

# A genomic duplication in *Arabidopsis thaliana* contains a sequence similar to the human gene coding for SAP130

Elena Casacuberta<sup>a,b</sup>, Pere Puigdomènech<sup>a</sup>, Amparo Monfort<sup>a,\*</sup>

<sup>a</sup> Departament de Genètica Molecular, Institut de Biologia Molecular de Barcelona, CID-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

<sup>b</sup> Department of Biology, Massachusetts Institute of Technology, Cambridge 02139, USA

Received 23 February 2001; accepted 26 March 2001

**Abstract** – The *Arabidopsis thaliana* gene orthologue to human SAP130 is contained in a perfect genomic duplication of 4.3 kb. The 1 217-aa AtSAP130 protein is 73 % similar to the human gene that is also present in the *Caenorhabditis elegans*, *Drosophila melanogaster* and *Sacharomyces pombe* genomes. The duplicated sequence includes the coding region, two introns, a 32-bp segment in the 5' non-coding region and 420 bp of 3' non-coding region that contains an intron. Sequence homology with ESTs suggests that at least one of the gene copies is transcribed. The presence of the 4.5-kb inverted duplicated sequence has been verified in thirteen ecotypes of *A. thaliana*. Size variation was detected relative to the Columbia ecotype and corresponds to the deletion of a 3' non-coding region that is flanked by a 110-bp tandem repeat. The AtSAP130 duplication appears to be a recent event in the *Arabidopsis* species evolution. © 2001 Éditions scientifiques et médicales Elsevier SAS

*Arabidopsis thaliana* / genomic duplication / spliceosomal associated proteins

aa, amino acids / DDB, DNA-damaged binding factor / EST, expressed sequenced tag / SAP, spliceosomal-associated protein

## 1. INTRODUCTION

*Arabidopsis thaliana*, a member of the crucifer family, has become an important model species to study different aspects of plant biology. Its small genome, 125 Mb, has made it possible to obtain the complete sequence and join, as the first flowering plant genome, the other eukaryotic sequenced genomes of *Sacharomyces cerevisiae*, the nematode model *Caenorhabditis elegans* and *Drosophila melanogaster*. The *Arabidopsis* genome contains 25 498 genes that are spaced on average every 4.5 kb [21, 29]. Approximately 60 % of identified genes matched expressed sequenced tags (EST).

As the *Arabidopsis* genome has been sequenced, several different studies have pointed out the high number of regions that are duplicated [2, 17, 19, 22]. These regions belong either to large segmental duplications or clustered gene families, and whether they

come from a possible ancient tetraploid genome or from independent segmental duplication events is still unclear [29]. This duplication rate is larger than any reported in the *C. elegans* genome [30]. As gene duplications are often accompanied by functional divergence, the identification of such duplicated structures in the plant lineage may help identify new gene functions in plants. In *Arabidopsis* and probably in other plants also, many of the duplications occurred in tandem, generating closely linked gene families that may provide the genome size variation in plants and some of the allelic diversity that is useful in agricultural differentiation.

RNA splicing is an essential cellular process that occurs mainly in the spliceosome, the organelle in which the splicing and excision reactions that remove introns from precursor messenger RNA molecules occur. The spliceosome is formed by five small nuclear RNA molecules (U1, U2, U4, U5, U6) and a protein complex that contains SF3 splicing factors and spliceosomal-associated proteins (SAPs) [7]. Two groups of splicing factors SF3a and SF3b have been

\*Correspondence and reprints: fax +34 932045904.  
E-mail address: amvgmp@ibmb.csic.es (A. Monfort).

cloned in mammals [15], and in yeast it has been shown that the SF3a products are essential for viability. The human SAP49, SAP130, SAP145 and SAP155 correspond to SF3b factors. Das et al. [7] have shown that SAPs130, 145 and 155 of mammals are present in a protein complex essential for splicing.

The organisation, function and regulation of splicing in plants has a number of differential specificities compared to the animal mRNA splicing process [3, 25]. The similarity of plant snRNPs (small nuclear RiboNucleoProteins) with those from vertebrates has been reported, and more specifically the homologues of U6atac and U12snRNA have been found in *A. thaliana* showing a complete conservation of sequences and structures involved in splicing [24]. The splicing factor SF2 detected in Arabidopsis, atSRp30, for instance is 80 % similar to human SF2 [18]. In this work, we show that in *A. thaliana* the gene orthologue to human SAP130 is contained in a perfect genomic duplication of 4.5 kb. This gene is also present as a single copy gene in the *C. elegans*, *D. melanogaster* and the *S. pombe* genomes. The analysis of the duplication in thirteen ecotypes of *Arabidopsis* shows size variation. The recent duplication of the SAP130 gene is discussed.

