Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the *APETALA2* locus on chromosome 4

Nancy Terryn^a, Leo Heijnen^b, Annick De Keyser^a, Martien Van Asseldonck^{1,b}, Rebecca De Clercq^a, Henk Verbakel^b, Jan Gielen^a, Marc Zabeau^b, Raimundo Villarroel^a, Taco Jesse^b, Pia Neyt^a, René Hogers^b, Hilde Van Den Daele^a, Wilson Ardiles^a, Christine Schueller^c, Klaus Mayer^c, Patrice Déhais^a, Stephane Rombauts^a, Marc Van Montagu^a, Pierre Rouzé^{a,d,*}, Pieter Vos^b

^aLaboratorium voor Genetica, Departement Genetica, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Universiteit Gent, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium

^bKeygene N.V., Postbus 216, AgroBusiness Park 90, Keyenbergseweg 6, NL-6700 AE Wageningen, The Netherlands ^cMunich Information Center for Protein Sequence, Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany ^dLaboratoire Associé de l'Institut National de la Recherche Agronomique (France), Universiteit Gent, B-9000 Ghent, Belgium

Received 9 January 1999

Abstract As part of the European Scientists Sequencing *Arabidopsis* program, a contiguous region (396 607 bp) located on chromosome 4 around the *APETALA2* gene was sequenced. Analysis of the sequence and comparison to public databases predicts 103 genes in this area, which represents a gene density of one gene per 3.85 kb. Almost half of the genes show no significant homology to known database entries. In addition, the first 45 kb of the contig, which covers 11 genes, is similar to a region on chromosome 2, as far as coding sequences are concerned. This observation indicates that ancient duplications of large pieces of DNA have occurred in *Arabidopsis*.

© 1999 Federation of European Biochemical Societies.

Key words: Arabidopsis thaliana; APETALA2; Genome; Sequencing

1. Introduction

In plant molecular biology and genetics, *Arabidopsis thali-* ana has long been recognized as a model organism [1]. By the end of 1993, a project designated European Scientists Sequencing *Arabidopsis* (ESSA) was initiated with the aim of sequencing large fragments of the *A. thaliana* genome. Currently, a whole international team is working on the completion of that genome [2].

*Corresponding author. Fax: (32) (9) 264 5349.

E-mail: pirou@gengenp.rug.ac.be

¹Present address: University Hospital Nijmegen, Department of Human Genetics, Geert Grootteplein 24, NL-6500 HB Nijmegen, The Netherlands.

Abbreviations: AFLP, amplified fragment length polymorphism (is a trademark in the Benelux); BAC, bacterial artificial chromosome; EST, expressed sequence tag; YAC, yeast artificial chromosome

The genomic sequence of the contig has been submitted to the GenBank under accession numbers Z99707 and Z99708 (with 8361-bp overlap). Cognate cDNAs have accession numbers AJ002596, AJ002597, and AJ002598.

Within the ESSA program, most effort was concentrated initially on chromosome 4 around the *FCA* locus [3] and the *APETALA2* (*AP2*) locus. The *ap2* mutant had been described, mapped, and cloned before [4–6]. On the most recent map of Dean and Lister [7], *AP2* was found between the markers m214 and g2486 at 93.2 cM (the whole chromosome 4 being 116 cM).

Results of the sequencing and analysis of this latter region, as the result of cooperation between two laboratories, will be discussed here. This region represents the second largest contig of *Arabidopsis* being analyzed in detail, preceded by the 2 Mb *FCA* contig [3].

2. Materials and methods

2.1. Isolation and subcloning of an AP2-containing yeast artificial chromosome (YAC)

The CIC YAC clone 7A10, size 420 kb (from the CIC A. thaliana (L.) Heynh. ecotype Columbia library) was isolated by using AFLP (filed by Keygene N.V.) markers [8,9] derived from sequences of the AP2 gene and subcloned in the cosmid vector pCLD04541 [10], using a partial Sau3AI digest. To isolate YAC-specific cosmids, a total number of 16000 Escherichia coli clones were hybridized to gel-purified YAC DNA. Approximately 450 YAC-specific clones were identified. AFLP fingerprinting [9] was used to build a cosmid contig. This approach resulted in a cosmid contig of approximately 260 kb.

2.2. Isolation of bacterial artifical chromosome (BAC) clones extending the cosmid contig

The BAC clones were isolated through hybridization of the BAC library with a probe containing the inserts of all cosmids of the 260-kb contig, followed by contig building by AFLP. All AFLP markers present in the cosmid contig were also present in the BAC contig, indirectly proving colinearity of the cosmid contig with the genomic sequence. TAMU 8H13 and TAMU 10C14, which overlapped with the cosmid contig, were selected for further sequencing (Fig. 1).

