Theor Appl Genet (1993) 86:880–888

THEORETICAL AND APPLIED GENETICS



# Duplicate sequences with a similarity to expressed genes in the genome of *Arabidopsis thaliana*

**J. M. McGrath*, M. M. Jancso, E. Pichersky**

Department of Biology, University of Michigan, Ann Arbor MI 48109-1048, USA

Received: 4 December 1992 / Accepted: 4 January 1993

**Abstract.** The proportion of non-tandem duplicated loci detected by DNA hybridization and the segregation of RFLPs using 90 independent randomly isolated cDNA probes was estimated by segregation analysis to be 17%. The 14 cDNA probes showing duplicate loci in progeny derived from a cross between *Arabidopsis-thaliana* ecotypes 'Columbia × Landsberg *erecta*' detected an average of 3.6 loci per probe (ranging from 2 to 6). The 50 loci detected with these 14 probes were arranged on a genetic map of 587 cM and assigned to the five *A. thaliana* chromosomes. An additional duplicated locus was detected in progeny from a cross between 'Landsberg *erecta* × Niederzenz'. The majority of duplicated loci were on different chromosomes, and when linkage between duplicate locus pairs was detected, these loci were always separated by at least 15 cM. When partial nucleotide sequence data were compared with GENBANK databases, the identities of 2 cDNA clones which recognized duplicate unlinked sequences in the *A. thaliana* genome were determined to encode a chlorophyll *a/b*-binding protein and a *beta*-tubulin. Of the 8 loci carrying *beta*-tubulin genes 6 were placed on the genetic map. These results imply that gene duplication has been an important factor in the evolution of the *Arabidopsis* genome.

**Key words:** cDNA clones – RFLP – Genetic mapping – *Beta*-tubulin – Gene duplication

Communicated by H. F. Linskens

* *Present address:* Department of Plant Pathology, 1630 Linden Dr., University of Wisconsin, Madison, WI 53706, USA
*Correspondence to:* J. M. McGrath

## Introduction

The divergence of duplicate genes presents the organism with a chance to acquire novel gene functions. Selection for the original function would maintain one copy of a duplicate gene pair relatively unchanged, while the other copy could evolve to encode a product with altered biochemical, structural, developmental or regulatory functions. The potential of gene duplication has long been recognized (Haldane 1932; Ingram 1961; Ohno 1970), but the rate by which new genetic functions arise depends on a number of poorly understood parameters, including the proportion of duplicate genes per genome and the particular lineages and gene pairs.

Traditionally, duplicate genes have been defined genetically as those gene pairs that are identical in function, thus giving rise to the same phenotype. However, recent investigations using DNA hybridization and cloning techniques have uncovered many instances of groups of genes, known as 'gene families', that encode similar but distinct proteins within the same genome [e.g., in *Arabidopsis thaliana*, genes encoding *beta*-tubulins (Snustad et al. 1992), chlorophyll *a/b*-binding proteins (McGrath et al. 1992), ribulose-1,5-bisphosphate carboxylases (Krebbers et al. 1988), S-adenosylmethionine synthetases (Peleman et al. 1989) and 12S seed storage proteins (Pang et al. 1988)], whose individual members may have different functions, but from nucleotide sequence comparisons it is clear that they arose through gene duplication and divergence. Depending on the experimental conditions, many members of a gene family can often be detected as numerous bands on Southern blots after hybridization with a cloned gene.

Restriction fragment length polymorphisms (RFLPs) exhibited by different individuals can be exploited to determine linkage relationships between gene family members when combined with genetic analysis of segregating populations. In crops such as maize (Helentjaris et al. 1988), soybean (Keim et al. 1990) and cultivated Brassicas (Slocum et al. 1990; Song et al. 1991), where whole or partial genome duplication has been inferred from cytogenetic studies, an examination of genetic maps based

on RFLP markers show that many probes disclose multiple loci (ca. 25% or more), and conserved linkage groups can be detected among different chromosomes. In contrast, in RFLP maps developed in species where cytogenetics has uncovered no evidence of chromosome-level duplication, such as tomato (Bernatzky and Tanksley 1986a), rice (McCouch et al. 1988), lettuce (Landry et al. 1987) and *Arabidopsis thaliana* (Chang et al. 1988; Nam et al. 1989), duplicated loci have been detected infrequently ($< 10\%$), and the majority of these duplicate loci appear to be unlinked. In either type of species, however, probes and hybridization conditions which generate complex patterns were generally not used in the development of RFLP maps simply to aviod complications arising in their analysis. Thus, the level of duplicated sequences reported in various plant species may be underestimated.

In the investigation reported here we have used randomly chosen cDNA clones to assess the extent and distribution of duplicated loci in the genome of *Arabidopsis thaliana*. We chose cDNA clones because they represent the transcripts of expressed genes, and duplicate loci detected by cDNA clones are candidate members of gene families. We chose *Arabidopsis thaliana* because it has one of the smallest genome sizes among the angiosperms (Leutwiler et al. 1984; Arumuganathan and Earle 1991), and it was of interest to examine its level of sequence duplication. Also, other members of the Crucifereae have been similarly characterized with cDNA clones as RFLP markers, such as *Brassica oleracea* (Landry et al. 1992) and *Brassica rapa* (syn. *campestris*) (McGrath and Quiros 1991), and have been shown to contain many duplicated loci. In contrast to *Brassica*, however, there is no evidence that whole or partial linkage groups are duplicated in *A. thaliana* (Koornneef et al. 1983; Chang et al. 1988; Nam et al. 1989; Reiter et al. 1992). Thus, the level of duplicated cDNA loci in *A. thaliana* should best approximate a basal level of duplicated genes among crucifers.