## 2. RESULTS

### 2.1. A 4.3-kb duplication in BAC T26I12 from chromosome 3

In the course of sequencing the BAC T26I12 from the genomic project of *A. thaliana* chromosome 3 long arm [11], a perfect inverted duplication was detected. The duplication is 4 290 bp long and is separated by a 1 919-bp fragment. The identity between the duplicated sequences is higher than 99.98 %. Two fragments of 4.5 kb were amplified separately with specific primers designed from the flanking regions and sequenced separately by primer walking and the presence of the duplication has been verified with PCR amplifications and sequences in the genomic DNA of the Columbia ecotype (data not shown).

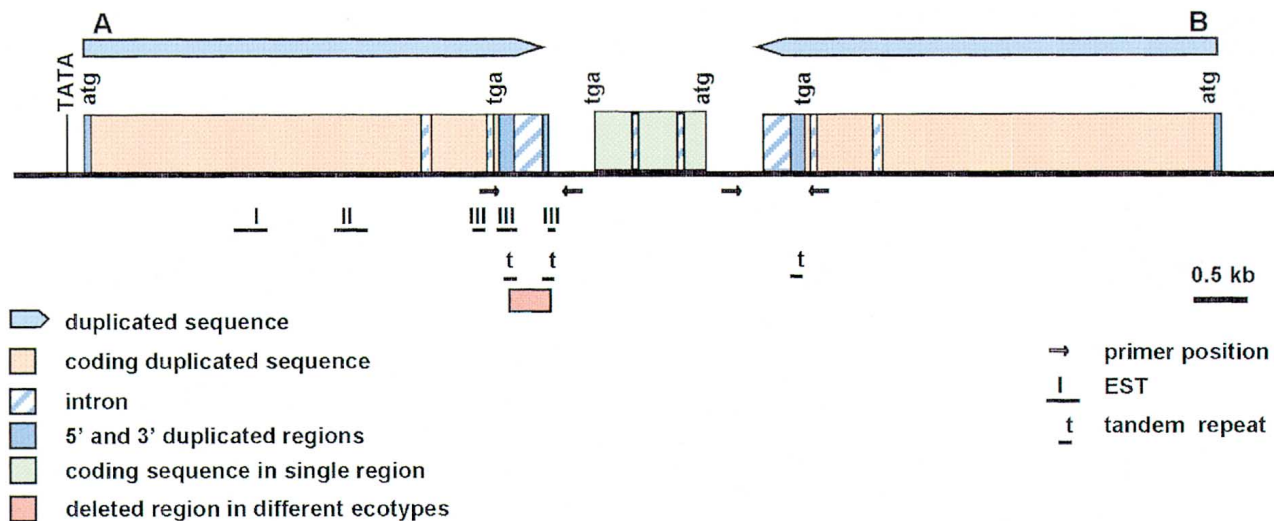
This duplicated region contains a putative coding sequence (1 214 amino acid (aa) residues long), 32 bp of 5'-region, two introns in the coding region, and a fragment of the 3' non-coding region (420 bp), including part of an intron of this 3' non-coding region (figure 1). The 3' non-coding region contains a 110-bp sequence that is repeated 400 bp downstream of one of the copies (figure 1, copy A). The repetition of this sequence is the main difference observed between the

two copies as we have found only two nucleotide differences in the coding region. The first one is located in the codon for residue 557 of the protein (see figure 2), where an A is substituted by a C producing a Ser to Arg change. In the single EST detected in this region by searching the databases (see below), the nucleotide found in this position is A. The second difference is in residue 721 where a G is substituted by a T, but the change is conservative in the amino acid sequence. No EST sequences have been found in this region of the gene. The proximal 5'-region, 32 bp long, is present in the two copies described of the duplicated gene. The 5'-distal region including position -33 to -200 shows 70 % homology between the two copies. The two putative coding regions (figure 1, copies A and B) have been annotated as T26I12\_100 and T26I12\_80 in EMBL Acc. No. AL132954. The sequence of BAC T26I12 is involved in a large duplication region between chromosomes 3 and 2 [29]. However, there are no copies of the AtSAP130 gene in chromosome 2.

### 2.2. Characterisation of AtSAP130

Public databases, both protein and nucleotide, were searched for similar sequences using a 4.3-kb fragment, corresponding to one of the duplicate regions. The Blastx program found significant similarities with a family of spliceosomal proteins. We identified a *Homo sapiens* SAP130 gene coding for a protein of 1 217 aa (CAB56791) that shows a 55 % identity and 73 % similarity with our sequence. The full-length human cDNA was cloned after screening with a peptide product of protein SAP130 digestion [7]. A similar level of identity (48 %) was detected with a gene from *C. elegans* chromosome 3, coding for a 1 220-aa protein (AC 044985), and *D. melanogaster* gene (AE003469), both obtained in the corresponding genomic sequence project. The sequence also shares identity with cDNA sequences matching this region. Moreover, the gene Prp12p/SAP130 of *S. pombe* (BAA86918) coding for a 1 206-aa protein shows 40 % identity and 60 % similarity with the Arabidopsis sequence. This yeast protein has been described as SAP130-like. Only plant ESTs were found to have similarity with the Arabidopsis sequence.