2.3. Construction of cosmid and BAC subclones

DNA from cosmids and BAC clones was sheared either by sonication (Misonix Inc., Farmingdale, NY; type XL2020) or by nebulizing (Lifecare Hospital Supplies, Harborough, UK). A 1.8–2.2-kb end-repaired fraction was isolated and ligated in pUC18/Smal/BAP (Pharmacia, Uppsala, Sweden) or a modified pUC19 vector (BamHI/Sall fragment replaced by a StuI-, SpeI-, and SalI-containing fragment). Individual colonies of the transformation were grown in 96- or 384-well microtiter plates.

0014-5793/99/\$19.00 $\ensuremath{\mathbb{C}}$ 1999 Federation of European Biochemical Societies. All rights reserved.

PII: S0014-5793(99)00097-6

2.4. Sequencing strategy

The sequence of the partially overlapping cosmids 3A6, 4B6, 4E12, 5C9, 2H2, 5E2, 2F3, 2H7, and BAC TAMU 10C14 was determined in a non-random approach by sequencing the cosmid or BAC ends, as well as a few random subclones. Primers were designed to isolate primer-specific clones from pools of subclones. A random sequencing approach was taken for cosmids 2C7, 3A6, 4F4, 3F11, 6B5, 5E3, 4G12, 4E3, and BAC TAMU 13H8. The ABI PRISM dye terminator cycle sequencing ready reaction kit was used mostly (Perkin-Elmer, Foster City, CA). The reaction products were analyzed on an ABI Prism 377. The sequence with the corresponding electropherograms were assembled into contigs using a home-made computer program called Sequence Assembly Facility Environment (SAFE; [11]) or the 1994 version of the Staden sequence analysis program [12]. The mean redundancy of the assembled sequence is 5.

2.5. Sequence comparisons

Sequence analysis was done initially by the Martinsried Institute for Protein Sequences (MIPS) center (Martinsried, Germany) and refined by the bioinformatics team of the Laboratory of Genetics (Ghent, Belgium). To this end, the sequence was cut into pieces of 7000 bp with 1000-bp overlaps. Each piece was submitted to a BLASTX/non-redundant protein and BLASTN/non-redundant DNA search as well as a BLASTN/EST search (mostly on the Beauty BCM server). From the resulting files, the homologues with the highest score and a reliable annotation were looked for.

When a homologue was found with a high score (P(N) < E-100) and with homology over its whole length, its protein sequence was followed to check whether the transcript did not show any frameshifts or stop codons and whether the intron borders were correct. When any doubt arose, NetGene2 predictions were used [13,14].

For genes with weak homologies, different prediction programs were used according to the specific problems: NetStart when the start codon was not obvious, GeneMark [15] together with GenScan [16] for exon predictions, GeneMark for exon frame predictions, GenScan to check that small exons were not missed, and NetGene2 for intron border prediction. When an expressed sequence tag (EST) was found through BLASTN/EST (most of the time for the 5' or 3' parts of the genes), the EST was used to complement the information obtained by BLASTX.

When no homologues were found through BLASTX, and no ESTs were available, we relied upon prediction programs, namely GenScan and GeneMark together to locate potential exons, NetGene2 for the

intron borders, and NetStart [17] for the position of the start codon. In all cases, the coding sequence (CDS) was reconstructed, translated, and submitted to BLASTP, for a last homology search and check for gaps.

Updated FASTA searches done at MIPS on the genes in this contig can be found at http://speedy.mips.biochem.mpg.de/arbi/data/ap2_contig.html. Full annotations as well as various genomic features can also be found at the Ghent site (http://spider.rug.ac.be/public/seq/ap-2.html).

3. Results and discussion

3.1. General overview of the contig

Fig. 1 gives an overview of the clones used for the sequencing program. Initially cosmid clones were used, but during the project BAC clones became available and these were used to continue the contig sequencing (see Section 2).

The 103 predicted genes that are present in this region are summarized in Tables 1 and 2 and in Fig. 2. Table 1 provides information on their putative function, highest related entry in the public databases as well as EST sequences that correspond to the putative genes whereas in Table 2 the genes are classified based on their homology.

In conclusion, the putative role of approximately half of the genes could be established by sequence similarity to known genes. These genes have been classified into 15 classes according to their putative cellular role that will be used to describe genes identified in the genome program [3]. This list is preliminary and new categories and subcategories will be added as more of the genome is sequenced (updated version available at http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html).