## Materials and methods

### *Plant material*

Ecotypes of *A. thaliana* used in this study included 'Blanes', 'Büchen', 'Columbia', 'Dijon', 'Estland', 'Landsberg *erecta*', 'Niederzenz', 'Rschew' and 'Wassileskija' (Kranz and Kirchheim 1987). However, our primary focus was on 'Columbia', 'Landsberg *erecta*' and 'Niederzenz'. A total of 54 $F_2$ plants derived from a single $F_1$ individual of the 'Landsberg *erecta* × Niederzenz' cross and 60 $F_2$ plants derived from a single $F_1$ individual of the 'Columbia × Landsberg *erecta*' cross were allowed to self. Approximately 150 $F_3$ seeds from each individual $F_2$ plant were grown for DNA extraction.

### *Genomic blots*

DNA was extracted as described (McGrath et al. 1992) and digested (5 µg per sample) with one of the four restriction enzymes, *Bgl*II, *Eco*RI, *Hin*dIII and *Xba*I (all enzymes from Boehringer Mannheim Biochemicals). Occasionally, *Dra*I, *Eco*RV, *Sac*I or *Hae*III were used. Gel electrophoresis (0.8% agarose in TAE) and alkaline-blotting to Hybond-N membranes (Amersham) were performed as previously described (McGrath et al. 1992).

Two types of analyses were performed. In the first case, test blots consisting of the three primary ecotypes digested with the four enzymes (see above) were used to uncover RFLPs and to obtain an estimate of copy number. Subsequently, putative duplications which also showed polymorphism were examined for segregation of each RFLP in the pooled $F_3$ families (which represent the $F_2$ population).

Hybridization was carried out at 65 °C in 5 × 2 SSPE, 5 × Denhardt's solution, 0.5% SDS and 0.1 mg/ml fish sperm DNA, with two washes in 2 × SSC, 0.75% SDS at 65 °C (McGrath et al. 1992). Test blots were reused up to ten times after the probes had been strip-washed in boiling 0.1 × SSC, 1% SDS, and the $F_2$ blots were reused up to six times.

### *Probe preparation and nucleotide sequence analysis*

A cDNA library constructed in lambda ZAP II (Stratagene) from 3-day-old above-ground tissues of 'Columbia' was kindly provided by Dr. J. Ecker. The plasmid rescue protocol (Stratagene) was performed on an aliquot of the 1.0- to 2.0-kb size-fractionated library. Inserts were recovered from 94 plasmid clones after digestion with *Eco*RI and labelled with [$^{32}$P] to a high specific activity using random sequence oligo-nucleotide primers.

Nucleotide sequences were determined enzymatically with the aid of the Sequenase DNA sequencing kit as described (United States Biochemical). Sequences were checked with the GENBANK nucleotide sequence database (release 69) and the associated polypeptide sequence database for similarity to previously characterized sequences.

### *Linkage analysis*

When probes which detected duplicated loci were used, the segregation patterns were generally complex. To discriminate alleles at a locus, we scored each segregating band from among the many bands segregating as present or absent in each $F_3$ family. The data was then examined for goodness-of-fit to a 3:1 segregation ratio and for linkage. A recombination frequency of 0.0 indicated which bands were allelic (if inherited from different parents) or the same allele (if inherited from the same parent). If allelism was detected, scores were combined and re-tested for goodness-of-fit to a 1:2:1 segregation ratio. In cases where allelism was not detected (i.e., a band segregating 3:1 could not be paired with another band segregating 3:1 at a recombination fraction of 0.0), additional restriction enzymes were tested in order to reveal an alternate allele; a procedure that was not always successful.

The LINKAGE-1 computer algorithm (Suiter et al. 1983) was used for segregation goodness-of-fit, and linkage analyses were performed using MAPMAKER (Lander et al. 1987). Chromosome identities were assigned with the aid of previously mapped RFLP markers from Chang et al. (1988) and the morphological character *erecta*.

## Results

### *Estimate of proportion of duplicated and single-copy sequences*

Each of 94 *Arabidopsis thaliana* cDNA clones was hybridized to a panel of genomic DNA from *A.*

*thaliana* ecotypes 'Landsberg *erecta*' and 'Columbia' and/or 'Niederzenz' digested with a series of restriction enzymes, including *Bgl*II, *Eco*RI, *Hin*dIII and *Xba*I. On the basis of the similarity of hybridization patterns and partial nucleotide sequence analysis, 3 clones were represented more than once in this population of plasmids. Of these, 2 were recovered twice (CAB and *beta*-tublin, see below), and a third unidentified sequence was recovered three times. Thus, 90 independent clones were tested.

A preliminary estimate of sequence duplication was obtained by examining the number of bands on autoradiographs in the following manner. Any probe which hybridized only to a single band for 1 or more restriction digests in all tested ecotypes was considered as a single locus (see, for example, Fig. 1A, B); 44 out of 90 clones (49%) showed this pattern. This estimate of 49% represents a minimum estimate of single-copy loci in the *Arabidopsis* genome. The remaining 46 probes (51%) detected 2 or more fragments in all of the probe-enzyme combinations tested and were considered to be potential duplicate loci (for example, Fig. 1C, D). This value represents the maximum estimate for the proportion of duplicated sequences similar to expressed genes among this population of clones.

### *Inheritance of duplicated sequences*

When polymorphism between parents was evident, as in the case of probe 559 (Fig. 1D, probe nomenclature follows from order of isolation on a particular date, i.e., 5/5/91 no. 9 = 559), the inheritance of the RFLPs was examined in the 'Columbia × Landsberg' $F_3$ population. Some RFLP markers which did not segregate in 'Columbia × Landsberg' were tested in a similar population derived from 'Landsberg × Niederzenz'. Fourteen probes disclosed 2 or more loci in 'Columbia × Landsberg' (Table 1), and an additional duplicated locus was identified in 'Landsberg × Niederzenz' (probe 171, whose multiple-banded pattern was identical between 'Columbia' and 'Landsberg'). Thus, 15 out of 90 (17%) cDNA clones revealed segregating duplicated loci in *A. thaliana*.

In 'Columbia × Landsberg', the 14 probes showing multiple segregating bands detected 50 loci, with a mean of 3.6 loci per probe (SD = 1.55). For probes which detected multiple loci, the average number of loci was significantly greater than 2 (Student's *t*- test, at the 0.001 level). Similarly, in the 'Landsberg × Niederzenz' $F_3$ population, the 5 probes tested revealed 13 loci (*mean* = 2.60, SD = 0.55, data not shown).