The clustal alignment of the AtSAP130 protein, with the human SAP130, the *D. melanogaster*, the *C. elegans* and Prp12p from *S. pombe* (figure 2) shares an average 50 % identity and 70 % similarity, both at the amino acid level. Moreover in three defined regions, the sequence is highly conserved: between residues 100 to 220, 80 % identity; between residues 425 to 625,



**Figure 1.** Schematic drawing of structural organisation of the duplicated 4.5-kb region of BAC T26I12. Boxes A and B indicate the duplicated AtSAP130 spliceosomal protein sequence. The 110-bp tandem repeat around the deleted region in some *Arabidopsis* ecotypes is marked by t. The central region contains the NAM gene. The roman numerals show the EST sequences: I: N96231; II: AA650848 and T42903; III: AV523424, AV538570, AI992647 and AA395122.

55 % identity; and also the C-terminal region from residue 1137 to the stop codon has 60 % identity.

The Blastn analysis within *A. thaliana* EST sequences has found seven ESTs that perfectly match our region (figure 1) in exon one (Acc. No. N96231, at position 1374 of the coding sequence; Acc. No. T42903 and AA650848 at position 2334 of the coding sequence), and exons two, three and the 3' non-coding duplicated region (Acc. No. AA395122, AV523424, AV538570 and AI992647 at position 3300) (figure 1). No homology was found with other *Arabidopsis* genomic regions. The analysis around other plants show homology with ESTs from tomato (AW031626), cotton AI728111 and *Medicago* (AW694651) with a similarity of sequences around 80 %.

The homology with human, *D. melanogaster*, *C. elegans* and *S. pombe* genes, and *Arabidopsis* ESTs allowed to define the intron and exon boundaries and amino- and carboxy-ends of the protein that is formed by 1 214 aa residues. The distribution of introns is very different between *A. thaliana* (two introns) and *C. elegans* (24 introns) and *D. melanogaster* (six introns). No evidence of a duplication event has been found in the complete genomic sequence of *D. melanogaster* or *C. elegans*.

AtSAP130 also shows similarity (27 %) with a group of proteins, members of a family of UV-damaged DNA repair binding complex that are defective in

xeroderma pigmentosum group E patients. Studies on the human sun-sensitive cancer-prone disease, xeroderma pigmentosum, with various levels of DNA repair deficiency suggest that a multienzyme complex is required for efficient repair in eukaryotes [5]. One of the compounds of DNA repair complex is a DNA-damaged binding factor (DDB). A family of cDNAs (DDB1 and DDB2) similar to this DNA binding factor of monkey cell line are found in human HeLa cells, and have been localised to human chromosome 11 [9]. In this group of proteins, genes from human [1] and monkey [27] are described but only one gene in plants, the *A. thaliana* UV-damaged DNA-binding protein-like (CAA17529) gene that has 27 % similarity with the AtSAP130 duplicated gene. The AtSAP130 region between residues 425 and 625 is conserved with the same region of UV-damaged DNA-binding proteins. DNA polymerases form a large protein family that has been proposed to have evolved after gene duplication during the evolution of eukaryotic and archeal bacteria [10].

### 2.3. Sequence analysis of the 1.9-kb region between the AtSAP130 duplication

A segment of 1.9 kb unique sequence is located between the duplicated sequence. It contains a gene with three exons (figure 1), coding for a 280-aa protein similar to the one encoded by the *Petunia hybrida*

Sequence alignment of protein domains across species: A.thaliana, H.sapiens, D.melanoga, C.elegans, S.pombe. Includes residue numbers (e.g., 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300, 320, 340, 360, 380, 400, 420, 440, 460, 480, 500, 520, 540, 560, 580, 600, 620, 640, 660, 680, 700, 720, 740, 760, 780, 800, 820, 840, 860, 880, 900, 920, 940, 960, 980, 1000, 1020, 1040, 1060, 1080, 1100, 1120, 1140, 1160, 1180, 1200, 1220, 1240, 1260) and residue counts for each species.