3.2. Annotation and re-annotation process

After the first MIPS automatic annotation, manual re-annotation showed discrepancies on about four genes out of

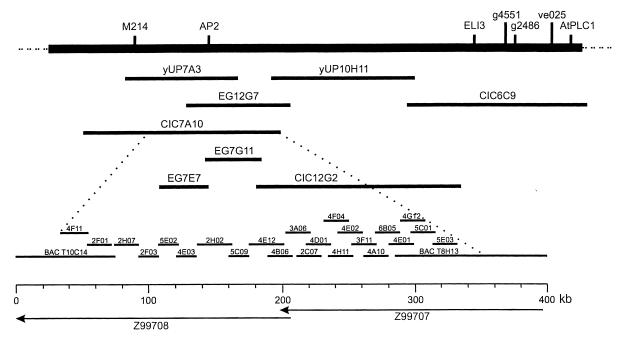


Fig. 1. Chromosomal location of the *AP2* gene. Overview of the cosmids and BACs used for sequencing. The upper line represents the chromosomal region (approximately 10 cM) with some known markers in the region. The sequenced cosmids and BAC clone are indicated as well as the region covered by the two database submissions covering this region (Z99707 and Z99708). The sequence Z99707 (206 440 bp) contains 50 genes (1–50) with an 8361-bp overlap with Z99708 (198 555 bp) that contains 53 genes (51–103).