In most cases, both parental alleles at a single locus were disclosed with a single restriction enzyme (Table 1). However, 11 of the 50 duplicate loci (22%) scored in 'Columbia × Landsberg' did not show one or the other parental alleles (Table 1). No bias in the parents was detected for loci scored in this way; 5 loci carried the 'Columbia' allele but not the 'Landsberg' allele, and 6 loci carried the 'Landsberg' allele but not the 'Columbia' allele. Although RFLP markers are generally expected to be co-dominant (i.e., both alleles at a locus can be scored), attempts to find the missing alleles by using a total of four restriction enzymes failed to uncover them, except for probe 415 (see below). In all cases, additional bands were observed which were monomorphic and could not be scored, hence an alternate allele could have been hidden if it co-migrated with one of these fragments.

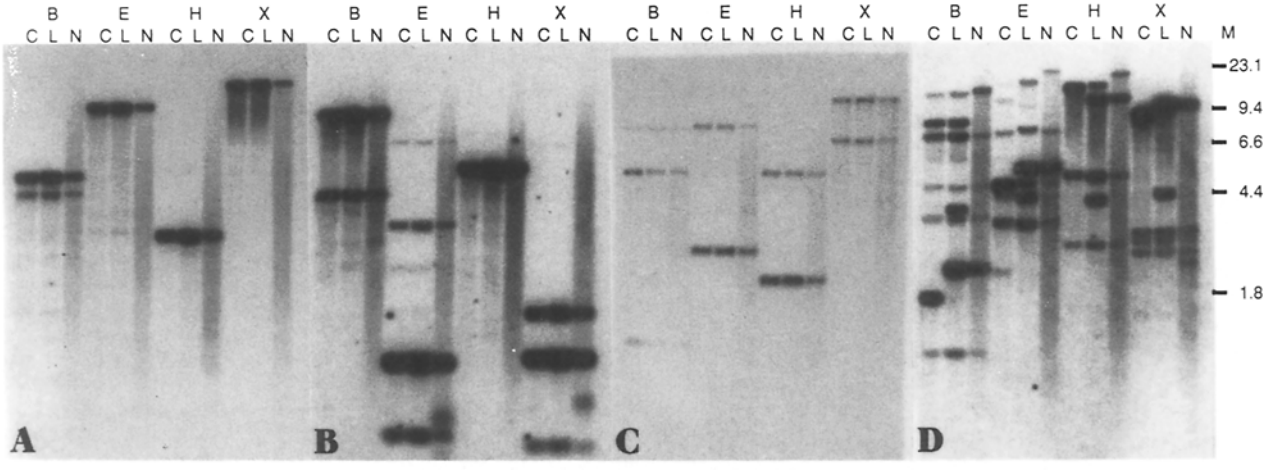


**Fig. 1A–D.** Detection of putative sequence duplications from test blots of 'Columbia' (*C*), 'Landsberg *erecta*' (*L*), and 'Neiderzenz' (*N*) digested with *Bgl*II (*B*), *Eco*RI (*E*), *Hin*dIII (*H*) or *Xba*I (*X*). A single band in all ecotypes for at least one enzyme was considered to be a single-copy locus, as shown in panels **A** (probe 241) and **B** (probe 243). Probes showing more than one band in all ecotypes for all enzymes were considered to be putative duplications, as indicated in panel **C** (probe 168) and **D** (probe 559)

**Table 1.** Segregation and chi-square analysis of duplicated RFLP loci detected with cDNA probes in *Arabidopsis thaliana*

| Marker name[a] | Homozygous Col | Heterozygous | Homozygous Ler | $\chi^2$ | Probability | Enzyme[c] |
|---|---|---|---|---|---|---|
| Cross Col × Ler | | | | | | |
| 173A | 9 | 32 | 13 | 2.44 | 0.29 | X |
| 173B | 10 | -44- | | 1.21 | 0.27 | X |
| 173C | 14 | 27 | 11 | 0.42 | 0.81 | X |
| 173D | 11 | -43- | | 0.62 | 0.43 | X |
| 177A (*Tub*9) | 8 | 27 | 18 | 3.79 | 0.15 | E |
| 177B (*Tub*6) | 13 | 26 | 15 | 0.22 | 0.89 | E |
| 177C (*Tub*5) | 18 | 20 | 16 | 3.77 | 0.15 | E |
| 177D (*Tub*1) | 10 | 24 | 20 | 4.37 | 0.11 | E |
| 177E (*Tub*4) | 15 | 28 | 11 | 0.67 | 0.72 | E |
| 710A (*Tub*2/3) | -39- | | 14 | 0.06 | 0.81 | E |
| 178A (*Lhb*1A) | 16 | 21 | 15 | 1.96 | 0.38 | X |
| 178B (*Lhb*1B) | 9 | 29 | 14 | 1.65 | 0.44 | X |
| 178C (*Lhb*2) | 12 | 26 | 14 | 0.15 | 0.93 | X |
| 415A | 7 | 27 | 17 | 4.10 | 0.13 | E, X |
| 415B | 8 | 29 | 15 | 2.58 | 0.28 | E, X |
| 415C | 10 | 25 | 18 | 2.58 | 0.27 | X |
| 415D | 14 | 24 | 15 | 0.51 | 0.78 | X |
| 415E | 10 | 30 | 13 | 1.26 | 0.53 | X |
| 415F | -40- | | 13 | 0.01 | 0.93 | E |
| 511A | 15 | 12 | 16 | 8.44 | 0.01* | B |
| 511B | 12 | -31- | | 0.19 | 0.66 | B |
| 511C | 6 | 25 | 17 | 5.13 | 0.08 | H |
| 515A | 8 | 24 | 20 | 5.85 | 0.05 | X |
| 515B | -34- | | 14 | 0.44 | 0.50 | X |
| 515C | 8 | 28 | 16 | 2.77 | 0.25 | X |
| 515D | -45- | | 8 | 2.77 | 0.10 | X |
| 555A | 11 | 31 | 10 | 1.96 | 0.38 | E |
| 555B | 7 | 31 | 14 | 3.81 | 0.15 | E |
| 559A | 11 | -39- | | 0.24 | 0.62 | H |
| 559B | 8 | 28 | 18 | 3.78 | 0.15 | H |
| 559C | 15 | 21 | 16 | 1.96 | 0.38 | E |
| 559D | 12 | -38- | | 0.03 | 0.87 | E |
| 571A | 13 | 34 | 7 | 4.96 | 0.08 | X |
| 571B | 10 | 25 | 18 | 2.58 | 0.27 | X |
| 574A | 6 | 25 | 13 | 3.05 | 0.22 | B |
| 574B | 5 | 22 | 7 | 3.18 | 0.20 | B |
| 579A | 10 | 22 | 20 | 5.08 | 0.08 | B |
| 579B | 10 | 30 | 14 | 1.26 | 0.53 | B |
| 579C | 9 | 32 | 13 | 2.44 | 0.29 | B |
| 579D | 15 | 24 | 15 | 0.67 | 0.72 | E |
| 596A | 9 | 28 | 13 | 1.36 | 0.51 | H |
| 596B | 10 | -40- | | 0.67 | 0.41 | H |
| 611A | 9 | 32 | 13 | 2.44 | 0.29 | X |
| 611B | 8 | 27 | 19 | 4.48 | 0.11 | X |
| 711A | 9 | 24 | 15 | 1.50 | 0.47 | H |
| 711B | 9 | 25 | 6 | 2.95 | 0.23 | B |
| 713A | 11 | 24 | 8 | 1.00 | 0.61 | B |
| 713B | 12 | 22 | 20 | 4.22 | 0.12 | X |
| 713C | 8 | 30 | 15 | 2.77 | 0.25 | X |
| 713D | -30- | | 12 | 0.29 | 0.59 | B |
| Cross Ler × Nd | | | | | | |
| | Ler | | Nd | | | |
| 171A[b] | -36- | | 19 | 2.67 | 0.10 | X |
| 171B[b] | 17 | 27 | 11 | 1.33 | 0.51 | X |

