 deviated at the 0.01 level. This suggested at least some degree of marker clustering when all markers (not just framework markers) are considered. In both of the tests for marker distribution, the unique value of the Poisson parameter for each linkage group and window precluded the use of a single test statistic to evaluate the extent of clustering.

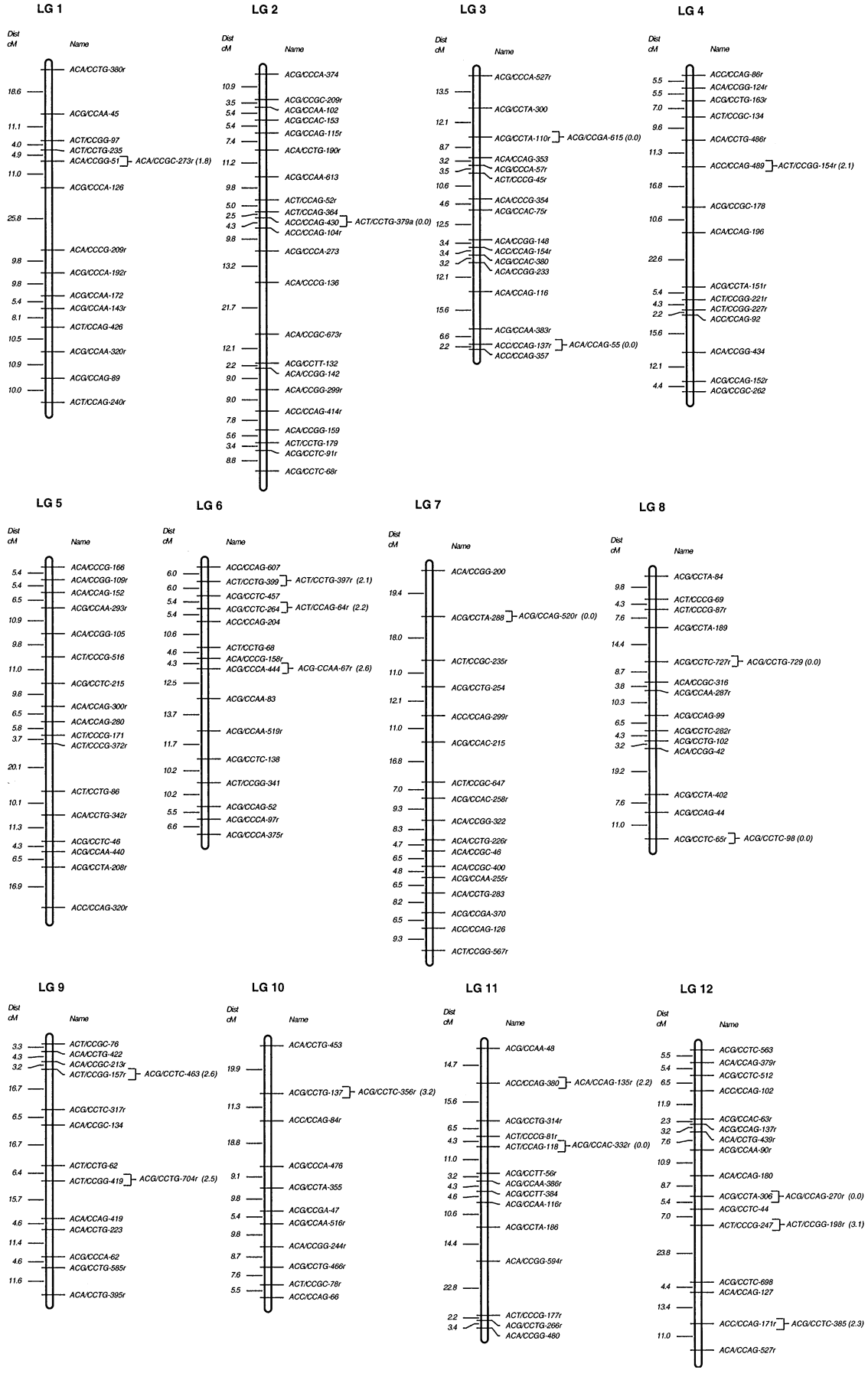
LG 1
Dist cM
Name
ACA/CCTG-380r
18.6
ACG/CCAA-45
11.1
ACT/CCGG-97
4.0
ACT/CCTG-235
4.9
ACA/CCGG-51
ACA/CCGC-273r (1.8)
11.0
ACG/CCCA-126
25.8
ACA/CCCG-209r
9.8
ACG/CCCA-192r
9.8
ACG/CCAA-172
5.4
ACG/CCAA-143r
8.1
ACT/CCAG-426
10.5
ACG/CCAA-320r
10.9
ACG/CCAG-89
10.0
ACT/CCAG-240r
LG 2
Dist cM
Name
ACG/CCCA-374
10.9
ACG/CCGC-209r
3.5
ACG/CCAA-102
5.4
ACG/CCAC-153
5.4
ACG/CCAG-115r
7.4
ACA/CCTG-190r
11.2
ACG/CCAA-613
9.8
ACT/CCAG-52r
5.0
ACT/CCAG-364
2.5
ACC/CCAG-430
ACT/CCTG-379a (0.0)
4.3
ACC/CCAG-104r
9.8
ACG/CCCA-273
13.2
ACA/CCCG-136
21.7
ACA/CCGC-673r
12.1
ACG/CCTT-132
2.2
ACA/CCGG-142
9.0
ACA/CCGG-299r
9.0
ACC/CCAG-414r
7.8
ACA/CCGG-159
5.6
ACT/CCTG-179
3.4
ACG/CCTC-91r
8.8
ACG/CCTC-68r
LG 3
Dist cM
Name
ACG/CCCA-527r
13.5
ACG/CCTA-300
12.1
ACG/CCTA-110r
ACG/CCGA-615 (0.0)
8.7
ACA/CCAG-353
3.2
ACG/CCCA-57r
3.5
ACT/CCCG-45r
10.6
ACA/CCCG-354
4.6
ACG/CCAC-75r
12.5
ACA/CCGG-148
3.4
ACC/CCAG-154r
3.4
ACG/CCAC-380
3.2
ACA/CCGG-233
12.1
ACA/CCAG-116
15.6
ACG/CCAA-383r
6.6
ACC/CCAG-137r
ACA/CCAG-55 (0.0)
2.2
ACC/CCAG-357
LG 4
Dist cM
Name
ACC/CCAG-86r
5.5
ACA/CCGG-124r
5.5
ACG/CCTG-163r
7.0
ACT/CCGC-134
9.6
ACA/CCTG-486r
11.3
ACC/CCAG-489
ACT/CCGG-154r (2.1)
16.8
ACG/CCGC-178
10.6
ACA/CCAG-196
22.6
ACG/CCTA-151r
5.4
ACT/CCGG-221r
4.3
ACT/CCGG-227r
2.2
ACC/CCAG-92
15.6
ACA/CCGG-434
12.1
ACG/CCAG-152r
4.4
ACG/CCGC-262
LG 5
Dist cM
Name
ACA/CCCG-166
5.4
ACA/CCGG-109r
5.4
ACA/CCAG-152
6.5
ACG/CCAA-293r
10.9
ACA/CCGG-105
9.8
ACT/CCCG-516
11.0
ACG/CCTC-215
9.8
ACA/CCAG-300r
6.5
ACA/CCAG-280
5.8
ACT/CCCG-171
3.7
ACT/CCCG-372r
20.1
ACT/CCTG-86
10.1
ACA/CCTG-342r
11.3
ACG/CCTC-46
4.3
ACG/CCAA-440
6.5
ACG/CCTA-208r
16.9
ACC/CCAG-320r
LG 6
Dist cM
Name
ACC/CCAG-607
6.0