W. Cytochrome P490 97,2460 Oytochrome P490 97,2460 Oytochrome P490 97,2460 Oytochrome P490 98,2311 P24466 Oytochrome P490 98,2311 P24466 Oytochrome P490 18,5417 P24466 Oytochrome P490 98,2311 P24466 Oytochrome P490 98,2311 P24466 Oytochrome P490 98,2311 P24466 Oytochrome P490 P34466 Oytochrome P490 98,247,748 P24466 Oytochrome P490 P34461 P24466 Oytochrome P490 P34461 P24466 Oytochrome P490 P34461 P34461 </th <th></th> <th>\mathbf{S}_{a}</th> <th>Function/ product</th> <th>Position</th> <th>Best homologue accession</th> <th>Name</th> <th>Species</th> <th>BLASTP score</th> <th>EST</th> <th>BLASTN/EST score</th> <th>Class^b</th> <th>Category^c</th>		\mathbf{S}_{a}	Function/ product	Position	Best homologue accession	Name	Species	BLASTP score	EST	BLASTN/EST score	Class ^b	Category ^c
Cyclochrome P450 Sylvestade Cyclochrome P450 avocado 1 6E-21 1745440 5 9E-47 2 Cyclochrome P450 cyclochrome P450 avocado 1 6E-21 1745440 5 9E-47 5 Debate	1	≥≥≥	Cytochrome P450 Cytochrome P450 Cytochrome P450	<1-560 978-2731 3115-5137	P24465 P24465 P24465	cytochrome P450 cytochrome P450 cvtochrome P450	avocado avocado avocado	1.6E-25 7.6E-60 1.5E-67	H76015 H76015 AA395149	2.1E-76 3.6E-65 1.3E-149	w 0 0	12 12 12 12
Unknown 13 24 2 20 755 Cult. monown I 754 1 755 AC002391 unknown I 754 1 755 AC002391 Unknown I 754 2 37 73 AC002391 AC002392 CLI-transporting nouse 3.3 E-14 ACC	•	≥≥∪		5994–7943 8851–11532 12186–12879	P24465 P24465	cytochrome P450 cytochrome P450	avocado	1.6E-21 1.0E-68	T43640 F13573 N38505	5.9E-47 6.4E-104 9.9E-131	1 K 7 K	1221
Cuinton-transporting 21332-2697 P03356 Cu-transporting mouse 33E-14 Z33731 2.0E-104 3 AfP-schladed 26 120-2782 Z54156 Meh-related ATP-ase T4321 5.5E-3 2 Meh-related 26 120-2782 Z54156 Meh-related ATB-ase Meh-related ATB-ase T4E-60 T4321 5.5E-3 2 Kinnes-lise receptor 21 610-2780 S04132 potential problems of the complex coloring ATB-ase Meh-related 4.5E-50 H6663 4.6E-34 5 Cold acclimation 5 1715-2257 U73216 cold acclimation T717ctcum activum 4.4E-56 AA65647 1.3E-35 3 Cold acclimation 5 1715-2257 U73216 cold acclimation T717ctcum claved from the cold acclimation T717ctcum AA66647 1.3E-35 2 Putative histone binding 53662-35 Mg A25680 nanchage protein T717ctcum claved from the cold acclimation T717ctcum claved from the cold acclimation AA66647 1.3E-35 2 Putation binding 10 cold acclimation		≥ ≥	Unknown	17 504–17 755 18 322–20 936	AC002391	unknown	Arabidopsis		D15421	5.7E-22	4 9	13
Wish-related 57 (2) 2.70 2. 2.541.5 Mish-related Arobidoposes 1426-69 14311 5.55.3 2.53.3 2.54.3 Wish-related Arobidopose 4.55.5 147665 4.65.3 2.53.3 2.54.3 2.54.3 2.54.3 2.54.3 2.54.3 2.54.3 3.54.3		≽	Cation-transporting ATP-ase	21 332–25 697	P30336	Cu-transporting ATPase	mouse	3.3E-14	Z33731	2.0E-104	3	7,22
Pseudo-gene 50097-51 500 S04132 photosystem II pea Disjli60 8.2E.22 3 Cold acclimation 51715-25.58 U.73216 confide acclimation Triticum accrition 4.4E.56 AA650647 1.3E.35 2 Cold acclimation 51715-25.78 V.73216 confide acclimation Triticum accrition 4.4E.56 AA650647 1.3E.35 2 Putative bistone binding 53 662-55.308 A2580 nuclear histone African clawed frog 3.3E.97 AA59462 1.9E.130 3 Putative cell division 58 172-60.259 U/80043 misato gene African clawed frog 3.3E.15 AA59462 1.9E.130 3 Unknown 65 821-66 868 K00193 Glu-tRNA Drosophila melano- 3.3E.15 AFE-110 3 Cilcu-tRNA 66 821-66 868 K00193 Glu-tRNA Drosophila melano- 3.3E-15 AFE-110 3 Pectitissterase 7 860-75 864 294058 pectinesterase tomato 1.6E-123 AA59456 1.9E-130 3	. , •	ບ ≱ ເ	Myb-related Kinase-like receptor	26 120–27 082 37 499–39 895 42 632–43 066	Z54136 P33543	Myb-related receptor kinase TMK1		4.2E-69 4.5E-50	T43211 H76836 H76663	5.5E-35 1.4E-135 4 6F-34	0 m v	4.19.04 10,01