* Distorted segregation ratio

Col, 'Columbia'; Ler, 'Landsberg *erecta*'; Nd, 'Neiderzenz'

[a] *Lhb* Photosystem II chlorophyll *a*/*b*-binding (CAB) protein genes, *Tub beta*-tubulin genes

[b] Scored in 'Ler × Nd' only

[c] Enzymes used to score the RFLP locus: B, *Bgl*II; E, *Eco*RI; H, *Hin*dIII; X, *Xba*I. When two enzymes are indicated, both were used to score a single locus (see text)

Probe 415 showed eight polymorphic bands on *Xba*I-digested DNA. Six bands showed allelism (indicating 3 loci) and two fragments (indicating 2 additional loci) whose alternate parent's allele could not be identified. However, *Eco*RI-digested samples disclosed a segregating RFLP for each alternate allele at both loci. In addition, the *Eco*RI-digested samples also showed another locus not previously detected in *Xba*I-digested DNAs (Table 1). In general, the use of additional restriction enzymes either uncovered additional loci (e.g., probes 415, 559, 579, 711 and 713, Table 1) or occasionally detected the same locus previously identified with another enzyme (e.g., locus 515A was polymorphic with three enzymes, and loci 579 A, B and C were each polymorphic with two enzymes). It is unlikely that many loci were missed for this set of duplicated sequences using the four enzymes, *Bgl*II, *Eco*RI, *Hin*dIII and *Xba*I.

### *Linkage between duplicated sequences*

The arrangement of duplicated loci in the *A. thaliana* genome is graphically represented as a genetic map in Fig. 2. A total of 64 loci were scored, including 50 duplicate loci and 8 single-copy loci. The morphological marker *erecta* and 5 previously mapped RFLP loci (Chang et al. 1988) were used to orient the map. Three loci (511B, 596B, 559D; each scored 3:1) failed to locate to a chromosome, and 2 others (173C and 571B) showed a loose affinity to chromosomes 4 and 5, respectively. The remaining 59 loci were arranged in five linkage groups that spanned a total of 587.2 centiMorgans (cM) of the *A. thaliana* genome (Fig. 2) Recombination values between adjacent markers ranged from 0.0 cM to 28.8 cM, with an average spacing between markers of 10.9 cM ($n = 54$, SD = 6.7 cM).

The majority of duplicate copies of a given sequence were not tightly linked with one another and were often located on different chromosomes. When linkage was apparent, no duplicate loci detected by the same probe were less than 15 cM apart. Among the closest were loci 415B and 415F, separated by 16.6 cM, and 173 A and 173D, linked by 19.4 cM (chromosomes 4 and 2, respectively; Fig. 2).

### *Tandem duplications*

Putative tandemly linked duplications were identified (but not included in the total derived from the genetic segregation analysis) based on the following assumption. An average gene (including promoter regions, gene sequence, introns and terminator regions) is approximately 3 kb in size. Thus, although two co-segregating fragments of 3 kb each could represent a single gene interrupted by a restriction site within the region detected by the probe, three or more fragments each greater than 3 kb are likely to represent tandem duplications. Two loci showed this pattern, 713 C and 712. Locus 713C showed a three-banded phenotype of linked fragments derived from 'Landsberg' on *Xba*I-digested DNA, the combined size of which exceeds 40 kb (Fig. 3). Similar three-banded phenotypes for this locus in 'Columbia' (and 'Landsberg') were also observed in *Eco*RI digests. This observation, and the difference in intensity between 713B and 713C (see Fig. 3), suggest that 713C contains tandemly duplicated genes at this locus.