ACT/CCTG-399
ACT/CCTG-397r (2.1)
6.0
ACG/CCTC-457
5.4
ACG/CCTC-264
ACT/CCAG-64r (2.2)
5.4
ACC/CCAG-204
10.6
ACT/CCTG-68
4.6
ACA/CCCG-158r
4.3
ACG/CCCA-444
ACG-CCAA-67r (2.6)
12.5
ACG/CCAA-83
13.7
ACG/CCAA-519r
11.7
ACG/CCTC-138
10.2
ACT/CCGG-341
10.2
ACG/CCAG-52
5.5
ACG/CCCA-97r
6.6
ACG/CCCA-375r
LG 7
Dist cM
Name
ACA/CCGG-200
19.4
ACG/CCTA-288
ACG/CCAG-520r (0.0)
18.0
ACT/CCGC-235r
11.0
ACG/CCTG-254
12.1
ACC/CCAG-299r
11.0
ACG/CCAC-215
16.8
ACT/CCGC-647
7.0
ACG/CCAC-258r
9.3
ACA/CCGG-322
8.3
ACA/CCTG-226r
4.7
ACA/CCGC-46
6.5
ACA/CCGC-400
4.8
ACG/CCAA-255r
6.5
ACA/CCTG-283
8.2
ACG/CCGA-370
6.5
ACC/CCAG-126
9.3
ACT/CCGG-567r
LG 8
Dist cM
Name
ACG/CCTA-84
9.8
ACT/CCCG-69
4.3
ACT/CCCG-87r
7.6
ACG/CCTA-189
14.4
ACG/CCTC-727r
ACG/CCTG-729 (0.0)
8.7
ACA/CCGC-316
3.8
ACG/CCAA-287r
10.3
ACG/CCAG-99
6.5
ACG/CCTC-282r
4.3
ACG/CCTG-102
3.2
ACA/CCGG-42
19.2
ACG/CCTA-402
7.6
ACG/CCAG-44
11.0
ACG/CCTC-65r
ACG/CCTC-98 (0.0)
LG 9
Dist cM
Name
ACT/CCGC-76
3.3
ACA/CCTG-422
4.3
ACA/CCGC-213r
3.2
ACT/CCGG-157r
ACG/CCTC-463 (2.6)
16.7
ACG/CCTC-317r
6.5
ACA/CCGC-134
16.7
ACT/CCTG-62
6.4
ACT/CCGG-419
ACG/CCTG-704r (2.5)
15.7
ACA/CCAG-419
4.6
ACA/CCTG-223
11.4
ACG/CCCA-62
4.6
ACG/CCTG-585r
11.6
ACA/CCTG-395r
LG 10
Dist cM
Name
ACA/CCTG-453
19.9
ACG/CCTG-137
ACG/CCTC-356r (3.2)
11.3
ACC/CCAG-84r
18.8
ACG/CCCA-476
9.1
ACG/CCTA-355
9.8
ACG/CCGA-47
5.4
ACG/CCAA-516r
9.8
ACA/CCGG-244r
8.7
ACG/CCTG-466r
7.6
ACT/CCGC-78r
5.5
ACC/CCAG-66
LG 11
Dist cM
Name
ACG/CCAA-48
14.7
ACC/CCAG-380
ACA/CCAG-135r (2.2)
15.6
ACG/CCTG-314r
6.5
ACT/CCCG-81r
4.3
ACT/CCAG-118
ACG/CCAC-332r (0.0)
11.0
ACG/CCTT-56r
3.2
ACG/CCAA-386r
4.3
ACG/CCTT-384
4.6
ACG/CCAA-116r
10.6
ACG/CCTA-186
14.4
ACA/CCGG-594r
22.8
ACT/CCCG-177r
2.2
ACG/CCTG-266r
3.4
ACA/CCGG-480
LG 12
Dist cM
Name
ACG/CCTC-563
5.5
ACA/CCAG-379r
5.4
ACG/CCTC-512
6.5
ACC/CCAG-102
11.9
ACG/CCAC-63r
2.3
ACG/CCAG-137r
3.2
ACA/CCTG-439r
7.6
ACG/CCAA-90r
10.9
ACA/CCAG-180
8.7
ACG/CCTA-306
ACG/CCAG-270r (0.0)
5.4
ACG/CCTC-44
7.0
ACT/CCCG-247
ACT/CCGG-198r (3.1)
23.8
ACG/CCTC-698
4.4
ACA/CCAG-127
13.4
ACC/CCAG-171r
ACG/CCTC-385 (2.3)
11.0
ACA/CCAG-527r