Cold acclimation 51715-2578 U73216 coll acclimation Triticum accitium 445-56 AA650647 13E-35 2 Protein Puttive histone binding S3 662-55 308 A25680 nuclear histone African clawed frog 3.3E-07 AA594562 1.9E-150 3 Thioredoxin S644-57848 P73920 Unknown C1570-60529 US0043 missto gene Drosophila melano- 3.5E-15 R64911 5.4E-110 3 Putnive cell division 61 577-62 335 C10-4RNA Drosophila melano- 3.5E-15 R64911 5.4E-110 3 Clat-RNA 6821-6868 K 00193 Glu-tRNA Drosophila melano- 3.5E-15 R64911 5.4E-110 3 Unknown 6821-6868 K 00193 Glu-tRNA Drosophila melano- 1.5E-123 H76007 1.3E-10 3 Pyctrosynitric byse 6821-6868 K 00193 Glu-tRNA Drosophila melano- 1.5E-123 H76007 1.3E-106 3 Pyctrosynitric byse 7352-74499 Z 24271 pir7a pir7a	-	C	Pseudo-gene	50 097-51 300	S04132	photosystem II oxygen-evolving	pea		D49160 RICE	8.2E-22	, w	2,2
Putative bistone binding \$33 662-553 08 A25680 nuclear histone African clawed frog 3.3E-07 A39-07 Appropriate inter-bistone A4F-23 AA394562 1.9E-130 3 Thioredoxin \$6454-57848 P73920 change protein Drosophila melano-regulator 3.5E-15 R64911 5.4E-110 3 Putative cell division \$1877-60333 L08043 missto gene Drosophila melano-regulator 3.5E-15 R64911 5.4E-110 3 Chkrown or 53.09 Cross-69 104 Glu-tRNA Drosophila melano-regulator 1.6E-123 R64911 5.4E-110 3 Chikrown or 53.09 Cross-69 104 Cross-69 104 Drosophila melano-regulator 1.6E-123 R64911 5.4E-110 3 Chikrown G 50.02-69 104 Cross-69 104 Drosophila melano-regulator 1.6E-123 R64911 5.4E-110 3 Pettinesterase T0 560-12864 Z34271 pir7a Cross-128 R75024 L752-7 R75024 L752-7 L7502 L752-7 L7502 L752-7 L750	-	C	Cold acclimation	51 715–52 578	U73216	complex cold acclimation	Triticum aestivum	4.4E-56	AA650647	1.3E-35	2	11,05
Thioredoxin 56454-57848 P739.0 thiodisulfide inter- change protein Synechocysis 44E-23 AA394562 1.9E-150 3 Putative cell division 8172-60529 U80043 missto gene Drosophila melano- ci 6300 3.5E-15 R64911 5.4E-110 3 Glu-t-RNA 66 821-60868 K00193 Glu-tRNA Drosophila melano- ci 6300 3 5.4E-10 5.4E-11 3 Glu-t-RNA 66 821-60868 K00193 Glu-tRNA Drosophila melano- gaster 1.6E-123 R64911 5.4E-110 3 Pectinesterase 70 560-72864 Z94058 pectinesterase tomato 1.6E-123 H76007 1.3E-106 5 Pectinesterase 70 560-72864 Z34271 pir7a rice 2.5E-27 RA65149 5.4E-20 3 Pectinesterase 75 567-78 69 Z34271 pir7a rice 2.5E-39 AA65149 5.E-20 3 Unknown 8 561-80 761 AC001229 unknown Caenorhabditis elegans 1.2E-13 1.2E-13 1.2E-	-	C	Putative histone bindin	g 53 662–55 308	A25680	nuclear histone	African clawed frog	3.3E-07			3	13
Putative cell division \$8172-60529 U80043 misate gence fregulator Drosophila melano- 3.5E-15 R64911 5.4E-110 3 regulator 16702-68 K 00193 Glu-tRNA Drosophila melano- 3.5E-15 R64911 5.4E-110 3 Glu-t-RNA 66 821-66 868 K 00193 Glu-tRNA Drosophila melano- 1.6E-123 H 76007 1.1E-30 6 Unknown 67029-69 104 Ectimesterase tomato 1.6E-123 H 76007 1.3E-106 2 Hydroxynitrile lyase 7360-72 864 Z94058 pertinesterase tomato 2.3E-27 1.3E-106 2 Reddo-gene 7360-73 869 Z34271 pir7b rice 2.5E-27 R 36024 2.5E-26 2 Reddo-gene 7360-73 869 Z34271 pir7b rice 2.5E-27 R 36024 3.5E-26 2 Unknown 1847-83 341 AC001229 unknown Caenorhabditis elegans 1.2E-08 7.3E-19 4 Unknown 89125-90346 A1	1	*	Thioredoxin	56 454–57 848	P 73920	thiodisulfide inter-	Synechocystis	4.4E-23	AA394562	1.9E-150	3	2,2