Probe 712 showed no variation between ecotypes 'Columbia' or 'Landsberg' with any of the tested en-

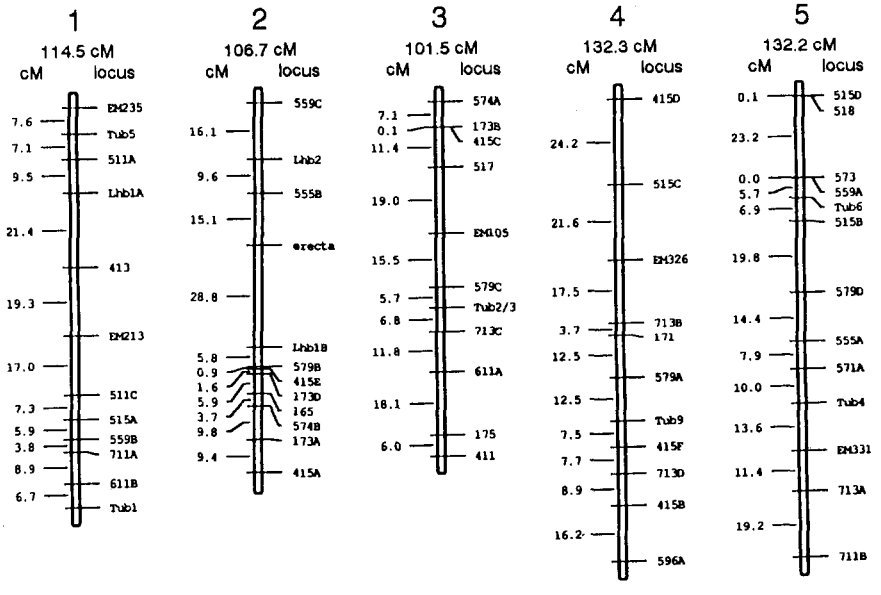


**Fig. 2.** Linkage relationships between duplicated RFLP loci as revealed by cDNA probes. Chromosome assignment (but not necessarily orientation), given at the *top*, was determined with previously mapped probes (prefixed *EM*, Chang et al. 1988) and the character *erecta*. Duplicated loci are suffixed with a *letter*; the remainder were single-copy loci. The map was created with MAPMAKER (Lander et al. 1987) for LOD scores greater than 3.0 using the Kosambi mapping function

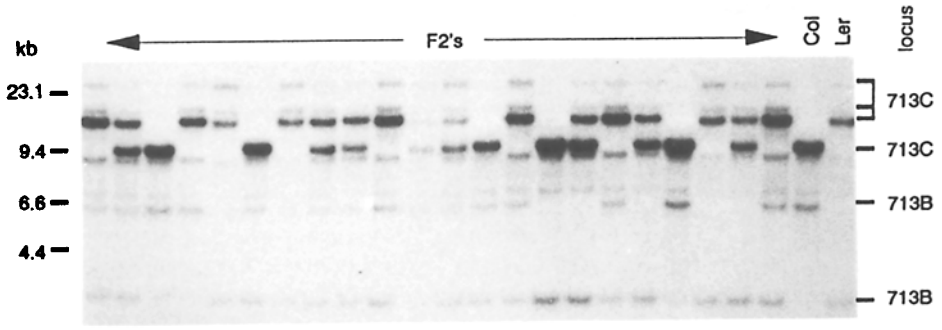


**Fig. 3.** Putative tandem duplication detected at locus 713C on *Xba*I digests of $F_3$ families. Three co-segregating fragments of approximately 15 kb, 17 kb and 24 kb derived from 'Columbia' (*Col*) suggest a tandem duplication at this locus [also present in 'Landsberg *erecta*' (*Ler*); see text]. An additional locus (713B) was also detected with this probe-enzyme combination

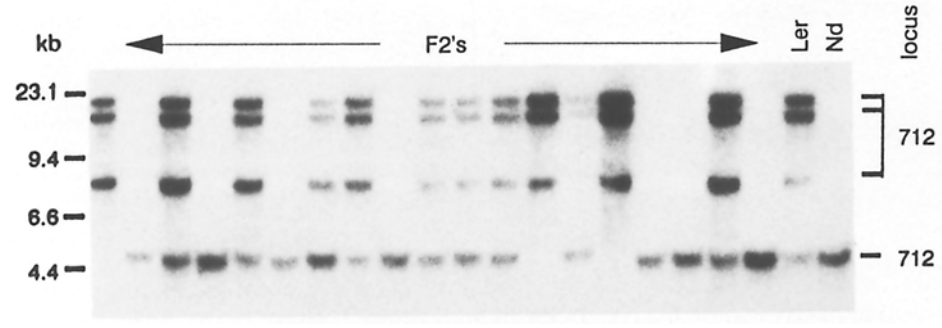


**Fig. 4.** Apparent ecotype-specific tandem duplication using probe 712 on *Eco*RI-digested samples from 'Landsberg *erecta*' (*Ler*, duplication present), 'Niederzenz' (*Nd*, duplication absent), segregating in the $F_3$ families of 'Ler × Nd'

zymes but did show three co-segregating bands of approximately 8 kb, 16 kb and 18 kb in both using *Eco*RI, which suggests that there are at least two copies at this locus in these ecotypes (Fig. 4). Ecotype 'Niederzenz' showed a single band of approximately 5 kb, suggestive of a single-copy locus. Analysis of the 'Landsberg × Niederzenz' $F_3$ families demonstrated that the 'Landsberg' fragments co-segregated as an allele of the 'Niederzenz' fragment. Thus, this locus represents an ecotype-specific tandem duplication within *A. thaliana*. The distribution of this duplication among accessions of *A. thaliana* was examined with *Eco*RI- and *Eco*RV-digested DNA from six additional ecotypes: 'Blanes', 'Büchen', 'Dijon', 'Estland' and 'Wassileskija'. Our results suggest that ecotypes 'Blanes', 'Estland' and 'Wassileskija' are similar to 'Niederzenz' in possessing a single copy in this locus. Ecotypes 'Dijon' and 'Rschew' (and perhaps 'Büchen') appeared to carry the duplication seen in 'Columbia' and 'Landsberg' (data not shown).