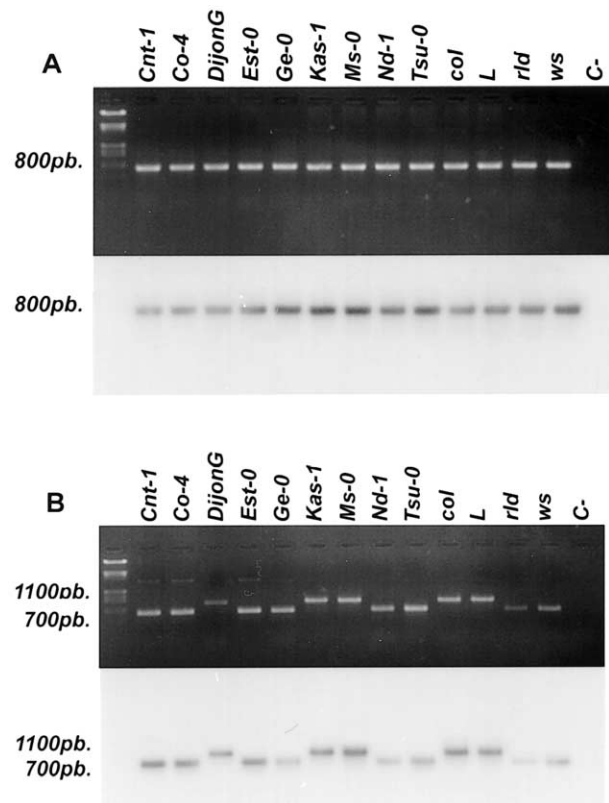
NAM (no apical meristem) gene. The NAM protein is required in *Petunia* for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries [26].

The sequence of BAC T26I12 is involved in a large duplicated region that contains one copy of the NAM gene in chromosome 2 but, as it has previously been mentioned, no AtSAP130 gene. Moreover in the Arabidopsis whole genome, we have found another copy of the NAM-like gene (AB0012243) in chromosome 5, and one corresponding Arabidopsis EST (AA395122).

#### 2.4. Polymorphism within the genomic duplication in thirteen ecotypes of *A. thaliana*

The presence of the AtSAP130 duplication in genomic DNA of *A. thaliana* Columbia ecotype was verified with PCR amplifications of the two copies separately. To detect the presence of this duplication in other ecotypes of Arabidopsis, we amplified the two copies separately from twelve ecotypes of this species. When we compared the 3'-end regions with the same region of gene AtSAP130 sequenced in the Columbia ecotype, we detected some differences in size (figure 3).

With an internal primer of the 3'-region and a primer of single sequence, it is possible to amplify the A and B copies of gene separately. This amplification in several ecotypes produce amplified bands of different sizes (figure 1). The B copy (amplified with primers 9 and 19) produce the same (800 bp) size fragment in all ecotypes (figure 3A). However, copy A (primers 14 and 19) amplification produces a fragment approximately 400 bp shorter than expected in ecotypes Cnt, Co4, Est, Gc-0, Nd-1, Tsu-0, rld and ws in relation to the copy A fragment amplified from *Arabidopsis* Columbia ecotype (1 100 bp) (figure 3B). The sequence obtained of the shorter fragments amplified shows a deleted region (compared to Columbia ecotype) of 378 bp flanked by a 110-bp tandem repeat sequence (figure 4). This short tandem repeat sequence is present once in the 3' non-coding region of AtSAP130 gene in every duplicated sequence, both A and B, and again near the 3' non-coding region of gene A, but not in gene B (figure 1). In the ecotypes Cnt, Co4, Est, Gc-0, Nd-1, Tsu-0, rld and ws, where the sequence between the tandem repeat is deleted (related to Columbia ecotype), the tandem repeat is only present once. This



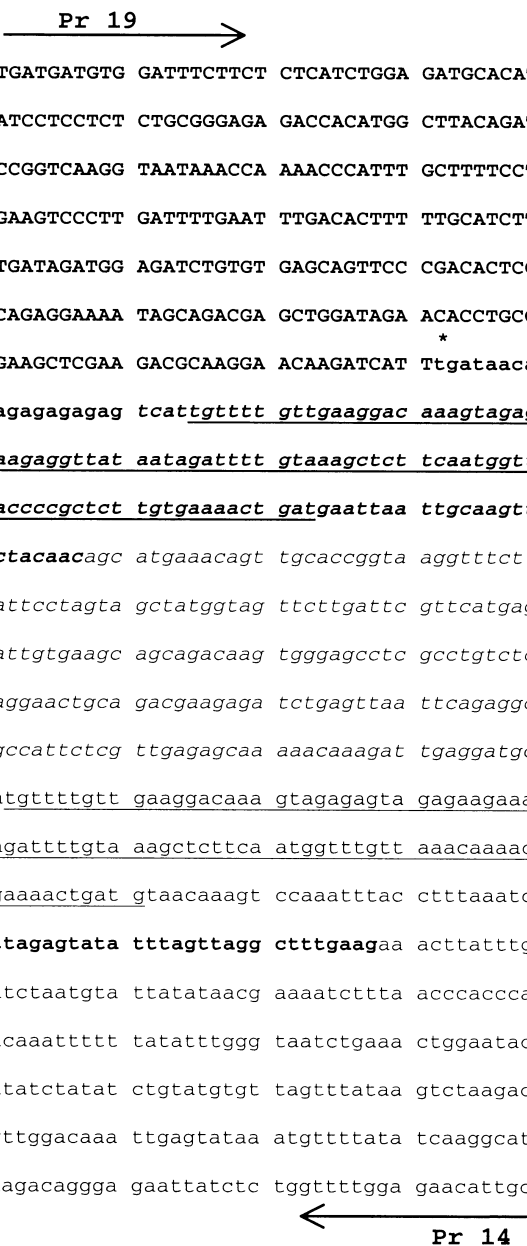
**Figure 3.** PCR amplification in EtBr stained gel of Arabidopsis ecotypes: Cnt-1, Co-4, DijonG, Est-0, Ge-0, Kas-1, Ms-0, Nd-1, Tsu-0, col (Columbia), L (Landsberg erecta), rld, and ws, C (PCR control). Autoradiograph of 15 min exposition of the nylon membrane hybridised at 65 °C with a genomic specific probe of BAC T26I12. **A**, PCR amplification with primers 9–19 corresponding to gene B (figure 1). **B**, PCR amplification with primers 14–19 corresponding to gene A (figure 1).