Table 2 Marker density by linkage group

Linkage group	Number of Markers (m_i)	Map length (cM)[a] (M_i)	Inferred LG length (cM)[a] (G_i)	Expected number of markers (λ_i)	Poisson two-tailed P-value[b]
1	45	137	154.8	45.15	0.530
2	53	162	179.8	52.45	0.488
3	43	114	131.8	38.44	0.252
4	45	131	148.8	43.40	0.424
5	38	144	161.8	47.19	0.100
6	52	116	133.8	39.03	0.027
7	44	160	177.8	51.86	0.153
8	39	108	125.8	36.69	0.373
9	36	104	121.8	35.53	0.491
10	26	106	123.8	36.11	0.049
11	51	117	134.8	39.32	0.042
12	36	129	146.8	42.82	0.167
Total:	508	1528	1741.6	508	

[a] Map lengths are in centiMorgans (cM), Kosambi function
[b] Poisson probability of having as many (for $m_i \geq \lambda_i$) or as few (for $m_i < \lambda_i$) markers as the observed number m_i in linkage group i, under the null hypothesis that the true marker density is the same for all linkage groups. As this is a two-tailed test, a p-value of 0.025 corresponds to a significance level of 0.05

We evaluated the degree of map coverage in several ways. Using the formula $c = 1 - e^{-2dn/L}$ (see Materials and methods) and estimating L at 1800 cM, an estimated 99.6% of the genome is within 10 cM of one of the 508 linked markers (Lange and Boehnke 1982). Using only the 183 framework markers, an estimated 87.5% of the genome is within 10 cM of a framework marker, and 98.4% is within 20 cM of a framework marker.

We estimated genome length using the Hulbert method (Chakravarti et al. 1991; Hulbert et al. 1988) with our modifications. A total of 3284 linked marker pairs were detected using a LOD threshold of 7.0 and letting n represent all 521 scored polymorphisms. The maximum map distance associated with the LOD score of 7.0 is approximately 22 cM, resulting in an unadjusted genome length estimate of 1814 cM Kosambi and an estimate of 1672 cM with the adjustment for chromosome ends. These estimates are both within 4.2% of the 1742 cM framework map length estimate.