*Identity of cDNA clones*

An average of 290 bp (ranging from 180 to 350 bp) of nucleotide sequence was obtained for 7 of the 14 clones that revealed duplicate loci. Two gene products were identified by comparisons with nucleotide and polypeptide sequences in the GENBANK and associated databases; one encodes a Photosystem-II Type-I CAB polypeptide (from 200 bp of partial sequence data of clone 178), and the other clone (from 330 bp of partial sequence from clone 177) encodes *beta*-tubulin (i.e., *Tub*5, P. Snustad, personal communication). The remaining sequences did not closely match any GENBANK sequence. Comparison of the CAB gene sequence we obtained with previously reported sequences shows this cDNA clone to be derived from one *Lhb*1A gene (Leutwiler et al. 1986). The inheritance of genes at this and related loci, but not their linkage relationships, has been previously reported (McGrath et al. 1992).

The inheritance of genes encoding *beta*-tubulin has not been previously described in *A. thaliana*, but most if not all of these genes have been isolated and sequenced (Marks et al. 1987; Oppenheimer et al. 1988; Snustad et al. 1992). Nine genes (*Tub*1 through *Tub*9) are present and expressed, and only one pair (*Tub*2 and *Tub*3) reside at the same locus (separated by 1 kb of noncoding sequence; Snustad et al. 1992). Comparison of *Eco*RI-digested 'Columbia' DNA probed with the

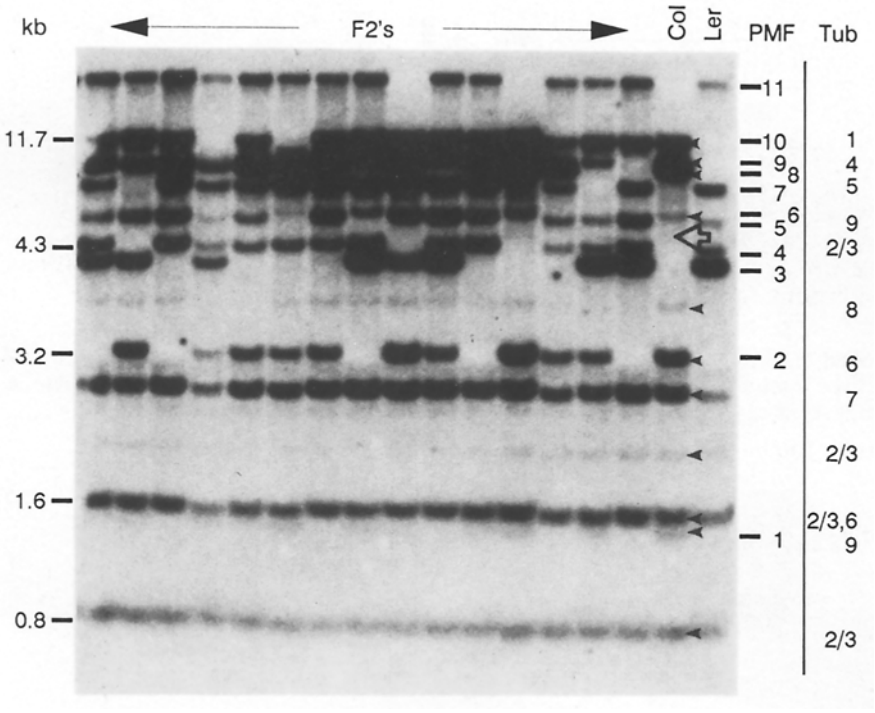


**Fig. 5.** Segregation of RFLPs at *beta*-tubulin loci detected by probe 177 in *Eco*RI-digested DNA from parents 'Columbia' (*Col*), 'Landsberg *erecta*' (*Ler*) $F_3$ families derived from 'Col × Ler'. Probe 177 was identified as the *beta*-tubulin (*Tub*) transcript of *Tub*5 (see text). Each of 11 polymorphic fragments (PMF) was scored for its presence or absence, and linkage analysis on these scores indicated that the co-segregating PMFs 1 and 6 were allelic to PMF 5 (corresponding to *Tub*9), PMFs 2 and 4 were alleles of *Tub*6, PMFs 3 and 8 were alleles of *Tub*5, PMFs 7 and 10 were alleles of *Tub*1 and PMFs 9 and 11 were alleles of *Tub*4. *Arrows* indicate *beta*-tubulin genes described by Snustad et al. (1992) in *Eco*RI-digested 'Columbia' DNA. *Open arrow* indicates fragment not identified with probe 177 but detected by an additional *beta*-tubulin cDNA (probe 710) segregating at the position of *Tub*2/3 (see text)

entire coding sequence and also with gene-specific probes (Snustad et al. 1992) with our *Eco*RI-digested 'Columbia × Landsberg' $F_3$ families probed with our *beta*-tubulin cDNA clones 177 and 710 allowed us to correlate the segregating loci with specific *Tub* genes (Fig. 5). Although we did not obtain sequence data for probe 710, all of the bands observed with this probe were identical to those of the sequenced probe 177, with one exception. When probe 710 was used (whose insert was 400 bp shorter than probe 177), a fragment corresponding to *Tub*2/3 was evident at 4.3 kb; this fragment was absent from blots in which probe 177 had been used (Fig. 5). This result, plus the observation that our two *beta*-tubulin probes hybridized with different intensities to *Tub*2/3 fragments, suggests that these two *beta*-tubulin cDNA clones are derived from different genes. Overall, our results indicate that six of the eight *beta*-tubulin loci mapped to six unlinked loci on four *A. thaliana* chromosomes (Fig. 2). Two genes (*Tub*7 and *Tub*8) did not segregate and were not mapped.