deleted sequence contains part of the 3'-end of the gene present in the EST AA395122 (figure 4).

### 3. DISCUSSION

A perfect genomic duplication of 4.5 kb that contains a coding sequence (AtSAP130) similar to human spliceosomal-associated protein SAP130 has been found in BAC T26I12 that is positioned in a non-centromeric or paracentromeric and non-telomeric region of the long arm of chromosome 3 of *Arabidop*

**Figure 2.** CLUSTAL alignment of *A. thaliana* AtSAP130 (T26I12\_100), human spliceosomal protein SAP130 (CAB56791), *D. melanogaster* (AE003469), *C. elegans* K02F2.3 (O44985) and *S. pombe* Prp12p (BAA86918). Both identical and similar amino acids in the five proteins are boxed in black and when only in three or four proteins are boxed in grey.



1 **TGATGATGTG GATTCCTTCT CTCATCTGGA GATGCACATG AGGCAGGAGT**  
 51 **ATCCTCCTCT CTGCGGGAGA GACCACATGG CTTACAGATC TGCATATTTT**  
 101 **CCGGTCAAGG TAATAAACCA AAACCCATTT GCTTTTCCTC CAAATCAAAA**  
 151 **GAAGTCCCTT GATTTTGAAT TTGACACTTT TTGCATCTTG TAACAGGACG**  
 201 **TGATAGATGG AGATCTGTGT GAGCAGTTCC CGACACTCCC AATGGACTTG**  
 251 **CAGAGGAAAA TAGCAGACGA GCTGGATAGA ACACCTGCGG AGATTCTGAA**  
 301 **GAAGCTCGAA GACGCAAGGA ACAAGATCAT Ttgataaacac acgaaaagag**  
 351 **agagagagag tcattgtttt gttgaaggac aaagtagaga gtagagaaga**  
 401 **aagaggttat aatagatttt gtaaagctct tcaatggttt gttaaacaaa**  
 451 **accccgctct tgtgaaaact gatgaattaa ttgcaagttt acttttgaat**  
 501 ***ctacaacagc atgaaacagt tgcaccggta aggtttcttc catattcagg***  
 551 ***attcctagta gctatggtag ttcttgattc gttcatgaga agcattgaga***  
 601 ***attgtgaagc agcagacaag tgggagcctc gcctgtctca acaaaagatc***  
 651 ***aggaactgca gacgaagaga tctgagttaa ttcagaggct cttaacacca***  
 701 ***gccattctcg ttgagagcaa aaacaaagat tgaggatgca ggtgagctaa***  
 751 **atgttttggt gaaggacaaa gtagagagta gagaagaaag aggttataat**  
 801 **agattttgta aagctcttca atggtttggt aaacaaaacc cgcctcttgt**  
 851 **gaaaactgat gtaacaaagt ccaaatttac ctttaaatct acaactttgt**  
 901 ****ttagagtata tttagttagg ctttgaagaa** acttatttgt tggttttaag**  
 951 atctaagtta ttatataacg aaaatcttta acccaccac aaatcaaac  
 1001 acaaattttt tatatttggg taatctgaaa ctggaataca cgggtgtttt  
 1051 atatctatat ctgatgtgt tagtttataa gtctaagacg ctaagaaaat  
 1101 gttggacaaa ttgagtataa atgttttata tcaaggcatc aaaagcaata  
 1151 aagacagga gaattatctc tggttttgga gaacattgc

**Figure 4.** Sequence of the 3'-end region of the gene amplified with primers 14–19 from the Columbia ecotype, that correspond to the A gene. Arrows indicate the primer position. Bold letters correspond to the EST homologue sequence, capital bold letters correspond to the coding sequence, the stop codon is marked by an asterisk, and small bold letters correspond to 3' non-coding sequence present in the EST. Underlined letters are the tandem repeat sequences, and italic letters correspond to the deleted sequence in some ecotypes. Normal letters correspond to single sequence.