Finally, we evaluated whether the 13 unlinked polymorphisms could be markers in genomic regions unsampled by the remaining markers. Six of these polymorphisms had segregation ratios highly distorted from the expected 1 : 1 ratio, with χ^2 test statistic values of 8.91 or greater, and were more suggestive of the 3 : 1 segregation ratio expected of a pair of unlinked comigrating fragments. Using MAPMAKER, we tested the 7 remaining unlinked polymorphisms for linkage to each other and to the 2 terminal framework markers of each linkage group at a permissive LOD threshold of 3.0. None of the 7 polymorphisms showed linkage to the terminal markers of any of the linkage groups. Three were loosely linked to each other, but support for the most likely order was very weak. We subsequently rechecked the RFLPscan images for these 7 polymorphisms. In one case, different fragments had been scored on different gels, and 5 of the other 6 polymorphisms were difficult to score confidently because of faint or variable-intensity bands and co-migrating fragments. We concluded that none of these unlinked polymorphisms were likely to be genuine markers outside of regions covered by the map.

Fig. 4 Final linkage map for *Pinus taeda* clone 7-56. Marker names ending with *r* are in reverse linkage phase to those not so designated. Alternate markers are placed to the *right* of the nearest framework marker, with the recombination fraction shown in *parentheses*

Population distribution of marker alleles

The diploid DNA from clone 7-51 included in the AFLP reactions was used to generate a preliminary estimate of the frequency at which 7-56 markers will also be segregating in an unrelated individual. Fragments corresponding to 171 polymorphisms segregating in 7-56 progeny were identified in 7-51, out of a total of 478 loci that could be confidently scored in 7-51. Some of these fragments will be homozygous in 7-51, and only the heterozygous fragments represent potential markers. If genotype frequencies are in Hardy-Weinberg proportions at each locus, the expected frequency of heterozygous fragments P_H is

$$P_H = 2[(P_A - V_p)^{1/2} - P_A],$$

where P_A is the observed frequency of band-absent phenotypes in a set of marker loci observed in diploid individuals, and V_p is the variance in band-present allele frequency among loci. V_p cannot be estimated from the data when only one diploid individual is observed, but

a reasonable range of values can be used in the equation. The estimate of P_A from the 7-51 data is $1 - 171/478 = 0.642$, and P_H estimates range from 0.318 with $V_p = 0$, to 0.290 with a standard deviation of 0.15 for band-present allele frequency ($V_p = 0.0225$), to 0.202 with a very large allele frequency standard deviation of 0.30 ($V_p = 0.09$). This also assumes that all corresponding fragments in 7-51 are actually homologous to the 7-56 fragments and that 7-51 is a typical individual. Consequently, these estimates are only preliminary and need to be verified by mapping other individuals using the same primer combinations.

Discussion

Map construction

Two persistent problems in genetic mapping have been the identification of optimal locus orders and the identification and correction of errors. Methods for identifying optimal locus orders without an exhaustive evaluation of every possible order include branch and bound (Thompson 1987), seriation (Buetow and Chakravarti 1987), and simulated annealing (Kirkpatrick et al. 1983). Only the branch and bound method is guaranteed to produce the best order, but an intractably large number of orders may need to be evaluated for large linkage groups (Liu 1998; Weir 1996). Several methods have been proposed to identify potential genotyping errors using likelihoods (Ehm et al. 1996; Lincoln and Lander 1992; Ott 1993). Newell et al. (1995) have proposed a distance geometry method that provides both a deterministic solution for optimal order and error estimates for placement of individual loci.

PGRI facilitates optimal locus ordering and the evaluation of order reliability by combining a simulated annealing algorithm with bootstrapping. The major advantage of bootstrapping is that the optimal order (if one clearly exists) is immediately apparent from the bootstrap matrix. Even though the simulated annealing algorithm frequently generated non-optimal orders, especially when ordering a large number of markers, the quality of the generated order could readily be evaluated from the bootstrap table and the markers could be reordered if necessary. As a result, locus ordering in PGRI was efficient even when large numbers of markers were being ordered at one time. The bootstrap matrix also allowed the immediate diagnosis and resolution of false linkage assignments, a situation that could be difficult to resolve by other methodologies. In contrast, the log likelihood comparisons from MAPMAKER offer a conventional algebraic measure of order support, around which standards for framework maps have been established (Keats et al. 1991). Our "framework" map does not strictly follow these standards, as we have included some locus combinations with interval support of less than 3 to improve coverage with both marker linkage phases. However, strict framework criteria could easily be met by dropping relatively few loci without affecting the overall integrity or genome coverage of the map.

The apparent optimal orders of framework markers were identical for all linkage groups in PGRI and MAPMAKER, although the implied reliability of orders is very different. Overall, a bootstrap support of 75–80% for a locus position tended to correspond to a log likelihood difference of about 3 for the favored order compared to the next most likely alternative. Log likelihood comparisons underestimate the error associated with locus orders as they only compare one alternative order at a time (Keats et al. 1991; Marques et al. 1998), and the likelihood ratio is not in itself a probability of type-I error. Plomion et al. (1995b) found that two independently constructed maps from the same individual contained order discrepancies in about 2% of the intervals when an interval support criterion of 3 was used. On the other hand, the bootstrap percentage for a given locus position is a conservative measure of reliability because order changes result from generation of non-optimal orders by the ordering algorithm (in this case simulated annealing) as well as from actual differences in optimal orders between bootstrap samples. Our replication of initial ordering for 2 linkage groups suggests that bootstrap confidence levels may underestimate the true confidence level for locus position by nearly 10% for some linkage groups.