Unknown 61577-62335 Cilu-tRNA Drosophila melano- 6 Of 201-68 MA K00193 Cilu-tRNA Drosophila melano- F15442 2.1E-30 6 Cilu-tRNA 66821-66 88 K00193 Cilu-tRNA Drosophila melano- 1.6E-123 H76007 1.3E-106 2 Unknown 67029-69 104 Colu-tRNA Drosophila melano- 1.6E-123 H76007 1.3E-106 2 Hydroxynitrile lyase 7360-75 864 Z34271 pir7a rice 2.5E-27 RA651482 2.5E-26 2 Pseudo-gene lydroxynitrile lyase 73724-78117 UG8339 nucleoporin rat rate 2.5E-27 RA651482 2.5E-26 2 Pseudo-gene lydroxynitrile lyase 75724-78117 UG8339 nucleoporin rat rate 2.5E-27 RA651482 2.5E-26 2 Ushrown 8148-83-341 AC001229 unknown Caenorhabditis elegans 1.2E-03 AA651330 6.9E-10 6 Ushrown 81056-00346 AL002296 patatin<	-	C	Putative cell division	58 172–60 529	U80043	change protein misato gene	Drosophila melano- gaster	3.5E-15	R64911	5.4E-110	33	3,22
Glu-t-RNA Glo 23-06 868 K00193 Glu-tRNA Drosophila melano-gaster F15442 2.1E-30 3 Unknown 67029-69 104 294058 pectinesterase tomato 1.6E-123 F15442 2.1E-30 6 Peutinesterase 70560-72 864 294058 pectinesterase tomato 1.6E-123 H76007 1.3E-106 2 Hydroxynitrile lyase 7362-74 699 234271 pir7a rice 2.3E-27 R3024 3.5E-26 2 Pseudo-gene 7456-75 569 234271 pir7a rice 2.3E-27 R3034 9.5E-20 2 Unknown 78 561-80 761 AC001229 unknown Caenorhabditis elegans 0.0E+00 AAA54170 6.9E-18 6 Unknown 81 847-8 341 AC001229 unknown Caenorhabditis elegans 0.0E+00 AAA54133 6.3E-01 6 Unknown 81 847-8 341 AC09128 unknown Caenorhabditis elegans 1.2E-08 7.21702 7.3E-19 7.3E-19 Unknown	-	C	Unknown	61 577–62 335							9	13
Unknown 67029-69 104 pectinesterase tomato 1.6E-123 F15442 2.1E-30 6 Pectinesterase 70 560-72 864 29 4058 pectinesterase tomato 1.6E-123 H76007 1.3E-106 2 Hydroxymitrile lyase 73 627-74 699 23 4271 pir7b rice 2.5E-27 R30024 9.5E-20 3 Hydroxymitrile lyase 74 956-75 869 23 4271 pir7b rice 2.5E-27 R30024 9.5E-20 3 Unknown 78 561-80 761 AC001229 unknown Caenorhabditis elegams 0.0E+00 AA651330 6.3E-01 4 Unknown 81 487-83 341 AC001229 unknown Caenorhabditis elegams 1.2E-08 7.3E-19 4 Unknown 81 821-88 511 AC0958 patatin tobacco 9.2E-57 H35955 2.5E-106 3 Unknown 93 476-90 968 Patatin homologue 100 626-102 768 Patatin tobacco 9.2E-57 H35957 7.5E-29 AB94671 6.1E-23		≽	Glu-t-RNA	66 821–66 868	K00193	Glu-tRNA	Drosophila melano-				3	4,19
Hydroxynitrile Jyase 73627–74699 Z34271 pir7a rice 3.3E-39 AA651482 2.5E-26 3 Pseudo-gene hydroxynitrile Jyase 74956–75 569 Z34271 pir7b rice 2.3E-27 R30024 9.5E-20 3 Unknown 78 561–80 17 U63839 nucleoporin rat 2.3E-02 AA394956 2.9E-117 4 Unknown 81 487–83 341 AC001229 unknown Caenorhabditis elegans 1.2E-08 AA394956 2.9E-117 4 Unknown 85 221–88 511 Z69793 unknown Caenorhabditis elegans 1.2E-08 T21702 7.3E-19 6 Unknown 89 125–90 346 P15478 patatin tobacco 2.2E-57 H35955 2.3E-106 5 Patatin homologue 108 042–110 423 AJ002596 patatin tobacco 1.1E-153 H35957 7.5E-29 3 Methionyl-amino-like 113 337–115 534 AL00883 methionyl amino-patatin tobacco 1.1E-153 AA39471 6.1E-23 6		≥ ≥	Unknown Pectinesterase	67 029–69 104 70 560–72 864	Z94058	pectinesterase	guster tomato	1.6E-123	F15442 H76007	2.1E-30 1.3E-106	7 9	13