## Discussion

Gene duplication may have been the mechanism of origin for almost all genes (Ohno 1970), so in that sense most genes are related to each other. However, most duplications likely occurred so long ago that neither inspection of the sequences nor experimental procedures such as DNA hybridization will detect these as duplications. Other, less ancient duplications can be inferred from inspection of protein sequences, but not by DNA hybridization, because of the degeneracy of the genetic code in specifying the amino acid composition. Therefore, the detection of gene duplication by the empirical method of DNA hybridization is less effective than a statistical analysis of sequence comparisons at the protein (or DNA) sequence levels. In addition to sequence similarity, the hybridization of a DNA probe to a duplicate copy depends on other experimental conditions (e.g., temperature, salt concentration, molecular weight distribution of unrestricted DNA and the number of times a blot has been re-used) and is thus difficult to predict. Also, the number and lengths of regions with high sequence identity between a probe and its target sequence, not the average sequence similarity between them, are the important controlling factors for the detection of duplicate sequences by DNA hybridization. Since DNA hybridization can detect only relatively recent gene duplications, the extent of duplications, and therefore an empirical definition of 'gene families', depends on the specific experimental techniques employed.

Duplicate gene copies may be identical only for a short time following the duplication event since both continue to evolve. Such duplicate genes may diverge substantially in their primary sequence and still maintain similar and overlapping functions. Also, both DNA hybridization experiments and nucleotide sequence determinations have uncovered a large number of related sequences for which little evidence is available concerning their function, yet they are clearly the

result of gene duplications. Thus, defining duplicate genes by (nearly) identical function or by DNA hybridization criteria covers only a subset of all gene duplications that exist in a genome.

In this study we have chosen DNA hybridization conditions such that the formation of extensively mismatched DNA duplexes is depressed (i.e., high temperature) but the retention of moderately mismatched hybrid duplexes is favored during the subsequent washing steps (i.e., high salt concentration). It is not possible from this data to conclude how related two sequences are because even a short stretch of identical nucleotides would be enough to give a detectable signal. However, we have chosen cDNA clones as probes because they represent an active protein-encoding portion of the genome, in contrast to other available DNA markers. Thus, at least one gene among the duplicated sequences detected by cDNA clones is expressed. Interestingly, the 2 cDNA clones which we positively identified, CAB and *beta*-tubulin, belong to gene families in which all or almost all of the genes are expressed in *A. thaliana* (Leutwiler et al. 1986; W. Terzaghi, personal communication; Snustad et al. 1992). We expect that many, and perhaps most, duplicate loci detected by these cDNA clones contain functional and expressed genes.

The duplicate sequences we have described in this report were widely distributed among all five *A. thaliana* chromosomes, with a few copies being linked by less than 20 cM. A similar distribution of duplicate loci in *A. thaliana* has been reported by Nam et al. (1989) for genes encoding glutamine synthetases, nitrate reductases and glyceraldehyde-3-phosphate dehydrogenases, as well as for duplicated genes encoding tryptophan synthases (Last et al. 1991) and anthranilate synthases (Niyogi and Fink 1992). Thus, non-linkage of gene duplications appears to be a common arrangement of duplicated genes in *A. thaliana* and in other plants as well. For instance, in *Brassica oleracea* and *B. campestris* (also in the crucifer family) the majority of duplicated RFLP loci detected by cDNA clones are unlinked (McGrath et al. 1990; McGrath and Quiros 1991; Landry et al. 1992). In an unrelated species, tomato, duplicated protein-encoding genes are also often unlinked (Vallejos et al. 1986; Bernatzky and Tanksley 1986b).

It is likely that some of the specific sequence duplications we have documented in *Arabidopsis thaliana* are also found in some of its close relatives, and therefore must have arisen prior to the divergence of *A. thaliana* from a common ancestor. Preliminary experiments utilizing these cDNA clones against the related species *Cardaminopsis arenosa* suggest that most of the duplicated loci in *A. thaliana* are duplicated in this species as well (data not shown). However, to determine the time of each duplication requires the progressive examination of increasingly more distant relatives for the presence or absence of a duplicated sequence.

Although we were able to infer the presence of two loci with tandem duplications from the RFLP data, we suspect that additional tandem duplications have gone undetected. In our initial screening, 46 of the 90 cDNA clones (51%) showed multiple fragments in all of the ecotypes and restriction digests, and therefore were considered to be putative sequence duplications. However, only 15 of these 46 cDNA clones showed multiple polymorphic bands that were subsequently determined by segregation analysis to be non-tandem duplicated sequences. With most of these 46 clones, we did not score multiple bands as tandem duplications unless they fulfilled specific requirements (i.e., at least three co-segregating bands, each greater than 3 kb; see Results). Thus, while the majority of duplicated sequences we report here are unlinked, this result does not preclude the existence of similar numbers of tandem duplications.

Since we have examined only 90 randomly isolated independent transcripts, these transcripts are likely to be derived from relatively highly expressed genes. For example, some CAB and tubulin transcripts have been shown to comprise more than 1% of total leaf mRNA, a level consistent with the isolation of multiple cDNA clones of these types from the cDNA library we used. However, there is no *a priori* reason to assume that genes expressed at lower levels should not show equivalent levels of sequence duplication as well. For example, several duplications of genes known to be expressed at lower lelvels have been described in *A. thaliana*, such as genes for tryptophan synthase (Last et al. 1991) and anthranilate synthase (Niyogi and Fink 1992).

Our understanding of the mechanisms which generate novel genetic diversity will require knowledge of the distribution of duplicated sequences in the genome, the frequency and timing of gene duplication events and the fates of duplicated sequences. Our observation that more than 15% of the genes in the *Arabidopsis thaliana* genome may be encoded by multiple loci suggests that the evolutionary potential of duplicate sequences is substantial.

*Acknowledgements.* We gratefully thank Dr. J. Ecker for the *A. thaliana* cDNA library, R. Wilson for providing $F_3$ seeds of 'Landsberg × Niederzenz' and Dr. J. Scheifelbein for assistance in making crosses. This work was funded by a grant from NSF to EP.

*Note added in proof:* The clones described here have been deposited in, and can be obtained from, the Arabidopsis Biological Resource Center at Ohio State University, 1735 Neil Avenue, Columbus, OH 43210 USA.