We did not apply a systematic error detection algorithm, such as that of Lincoln and Lander (1992) in the current versions of MAPMAKER, to identify and correct individual scoring errors. We were more interested in identifying and dropping altogether loci with excessive scoring errors rather than correcting individual scores, and our approach of searching for excess double recombinants using the recombination matrix served this purpose. Our rationale was that error rates tend to reflect the difficulty of scoring particular markers, so markers scored with few errors are likely to be scored more accurately in future data sets as well. Alternative approaches to ordering loci and evaluating marker quality are needed in standard mapping software. The distance matrix approach of Newell et al. (1995) in particular may be worthy of further evaluation.

Map length and coverage

Several lines of evidence indicate virtually complete genome coverage for our map. These include coverage estimates of nearly 100% based on the number of markers; identification of 12 linkage groups, equal to the *Pinus* chromosome number; close agreement between map length and the Hulbert genome length estimator; and a lack of unlinked polymorphisms that are

credible markers. Our estimates of genome coverage based on number of markers predict that 98.4% of the loblolly pine genome is within 20 cM of a framework marker. Estimates based on the number of markers will underestimate coverage if markers are spaced systematically. As the process of selecting framework markers results in a somewhat systematic marker distribution, our estimate probably represents a lower bound of framework map coverage.

The coalescence of our AFLP linkage map into 12 strongly supported linkage groups contrasts with the 29 linkage groups obtained by Paglia et al. (1998) in Norway spruce ($n = 12$), and the 25 linkage groups of Travis et al. (1998) in pinyon pine. Paglia et al. constructed their linkage map from 366 AFLP fragments, 20 selective amplification of microsatellite polymorphic loci (SAMPL) fragments, and 61 microsatellites, and Travis et al. used 542 AFLP markers. However, both studies used smaller sample sizes (72 and 40 megagametophytes, respectively), which would give less power to detect statistically significant linkages. In addition, Paglia et al. used the methylation-sensitive *Pst*I in place of *Eco*RI in their AFLP restriction digests to reduce the number of bands obtained in the large spruce genome. Paglia et al. speculate that the resulting markers are concentrated in non-randomly distributed hypomethylated regions. *Eco*RI sites may be more randomly distributed over the genome, leaving fewer large gaps in map coverage and facilitating the coalescence of linkage groups. We did observe some clustering of markers within linkage groups. While the non-random process of selecting framework markers may have influenced our test for marker clustering, this is unlikely to explain the number of observations that deviated from expectations at the 1% level. Studies in *Drosophila* and mouse indicate that non-random variation in marker distribution on genetic maps is due mostly to heterogeneous recombination rates rather than differences in the physical distribution of markers (Lyon 1976; Nachman and Churchill 1996), and this may explain our results. We did not observe the extreme degree of centromeric clustering of AFLP markers reported by Young et al. (1998) in rainbow trout.

This study establishes a firm estimate of the genome length of pine. Our estimates of genome length, based both on map length and the adjusted Hulbert estimate, suggest a genome length of approximately 1700 cM Kosambi. Other published estimates in *Pinus* spp. range from about 1300 cM to more than 3000 cM (Echt and Nelson 1997; Kubisiak et al. 1995; Nelson et al. 1993, 1994; Plomion et al. 1995a, b; Travis et al. 1998). These discrepancies may be due in part to the choice of map function (Echt and Nelson 1997) and differences in recombination rates between pollen and seed parents (Groover et al. 1995; Plomion and O'Malley 1996). Echt and Nelson (1997) obtained estimates close to 2000 cM Kosambi for three species of *Pinus* by using a set of standardized criteria. Estimates based on chiasmata frequency suggest a genome length closer to 1500 cM (Plomion et al. 1995b; Saylor and Smith 1966).

Estimates of genome length and map distances between markers are important for the estimation of gene effects, integration of genetic and physical maps, and evaluation of map coverage. Consequently, it is important to minimize biases that can influence these estimates. Simulation studies show that the Hulbert estimator tends to overestimate genome length (Chakravarti et al. 1991). The upward bias may be due in part to ignoring the effect of chromosome ends. We have introduced an adjustment for this bias that does not require use of the more computationally intensive maximum likelihood estimator developed by Chakravarti et al. (1991). Using this adjustment shortened our genome length estimate by about 8%. Genotyping errors also cause substantial inflation of map length estimates (Buetow 1991; Shields et al. 1991), and they will inflate the Hulbert estimator as well. Genotyping errors are probably a factor in all estimates of pine genome length to date, especially when all of the scored markers are included in the data set. We attempted to minimize the contribution of scoring errors to the framework map by starting with a large initial number of markers and dropping markers that showed excessive double recombinations with flanking markers. Nevertheless, our framework map length may still be somewhat inflated by remaining errors. The non-random clustering of markers, on the other hand, may bias the genome length estimates downward.