Hydroxymutic tyase 75724-78117 U63839 nucleoporin rat 2.3E-02 AAA231701 6.9E-18 6 Unknown 78 561-80761 AC001229 unknown Caenorhabditis elegans 0.0E+00 AAA34956 2.9E-117 4 Unknown 81 487-83 341 AC001229 unknown Caenorhabditis elegans 1.2E-08 7.3E-19 4 Unknown 85 821-88 511 Z69793 unknown Caenorhabditis elegans 1.2E-08 7.3E-19 4 Unknown 93 476-90 968 Platatin homologue 100 626-102 768 Platatin Platatin 100 626-102 768 Platatin AJ002596 patatin 100 626-102 768 Platatin 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106 100 626-106		≽ ≽	Hydroxynitrile lyase Pseudo-gene	73 627–74 699 74 956–75 569	Z34271 Z34271	pir7a pir7b	rice rice	3.3E-39 2.5E-27	AA651482 R30024	2.5E-26 9.5E-20	3.2	9,01 11,02 11,02
Unknown 81487–83341 AA651330 6.3E-01 6 Unknown 85821–88511 Z69793 unknown Caenorhabditis elegans 1.2E-08 T21702 7.3E-19 4 Unknown 89125–90346 Patatin tobacco 9.2E-57 H35955 2.3E-128 6 Unknown 103 950–106 111 AJ002596 patatin tobacco 4.7E-228 H35956 2.6E-106 3 Patatin homologue 108 042–110 423 AL00883 methionyl amino- yeast 8.8E-70 AA394671 6.1E-23 3 Methionyl-amino- 110 399–112 509 AL00883 methionyl amino- yeast 8.8E-70 AA394671 6.1E-23 2 Unknown 113 337–115 534 P1210 caltractin Atriplex nummularia 2.0E-58 AA394779 2.0E-154 5 Caltractin-like 122 304–123 263 P41210 Caltractin Atriplex nummularia 2.0E-58 AA395741 6.5E-136 2		CC	hydroxynitrue iyase Unknown Unknown	75 724–78 117 78 561–80 761	U63839 AC001229	nucleoporin unknown	rat Caenorhabditis elegans	2.3E-02 0.0E+00	AA231701 AA394956	6.9E-18 2.9E-117	o 4 <i>i</i>	13
Unknown 93476–90968 patatin tobacco 9.2E-57 H35955 2.3E-128 6 Patatin homologue 100 626–102 768 P15478 patatin tobacco 4.7E-228 H35955 2.6E-106 3 Patatin homologue 103 950–110 611 AJ002596 patatin tobacco 1.1E-153 H35957 7.5E-29 3 Methionyl-amino- 110 599–112 509 AL008883 methionyl amino- yeast 8.8E-70 AA394671 6.1E-23 2 Deptidase-like 113 337–115 534 peptidase-like Loknown AA394779 2.0E-154 5 Unknown 120 506–121 725 P41210 caltractin Atriplex nummularia 2.0E-58 AA395741 6.5E-136 2		≥ ≥ ∪	Unknown Unknown Unknown	81 487–83 341 85 821–88 511 89 125–90 346	Z69793	unknown	Caenorhabditis elegans	1.2E-08	AA651330 T21702	6.3E-01 7.3E-19	040	13 13
Fataun homologue 108 042–110 423 AJ00.29% paratin tobacco 1.1E-133 H.5957 7.5E-29 3 Methionyl-amino- 110 599–112 509 AL008883 methionyl amino- yeast 8.8E-70 AA394671 6.1E-29 3 peptidase-like 113 337–115 534 peptidase 6 6 6 Unknown 120 506–121 725 AL1210 caltractin Atriplex nummularia 2.0E-58 AA395741 6.5E-136 2		∪ ≽ ≽ }	Unknown Patatin homologue Patatin homologue	93 476–90 968 100 626–102 768 103 950–106 111	P15478 AJ002596	patatin patatin	tobacco	9.2E-57 4.7E-228	H35955 H35956	2.3E-128 2.6E-106	975	13 6,2 6,2
Unknown 120 506–121 725 Caltractin-like 122 304–123 263 P41210 caltractin Atriplex numnularia 2.0E-58 AA395741 6.5E-136 2	-	≥ ບ	Patatin homologue Methionyl-amino- pentidase-like	108 042–110 423 110 599–112 509	AJ002596 AL008883	patatin methionyl amino- nentidase	tobacco yeast	1.1E-153 8.8E-70	H35957 AA394671	7.5E-29 6.1E-23	z 0	6,2 6,07
		000	Dorwan Unknown Unknown Caltractin-like	113 337–115 534 120 506–121 725 122 304–123 263		popularise caltractin	Atriplex nummularia	2.0E-58	AA394779 AA395741	2.0E-154 6.5E-136	5 2 9	13 13 9,04