## References

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Bernatzky R, Tanksley SD (1986a) Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. Genetics 112:887–898

Bernatzky R, Tanksley SD (1986b) Genetics of actin-related sequences in tomato. Theor Appl Genet 72:314–321

Chang C, Bowman JL, DeJohn AW, Lander ES, Meyerowitz EM (1988) Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. Proc Natl Acad Sci USA 85:6856–6860

Haldane JBS (1932) The causes of evolution. Longmans, Green and Co, London

Helentjaris T, Weber D, Wright S (1988) Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. Genetics 118:353–363

Ingram VM (1961) Gene evolution and the hemoglobins. Nature 189:704–708

Keim P, Diers BW, Olson TC, Shoemaker RC (1990) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742

Koornneef M, van Eden J, Hanhart CJ, Stam P, Braaksma FJ, Feenstra WJ (1983) Linkage map of *Arabidopsis thaliana*. J Hered 74:265–272

Kranz AR, Kirchheim B (1987) Genetic resources in Arabidopsis. Arabidopsis Inf Serv 24:1–349

Krebbers E, Seurinck J, Herdies L, Cashmore AR, Timko MP (1988) Four genes in two diverged subfamilies encode ribulose-1,5-bisphosphate carboxylase small subunit polypeptides of *Arabidopsis thaliana*. Plants Mol Biol 11:745–749

Lander EC, 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

Landry BS, Kesseli RV, Farrara B, Michelmore RW (1987) A genetic map of lettuce (*Lactuca sativa* L.) with restriction length polymorphism, isozyme, disease resistance and morphological markers. Genetics 116:331–337

Landry BS, Hubert N, Crete R, Chang MS, Lincoln SE, Etoh T (1992) A genetic map for *Brassica oleracea* based on RFLP markers detected with expressed DNA sequences and mapping of resistance genes to race 2 of *Plasmodiophora brassicae* (Woronin). Genome 35:409–420

Last RL, Bissinger PH, Mahoney DJ, Radwanski ER, Fink GR (1991) Tryptophan mutants in *Arabidopsis*: the consequences of duplicated tryptophan synthase $\beta$ genes. Plant Cell 3:345–358

Leutwiler LS, Hough-Evans BR, Meyerowitz EM (1984) The DNA of *Arabidopsis thaliana*. Mol Gen Genet 194:15–23

Leutwiler LS, Meyerowitz EM, Tobin EM (1986) Structure and expression of three light-harvesting chlorophyll *a/b*-binding protein genes in *Arabidopsis thaliana*. Nucleic Acids Res 14:4051–4064

Marks MD, West J, Weeks DP (1987) The relatively large *beta*-tubulin gene family of *Arabidopsis* contains a member with an unusual transcribed 5′ noncoding sequence. Plant Mol Biol 10:91–104

McCouch SR, Kochert G, Yu ZH, Wang ZY, Khush GS, Coffman WR, Tanksley SD (1988) Molecular mapping of rice chromosomes. Theor Appl Genet 76:815–829

McGrath JM, Quiros CF (1991) Inheritance of isozyme and RFLP markers in *Brassica campestris* and comparison with *B. oleracea*. Theor Appl Genet 82:668–673

McGrath JM, Quiros CF, Harada JJ, Landry BS (1990) Identification of *Brassica oleracea* monosomic alien chromosome addition lines with molecular markers reveals extensive gene duplication. Mol Gen Genet 223:198–204

McGrath JM, Terzaghi WB, Sridhar P, Cashmore AR, Pichersky E (1992) Sequence of the fourth and fifth Photosystem II Type I (*lhbA*) chlorophll *a/b*-binding protein genes of *Arabidopsis thaliana* and evidence of a full complement of the extended CAB gene family. Plant Mol Biol 19:725–733

Nam HG, Giraudat J, den Boer B, Moonan F, Loos WDB, Hauge BM, Goodman HM (1989) Restriction fragment length polymorphism map of *Arabidopsis thaliana*. Plant Cell 1:699–705

Niyogi KK, Fink GR (1992) Two anthranilate synthase genes in Arabidopsis: defense-related regulation of the tryptophan pathway. Plant Cell 4:721–733

Ohno S (1970) Evolution by gene duplication. Springer, Berlin Heidelberg New York

Oppenheimer DG, Haas N, Silflow CD, Snustad DP (1988) The *beta*-tubulin gene family of *Arabidopsis thaliana*: preferential accumulation of the *beta*-1 transcript in roots. Gene 63:87–102

Pang PP, Pruitt RE, Meyerowitz EM (1988) Molecular cloning, genomic organization, expression and evolution of 12S seed storage protein genes of *Arabidopsis thaliana*. Plant Mol Biol 11:805–820

Peleman J, Saito K, Cottyn B, Engler G, Seurinck J, Van Montagu M, Inzé D (1989) Structure and expression analyses of the S-adenosylmethionine synthetase gene family in *Arabidopsis thaliana*. Gene 84:359–369

Reiter RS, Williams JGK, Feldmann KA, Rafalski JA, Tingey SV, Scolnik PA (1992) Global and local genome mapping in *Arabidopsis thaliana* by using recombinant inbred lines and random amplified polymorphic DNAs. Proc Natl Acad Sci USA 89:1477–1481

Slocum MK, Figdore SS, Kennard WC, Suzuki JY, Osborn TC (1990) Linkage arrangement of restriction fragment length polymorphism loci in *Brassica oleracea*. Theor Appl Genet 80:57–64

Snustad DP, Haas NA, Kopczak SD, Silflow CD (1992) The small genome of Arabidopsis contains at least nine expressed $\beta$-tubulin genes. Plant Cell 4:549–556

Song KM, Suzuki JY, Slocum MK, Williams PH, Osborn TC (1991) A linkage map of *Brassica rapa* (syn. *campestris*) based on restriction fragment length polymorphism loci. Theor Appl Genet 82:296–304

Suiter KA, Wendel JF, Case JS (1983) Linkage-1: a pascal computer program for the detection and analysis of genetic linkage. J Hered 74:203–204

Vallejos CE, Tanksley SD, Bernatzky R (1986) Localization in the tomato genome of DNA restriction fragments containing sequences homologous to the rRNA (45S), the major chlorophyll *a/b*-binding polypeptide and the ribulose bisphosphate carboxylase genes. Genetics 112:93–105