Map utility

This map should be useful for merging linkage groups on existing loblolly pine maps and developing consensus maps by virtue of its complete coverage and correct number of linkage groups. The distribution of 7-56 polymorphic fragments in an unrelated individual (7-51) suggests that about a quarter of these markers are likely to be segregating in any given loblolly pine family. Nearly 90% of co-migrating AFLP polymorphisms scored in different potato genotypes appeared to be homologous, as evidenced by mapping to the same regions and sequence identity (Rouppe van der Voort et al. 1997). Identifying two or more homologous segregating markers per linkage group will establish map synteny and alignment between different individuals. This should be easily achievable given the large number of available markers. This map should also provide a useful framework for locating multiallelic markers such as microsatellites as they become available, as discussed by Paglia et al. (1998).

We plan to use this linkage map for mapping expressed sequence tags (ESTs), known genes, quantitative trait loci (QTLs), and viability loci in a family derived by self-pollination of clone 7-56. Dominant

markers have been shown to have low information content for mapping QTLs in F_2 or self families (Liu 1998), but this assumes that all markers are in a single linkage phase. Dominant and codominant markers are equally informative for linkage mapping in haploid genomes as we have done, or with a backcross or pseudo-testcross design (Grattapaglia and Sederoff 1994). In a simulation study, Jiang and Zeng (1997) estimated the informativeness of dominant markers for QTL mapping in F_2 populations, relative to codominant markers, using a Markov chain method to estimate conditional marker genotype probabilities. They found little loss of power or precision when dominant markers of both linkage phases were equally represented and the linkage map was already known. We have sought to maximize coverage with both marker linkage phases in constructing this map, so these circumstances will be largely satisfied in our subsequent QTL mapping.

Acknowledgements We express our gratitude to Ron Sederoff for his guidance and encouragement, to Glen Dale and Maria-Teresa Cervera for invaluable help with the AFLP technique and adapting it for pine, to Jeff Harford of Li-Cor for assistance in adapting the AFLP protocol for the Li-Cor automated sequencing system, and to Rongling Wu and Brian Palmer for assistance with DNA preparations. We thank two reviewers for helpful comments on an earlier draft of this manuscript. This work was supported by funding from the North Carolina State University Forest Biotechnology Industrial Associates Consortium, the National Institutes of Health (Grant GM45344-06) and by a USDA National Needs Graduate Fellowship to D.L.R.

All experiments comply with the current laws of the United States of America.

Appendix

Calculation of probability threshold for two-point linkage assignments

Let U be the event that a randomly chosen pair of loci is unlinked, i.e., that they reside on different chromosomes. Also, let T be the event that a test statistic exceeds the critical value for declaring linkage at level α. Finally, we define U' and T' as the complement of U and T, respectively.

The goal in defining the appropriate level α is to minimize to some acceptable level (for example, 0.05) the probability that any pair of unlinked loci in the data set will falsely be identified as linked, which would result in the merging of two chromosomes into a single linkage group. This would be the conditional event $U^*|T$, where $U^* = \cup_i(U_i|T)$ over all i pairs of loci. While it may seem intuitive to treat α as $P(U|T)$, α is instead correctly interpreted as $P(T|U)$, which is the probability that the test statistic exceeds the critical value for an unlinked pair of loci. By Bayes Theorem:

$$P(U|T) = \frac{P(T|U)\ P(U)}{P(T)} = \frac{P(T|U)\ P(U)}{P(T|U)\ P(U) + P(T|U')\ P(U')}, \quad (1)$$

as originally shown by Morton (1955; Ott 1991).

To estimate $\alpha = P(T|U)$, we need estimates of the other terms in Eq. 1. It is convenient to estimate $P(U|T)$ using the relationship $P(U^*|T) = 1 - [1 - P(U|T)]^m \approx mP(U|T)$, where m is the number of unlinked locus pairs in the set of marker loci. This estimate is conservative, as the m unlinked locus pairs are not all independent. If markers have an equal probability of being on any chromosome, the expected value for m is $n^2(C-1)/2C$, where n is the total number of marker loci and C is the haploid chromosome number. If a is the desired value for $P(U^*|T)$, then

$$P(U|T) \approx a/m = \frac{2aC}{n^2(C-1)}.$$

$P(U)$ is approximately $(C-1)/C$, and $P(U') \approx 1/C$, provided that markers have nearly equal probabilities of being located on any chromosome. $P(T|U')$ is the power to detect true linkage, or $1-\beta$, where β is the probability of a type-II error. If d is the threshold map distance corresponding to T, and L is the total genome length in map units, then $1-\beta \approx 2dC/L$.