*sis thaliana*. The human SAP130 is part of a protein complex with SAP49, SAP145 and SAP155, that within U2snRNP, is present in nuclear extracts as a component of the spliceosomal protein complex SF3b [7]. The function of this spliceosome is to splice a rare class of introns [28]. Another duplicated gene related to splicing is a U3RNA small nucleolar RNA protein essential for growth in yeast and is required for very early pre-RNA processing events in yeast and in

mammals. In human chromosome 17p11.2, there are three U3 genes, with two located within an inverted duplication of nearly 45 kb. The U3 genes also share significant upstream and downstream similarity suggesting a recent evolutionary origin [12].

Little is known about the splicing mechanism in plants. Differences exist in the process of intron recognition since animal introns are not processed in plant tissues [4], but generally, it is assumed that the

basic mechanisms are similar to yeast and human splicing [16]. In the Arabidopsis genome, sequences coding for the SF3a splicing factor, U2B and two spliceosomal-associated proteins, SAP49 and SAP62, have been found by similarity. The splicing factor atSRp30 is 80 % similar to human SF2/ASF splicing factor, indicating the high degree of conservation of these gene sequences [18]. Similar identity has been demonstrated for the helicase family between *C. elegans* and human genomes [31]. The high similarity (70 %) of SAPs of *Arabidopsis*, human, *C. elegans*, *D. melanogaster* and *S. pombe* SAP130 genes demonstrate that this is also a highly conserved family of proteins in eukaryotic genomes. The homology between a group of DNA repair polymerases, DDBs, and AtSAP130 has also been detected. However, no DDB protein has been found in the human spliceosomal complex [7], suggesting a different function for the human DDB and SAP130 genes.

The gene copy AtSAP130A has a putative TATA box and it matches exactly with EST sequences, while the AtSAP130B has one nucleotide difference with the EST known sequence. These data would indicate that the AtSAP130A is a transcribed copy, but it does not exclude the possibility that both genes A and B are transcribed under different conditions. The region upstream of the 32-bp 5'-duplicated sequence contains a number of conserved short sequences. In any case, *Arabidopsis* is the first genome where this gene (SAP130) is duplicated, having been found only once in human, *D. melanogaster*, *C. elegans* and yeast genomes, indicating that it is not part of a multigene family but an isolated genomic duplication event. It is possible that this recent duplication has not included most of the regulatory sequences and therefore only one of the genes (copy A) is effectively transcribed. The upstream and downstream regions of the genes do not show any repetitive structure that would indicate putative recombination regions that would have resulted in this inverted duplication. The AtSAP130 duplication appears to be a recent event in evolution, since the DNA sequence is highly conserved.

In spite of this recent event, the duplication happened before the definition of the different ecotypes. Further variability appears after this duplication as some of the ecotypes studied present a 3' non-coding deletion that make it possible to distinguish the two copies of gene. A similar high level of conservation in the *A. thaliana* genome is shown by the family of PAI tryptophan biosynthetic genes present in four copies and suggesting that the amplification of this gene family has been a recent event in the evolution of the

species [20]. The DNA sequence is highly conserved among the different members of these two families, AtSAP130 and PAI genes. Differences and rearrangements are found in both genes when a study in a broad number of ecotypes is performed. These findings are in agreement with the theory that Arabidopsis is a species in expansion and of recent evolution [14].

This region is involved in a large duplicated region between chromosomes 2 and 3. The almost perfect inverted duplication (> 99.98 %) in a proximal region (< 2 kb) indicates that it is a recent success of a duplication in the genome. The inverted duplicate sequence containing AtSAP130 is separated by a fragment of 1.9 kb that contains an AtNAM-like (no-apical-meristem) gene. It should also be noted that the AtNAM-like gene and the regions adjacent to the AtSAP130 genes are present in the duplicated region in chromosome 2. The presence of the gene and its duplication in this region of chromosome 3, or its absence in chromosome 2, may be more recent than the large genomic duplication occurring in Arabidopsis and it might indicate that individual genes may jump between different genomic regions of the plant.

The inverted duplicated sequence has a different size in different ecotypes of *A. thaliana* caused by a 400-bp deletion or insertion in the 3' non-coding region. The deletion is flanked by a 110-bp tandem repeat. In the case of AtSAP130A, only one copy of this tandem repeat is present in some ecotypes that present the deletion. The gene AtSAP130B has only one copy of the 110-bp tandem repeat in all ecotypes analysed. The deleted region approximately corresponds to the intron present in the 3' non-coding region. The deletion of the sequence flanked by the 110-bp tandem repeat suggests the possibility that this short sequence repeated in tandem could make a secondary structure or a false match that resulted in a deletion by a recombination, a situation often found between short repetitive sequences. Recombination is an important mechanism in evolution helping to create genetic variation in sexually reproductive organisms [6].