Using these approximations, Eq. 1 becomes:

$$\frac{a}{m} \approx \frac{\alpha(C-1)/C}{\alpha(C-1)/C + (1-\beta)/C} = \frac{1}{1 + (1-\beta)/((C-1)\alpha)}.$$

Substituting $2dC/L$ for $1-\beta$ and solving for α:

$$\alpha \approx \frac{2dCa}{(m-a)L(C-1)} \approx \frac{2dCa}{mL(C-1)} \approx \frac{4dC^2a}{n^2L(C-1)^2}. \quad (2)$$

L will usually be unknown at this stage and must be estimated. Also, d will be dependent on the value of α, which is being solved for, so an approximate value must be chosen. If desired, a new d can be chosen based on the calculated value of α, and the calculation repeated iteratively until the values for α converge. However, this is probably not warranted in most cases given the approximations involved in estimating $1-\beta$.

Adjustment of genome length estimate

In the method of Hulbert et al. (1988), as modified in method 3 of Chakravarti et al. (1991), genome length L is estimated by the formula,

$$\hat{L} = \frac{n(n-1)2d}{2k} = \frac{n(n-1)d}{k},$$

where n is the total number of markers, d is the map distance corresponding to the LOD threshold Z for declaring linkage, and k is the number of marker pairs linked at a LOD Z or greater. This formula assumes a window of $2d$ cM around each marker in which linked markers can be detected, which does not account for chromosome ends and thus will tend to overestimate map length. For markers within d cM of a chromosome end, the average position is $d/2$ cM from the chromosome end, so these markers have an average window size of $3d/2$ rather than $2d$. This assumes that marker locations follow a uniform probability distribution and that all chromosomes are at least $2d$ in length. The proportion of the genome in these regions is $2Cd/L$, where C is the haploid chromosome number. Accordingly, we also used an adjusted estimate for L:

$$\hat{L}_a = \frac{2Cd}{L}\frac{n(n-1)(3/4)d}{k} + \left[1 - \frac{2Cd}{L}\right]\frac{n(n-1)d}{k}. \quad (3)$$

As this estimate itself contains L, we set $L = \hat{L}_a$, multiply by $\hat{L}_a$ and rearrange to obtain the quadratic equation:

$$\hat{L}_a^2 - \frac{n(n-1)d}{k}\ \hat{L}_a - \frac{Cn(n-1)d^2}{2k} = 0.$$

Solving the quadratic equation for $\hat{L}_a$ and rearranging yields the solution:

$$\hat{L}_a = \frac{n(n-1)d}{2k}\left(1 + \left[1 - \frac{2Ck}{n(n-1)}\right]^{1/2}\right). \quad (4)$$

A second solution, in which the radical is subtracted rather than added, is artifactual.

References

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet 49: 985–994

Buetow KH, Chakravarti A (1987) Multipoint gene mapping using seriation. Am J Hum Genet 41: 180–201

Chakravarti A, Lasher LK, Reefer JE (1991) A maximum likelihood method for estimating genome length using genetic linkage data. Genetics 128: 175–182

Collins A, Teague J, Keats BJ, Morton NE (1996) Linkage map integration. Genomics 36: 157–162

Cooper DN, Krawczak M (1990) The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Hum Genet 85: 55–74

Devey ME, Fiddler TA, Liu B-H, Knapp SJ, Neale DB (1994) An RFLP linkage map for loblolly pine based on a three-generation outbred pedigree. Theor Appl Genet 88: 273–278

Devey ME, Bell JC, Smith DN, Neale DB, Moran GF (1996) A genetic linkage map for *Pinus radiata* based on RFLP, RAPD, and microsatellite markers. Theor Appl Genet 92: 673–679

Echt CS, Nelson CD (1997) Linkage mapping and genome length in eastern white pine (*Pinus strobus* L). Theor Appl Genet 94: 1031–1037

Edwards JH (1991) The reliability of locus orderings. Ann Hum Genet 55: 315–320

Ehm MG, Kimmel M, Cottingham RWJ (1996) Error detection for genetic data, using likelihood methods. Am J Hum Genet 58: 225–234

Elsner TI, Albertsen H, Gerken SC, Cartwright P, White R (1995) Breakpoint analysis: precise localization of genetic markers by means of nonstatistical computation using relatively few genotypes. Am J Hum Genet 56: 500–507

Falk CT (1992) Preliminary ordering of multiple linked loci using pairwise linkage data. Genet Epidemol 9: 367–375

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137: 1121–1137

Groover AT, Williams CG, Devey ME, Lee JM, Neale DB (1995) Sex-related differences in meiotic recombination frequency in *Pinus taeda*. J Hered 86: 157–158

Hulbert SH, Hott TW, Legg EJ, Lincoln SE, Lander ES, Michelmore RW (1988) Genetic analysis of the fungus, *Bremia lactucae*, using restriction fragment length polymorphisms. Genetics 120: 947–958

Jiang C, Zeng Z-B (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101: 47–58

Keats BJB, Sherman SL, Morton NE, Robson EB, Buetow KH, Cartwright PE, Chakravarti A, Francke U, Green PP, Ott J (1991) Guidelines for human linkage maps – an International System for Human Linkage Maps (ISLM 1990). Ann Hum Genet 55: 1–6

Kinlaw CS, Neale DB (1997) Complex gene families in pine genomes. Trends Plant Sci 2: 356–359

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220: 671–680

Kubisiak TL, Nelson CD, Nance WL, Stine M (1995) RAPD linkage mapping in a longleaf pine × slash pine F_1 family. Theor Appl Genet 90: 1119–1127

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174–181

Lange K, Boehnke M (1982) How many polymorphic genes will it take to span the human genome? Am J Hum Genet 24: 842–845

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics 14: 604–610

Liu B-H (1998) Statistical genomics: linkage, mapping and QTL analysis. CRC Press, Boca Raton, Fla.

Lyon MF (1976) Distribution of crossing-over in mouse chromosomes. Genet Res Cambridge 28: 291–299

Marques CM, Carocha VJ, Araujo JA, Ferreira JG, O'Malley DM, Liu B-H, Sederoff R (1997) Mapping in *Eucalyptus* for tree improvement – a comparison of the PGRI and MAPMAKER software In: IUFRO Conf Silvicul Improv Eucalypt, Salvador, EMBRAPA. Centro Nacional de Pesquisa de Florestas, Colombo, PR, Brazil, pp 116–122

Marques CM, Araujo JA, Ferreira JG, Whetten R, O'Malley DM, Liu B-H, Sederoff R (1998) AFLP genetic maps of *Eucalyptus globulus* and *E. tereticornis*. Theor Appl Genet 96: 727–737

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277–318

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, Quillen J, Sheffield VC, Sunden S, Duyk GM, Weissenbach J, Gyapay G, Dib C, Morrissette J, Lathrop GM, Vignal A, White R, Matsunami N, Gerken S, Melis R, Albertsen H, Plaetke R, Odelberg S, Ward D, Dausset J, Cohen D, Cann H (1994) A comprehensive human linkage map with centiMorgan density. Science 265: 2049–2054

Nachman MW, Churchill GA (1996) Heterogeneity in rates of recombination across the mouse genome. Genetics 142: 537–548

Nelson CD, Nance WL, Doudrick RL (1993) A partial genetic linkage map of slash pine (*Pinus elliottii* Engelm. var. *elliottii*) based on random amplified polymorphic DNAs. Theor Appl Genet 87: 145–151

Nelson CD, Kubisiak TL, Stine M, Nance WL (1994) A genetic linkage map of longleaf pine (*Pinus palustris* Mill.) based on random amplified polymorphic DNAs. J Hered 85: 433–439

Newell WR, Mott R, Beck S, Lehrach H (1995) Construction of genetic maps using distance geometry. Genomics 30: 59–70

Ott J (1991) Analysis of human genetic linkage, revised edn. Johns Hopkins University Press, Baltimore

Ott J (1993) Detecting marker inconsistencies in human gene mapping. Hum Hered 43: 25–30

Paglia GP, Olivieri AM, Morgante M (1998) Towards second-generation STS (sequence-tagged sites) linkage maps in conifers: a genetic map of Norway spruce (*Picea abies* K.). Mol Gen Genet 258: 466–478

Pfeiffer A, Olivieri AM, Morgante M (1997) Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). Genome 40: 411–419

Plomion C, O'Malley DM (1996) Recombination rate differences for pollen parents and seed parents in *Pinus pinaster*. Heredity 77: 341–350

Plomion C, Bahrman N, Durel CE, O'Malley DM (1995a) Genomic mapping in *Pinus pinaster* (maritime pine) using RAPD and protein markers. Heredity 74: 661–668

Plomion C, O'Malley DM, Durel CE (1995b) Genomic analysis in maritime pine (*Pinus pinaster*) – comparison of 2 RAPD maps using selfed and open-pollinated seeds of the same individual. Theor Appl Genet 90: 1028–1034

Rouppe van der Voort JNAM, van Zandvoort P, van Eck HJ, Folkertsma RT, Hutten RCB, Draaistra J, Gommers FJ, Jacobsen E, Helder J, Bakker J (1997) Use of allele specificity of comigrating AFLP markers to align genetic maps from different potato genotypes. Mol Gen Genet 255: 438–447

Saylor LC, Smith BW (1966) Meiotic irregularity in species and interspecific hybrids of *Pinus*. Am J Bot 43 : 453–468

Shields DC, Collins A, Buetow KH, Morton NE (1991) Error filtration, interference, and the human linkage map. Proc Natl Acad Sci USA 88 : 6501–6505

Thompson EA (1987) Crossover counts and likelihood in multipoint linkage analysis. IMA J Math Appl Med 4 : 93–108

Travis SE, Ritland K, Whitham TG, Keim P (1998) A genetic linkage map of pinyon pine (*Pinus edulis*) based on amplified fragment length polymorphisms. Theor Appl Genet 97 : 871–880

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23 : 4407–4414

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

Young WP, Wheeler PA, Coryell VH, Keim P, Thorgaard GH (1998) A detailed linkage map of rainbow trout produced using doubled haploids. Genetics 148 : 839–850