The nuclear genome of *A. thaliana* has an average 13 % of genes which are members of multigene families. Gene duplications can occur on a gene by gene basis, resulting in tandem repeats, or they can result from the duplication of larger chromosomal regions. As new genomic data is released and analysed, duplication regions, insertions, deletions and tandem duplications are found suggesting the role that these events could have played in the evolution of this genome [2]. A study of duplications in *C. elegans*

genes shows that about 40 % of genes analysed are involved in duplicated gene pairs [23], and 13 % of the gene pairs were separated by fewer than five genes. In Arabidopsis, 65 % of genes are present in more than one copy, and 17 % are present in tandem arrays with one unrelated gene as maximum among cluster members [29]. The large duplication regions in the Arabidopsis genome show a different copy number of genes, in each region, sometimes the gene is duplicated in tandem after region duplication, or it is lost. While AtNAM-like is duplicated once in each region, the AtSAP130 is absent in chromosome 2 but almost perfectly duplicated in chromosome 3 at the DNA (4.3 kb) and protein (1 214 aa) level. Gene duplications play a key role in genome evolution [13], but it is not known in what proportion gene by gene duplication or large genomic duplications contribute to this evolution. AtSAP130a and b, may be a good example of the mechanisms used by organisms with small genomes to generate different functions or different types of gene regulation.

#### 4. METHODS

The BAC clone T26112 (included in ESSAIII project, Arabidopsis genomic sequence chromosome 3) was sequenced after shotgun subcloning, and the presence of the duplicated region sequence was verified by primer walking on BAC DNA. The DNA sequence was assembled by the gap4 program from the STADEN package. Similarities to the EMBL, SWISS-PROT, and Arabidopsis ESTs were calculated using the BLAST algorithm. The protein alignment was made with the CLUSTALW program. DNA fragments were amplified by PCR from thirteen different *Arabidopsis thaliana* ecotypes, Columbia, Landsberg erecta, rld, ws, Co-4, Cnt-1, Est-0, Ge-0, Nd-1, Tsu-0, DijonG, Ms-0, Kas-1. Seeds of each ecotype were grown at 22 °C in a greenhouse with 8 h of light. Genomic DNAs were extracted using the protocol of Dellaporta et al. [8]. PCR amplifications were performed using primers 14–19 and 9–19 (see sequences below) with: 20 ng genomic DNA, 20 pmol each primer, 2 mM MgCl<sub>2</sub>, 0.2 mM each dNTP, in a total volume of 50 µL. The annealing temperature was 50 °C and extension time 90 s. The PCR products were visualised in 1 % agarose gels and transferred to a nylon membrane (Schleicher and Schuell) by standard procedures. The membrane was hybridised with a genomic specific probe of BAC T26112 and washed at high stringency (20 mM Na<sub>2</sub>PO<sub>4</sub>, pH 7.2, 1 % SDS, 1 mM EDTA at 65 °C).

Primer sequences:

Primer 9: 5'-GCAGGTACTTTTTTAAGTC-3'

Primer 19: 5'-GTGATGATGTGGATTCTTC-3'

Primer 14: 5'-GCAATGTTCTCCAAAACC-3'

**Acknowledgments.** The present work has been funded by grants from the European Union (Arabidopsis BIO4-CT98-0549 'EU Sequencing on Arabidopsis chromosome 3') and CICYT (grant BIO98-1838-CE). The work has been carried out within the framework of 'Centre de Referència de Biotecnologia de la Generalitat de Catalunya'.

#### REFERENCES

- [1] Alexander H., Lee S.K., Yu S.L., Alexander S., repE - the Dictyostelium homolog of the human xeroderma pigmentosum group E gene is developmentally regulated and contains a leucine zipper motif, *Nucleic Acids Res.* 24 (1996) 2295–2301.
- [2] Blanc G., Barakat A., Guyot R., Cooke R., Delseny M., Extensive duplication and reshuffling in the Arabidopsis genome, *Plant Cell* 12 (2000) 1093–1101.
- [3] Brown J.W., Arabidopsis intron mutations and pre-mRNA splicing, *Plant J.* 10 (1996) 771–780.
- [4] Brown J.W., Simpson C.G., Splice site selection in plant-mRNA splicing, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 49 (1998) 77–95.
- [5] Cleaver J.E., Kraemer K.H., Scriver C.R., Beaudet A.L., Sly W.S., Valle D. (Eds.), *The Metabolic Basis of Inherited Disease*, vol. 2, McGraw-Hill, New York, 1989, pp. 2949–2971.
- [6] Crow J.F., Kimura M., Evolution in sexual and asexual populations, *Am. Nat.* 99 (1965) 439–450.
- [7] Das B.K., Xia L., Palandjian L., Gozani O., Chyung Y., Reed R., Characterization of a protein complex containing spliceosomal proteins SAPs 49, 130, 145, and 155, *Mol. Cell. Biol.* 19 (1999) 6796–6802.
- [8] Dellaporta S.L., Wood J., Hicks J.B., *Molecular Biology of Plants: A Laboratory Course Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1984.
- [9] Dualan R., Brody T., Keeney S., Nichols A.F., Admon A., Linn S., Chromosomal localization and cDNA cloning of the genes (DDB1 and DDB2) for the p127 and p48 subunits of a human damage-specific DNA binding protein, *Genomics* 29 (1995) 62–69.
- [10] Edgell D.R., Malik S.B., Doolittle W.F., Evidence of independent gene duplications during the evolution of archaeal and eukaryotic family B DNA polymerases, *Mol. Biol. Evol.* 15 (1998) 1207–1217.
- [11] European Union Chromosome 3 Arabidopsis Sequencing Consortium, The Institute for Genomic Research & Kazusa DNA Research Institute, Sequence and analysis



- sis of chromosome 3 of the plant *Arabidopsis thaliana*, *Nature* 408 (2000) 820–822.
- [12] Gao L., Frey M.R., Matera A.G., Human genes encoding U3 snRNA associate with coiled bodies in interphase cells and are clustered on chromosome 17p11.2 in a complex inverted repeat structure, *Nucleic Acids Res.* 25 (1997) 4740–4747.
- [13] Gogarten J.P., Olendzenski L., Orthologs, paralogs and genome comparisons, *Curr. Opin. Genet. Dev.* 9 (1999) 630–636.
- [14] Hardtke C.S., Muller J., Berleth T., Genetic similarity among *Arabidopsis thaliana* ecotypes estimated by DNA sequence comparison, *Plant Mol. Biol.* 32 (1996) 915–922.
- [15] Kramer A., The structure and function of proteins involved in mammalian pre-mRNA splicing, *Annu. Rev. Biochem.* 65 (1996) 367–409.
- [16] Lazar G., Schaal T., Maniatis T., Goodman H.M., Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF, *Proc. Natl. Acad. Sci. USA* 92 (1995) 7672–7676.
- [17] Lin X., Kaul S., Rounsley S., Shea T.P., Benito M.I., Town C.D., et al., Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*, *Nature* 402 (1999) 761–768.
- [18] Lopato S., Kalyna M., Dorner S., Kobayashi R., Krainer A.R., Barta A., atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes, *Genes Dev.* 13 (1999) 987–1001.
- [19] Mayer K., Schuller C., Wambutt R., Murphy G., Volckaert G., Pohl T., et al., Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*, *Nature* 402 (1999) 769–777.
- [20] Melquist S., Luff B., Bender J., *Arabidopsis PAI* gene arrangements, cytosine methylation and expression, *Genetics* 153 (1999) 401–413.
- [21] Meyerowitz E.M., Today we have naming of parts, *Nature* 402 (1999) 731–732.
- [22] Paterson A.H., Bowers J.E., Burow M.D., Draye X., Elvik C., et al., Comparative genomics of plant chromosomes, *Plant Cell* 12 (2000) 1523–1539.
- [23] Semple C., Wolfe K.H., Gene duplication and gene conversion in the *Caenorhabditis elegans* genome, *J. Mol. Evol.* 48 (1999) 555–564.
- [24] Shukla G.C., Padgett R.A., Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants, *RNA* 5 (1999) 525–538.
- [25] Simpson G.G., Filipowicz W., Splicing of precursors to mRNA in higher plants: mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery, *Plant Mol. Biol.* 32 (1996) 1–41.
- [26] Souer E., van Houwelingen A., Kloos D., Mol J., Koes R., The no apical meristem gene of *Petunia* is required for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries, *Cell* 85 (1996) 159–170.
- [27] Takao M., Abramic M., Moos M. Jr, Otrin V.R., Wootton J.C., McLenigan M., Levine A.S., Protic M., A 127 kDa component of a UV-damaged DNA-binding complex, which is defective in some xeroderma pigmentosum group E patients, is homologous to a slime mold protein, *Nucleic Acids Res.* 21 (1993) 4111–4118.
- [28] Tarn W.Y., Steitz J.A., Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge, *Trends Biochem. Sci.* 22 (1997) 132–137.
- [29] The *Arabidopsis* genome initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature* 408 (2000) 796–815.
- [30] The *C. elegans* sequencing consortium, Genome sequence of the nematode *C. elegans*: A platform for investigating biology, *Science* 282 (1998) 2012–2018.
- [31] Villard L., Fontes M., Ewbank J.J., Characterization of xnp-1, a *Caenorhabditis elegans* gene similar to the human XNP/ATR-X gene, *Gene* 236 (1999) 13–19.