Table 1 (continued)

Overview of the 103 predicted genes in the 397-kb contig

No. S^a Function/ Position Bes

Overview	lew o.	of the 103 predicted genes	genes in the 39/-kb contig	ttig							
No.	\mathbf{S}_{B}	Function/ product	Position	Best homologue accession	Name	Species	BLASTP score	EST	BLASTN/EST score	Class ^b	Category ^c
37	υυ	Unknown Heat shock factor	123 922–125 057 125 978–127 024	U68017	HS factor HSF4	Arabidopsis	2.2E-188	R90161	0.0E+00	6	13 11,05
39	≱≽	Unknown Unknown	130 672–133 998							9	4.19.04 13 13
	: ∪ ≽	Ribonucleoprotein Kinase	138 243-140 333 143 838-147 989	S40774 AB000798	ribonucleoprotein protein kinase similar to NPK1	Xenopus Arabidopsis	7.8E-22 1.3E-23	T43343 Z26660	1.6E-05 8.9E-32) m m	4,22 10.04.04
43	C	Putative nicotinate phosphoribosyltransferase	148 171–150 743	Z99120	nicotinate phospho- ribosyltransferase	Bacillus subtilis	5.3E-47	N65739	3.2E-93	8	13
4 ;	C	Unknown	151 299–153 514	AC002330	•		2.9E-07	C21906	3.1E-08	9	13
	ე≽	AP2 Unknown	164 541–166 683 174 463–176 665	U12546 D63999	apetala2 dehydrogenase	Arabidopsis Synechocystis	1.6E-283 1.8E-126			3 -	4.19.04 13
	ე ≱	TINY-like Unknown	178 076–178 665	AJ002598	TINY	Árabidopšis	2.6E-182	AA404810 746429	4.0E-115 4.7E-76	7 ٧	4.19.04
	:≽∪	Cysteine proteinase Putative homeotic	191 501–192 989 193 958–198 256	P25251 A57632	cysteine proteinase BEL1	rape Arabidopsis	7.8E-198 6.3E-80	R84153	4.3E-81	273	6,13 4.19.04
51	≱ ≩	protein Unknown	8363–11049	AC003000	Cal Card	•	6.7E-89	N65197	2.4E-124	ss s	13
	\$	Olikilowii	13 /90-10 202	33/3//	probable memorane protein	yeası	4.4E-13	п//110	1.0E-104	n	CI
53	υυ	Unknown Unknown	16736–17917 19236–20105					AA605507 H77079 T45881	3.9E-120 7.2E-11 9.3E-147	9 9	13 13
	C	Unknown	21 622–22 727	U41558	unknown	Caenorhabditis elegans	7.5E-11	L46423	1.3E-48	4	13
99	C	Geranylgeranyl-pyro-	24 988–26 103	P34802	geranylgeranyl-pyro- phosphatase synthase	A rabidops is	1.0E-179	L37477	1.0E-08	1	20,2
57	*	Ubiquitin-conjugating	27 468–28 378	P52491	ubiquitin-conjugating	yeast	4.0E-34	T41550	0.0E+00	3	6,07
58	ე ≱	protein Unknown Unknown	31311–33255	AJ001694	protein membrane protein	Thermotoga maritima	3.1E-08	R 64792	1 0F-162	4 v	13
	: ≽	Glucosyltransferase	38 024-39 398	Q40287	UTP-glucose glucosyltransferase	cassava	9.6E-106			2 0	1,05
61 62	CC	Aminopepidase Phycoerythrin-like	39 636–42 894 43 401–44 974	AF038591 D45900	X-pro aminopeptidase Ledi-3	rat Lithospernum	5.9E-135 1.50E-49	N96008 R64949	3.0E-68 1.0E-158	2.2	6,13 12
63	≽	protein Putative homeotic	53 303–54 968	X94947	homeotic protein	<i>erythrorhizon</i> tomato	2.0E-23			3	4.19.04
49	≽	protein G-box binding GBF1	57 865–59 767	X63894	G-box binding GBF1	A rabidopsis	1.0E-162	H37673	3.0E-43	1	4.19.04
65 66 67 68 69* 70 71	X X X X X X X X X X X X X X X X X X X	Pseudo-gene Scarecrow homologue Putative storage protein Splicing factor Putative salt inducible Transport protein Unknown Putative transcription initiation factor	59 949–61 848 62 097–63 557 1 69 280–71 174 72 008–75 478 75 900–77 138 80 134–81 937 82 336–83 267 84 089–86 333	Q00765 U62798 Z54364 S20250 AJ002597 Q10286	scarecrow vicilin splicing factor putative salt inducible sugar transporter transcription initiation factor IIB	human 2.0E-20 Arabidopsis 1.0E-23 Matteuccia struthiopteris 4.0E-38 human 8.0E-70 Arabidopsis 0.0E+00 beetroot 1.6E-45 African clawed frog 5.0E-26	2.0E-20 1.0E-23 4.0E-38 8.0E-70 0.0E+00 1.6E-45 5.0E-26	N65163 Z48554 N96704 R65492 H36382	1.0E-175 4.0E-70 1.0E-112 1.0E-168 3.0E-12		1.3 4.19.04 6,2 4,22 11,05 7,07 13 4,19

Table 1 (continued)

Overview of the 103 predicted genes in the 397-kb contig

No. S^a Function/ Position Be

No. S. Principol. Function Position Species SIGNATION STATUTURES Consistency of the control		Claseb Category
73 W Unknown 9169-92429 US8068 sec14 rice 5.0E-21 74 W Unknown 9303-97272 US8068 sec14 rice 5.0E-21 75 W Unknown 9933-100 709 Cane finger Pecularies 2.7E-17 75 W Unknown 11622-108470 2.7276 L.EA Precademonus 2.7E-17 75 W Unknown 11662-108470 2.7276 L.EA Precademonus 2.0E-33 80 C Unknown 1200-124218 2.485833 unknown 1.0E-33 81 C Unknown 1200-124218 2.485833 unknown 1.0E-33 81 C Unknown 1200-124218 2.0E-34 1.0E-47 81 W Unknown 14311-144379 2.0E-35 2.0E-35 85 W Unknown 14311-14437 2.0E-48 2.0E-47 85 W Unknown 14311-14437 2.0E-31 2.0E-31 </th <th>EST BLASTN/EST C</th> <th></th>	EST BLASTN/EST C	
73 W Unknown 91109-92429 Cannel Not-Class Society Tick Pseudomonus 5.00-21 75 W Unknown 9393-107.07 Cannelling Cannelling Pseudomonus 2.00-21 75 W Unknown 10.241-104.07 12.8835 unknown 2.00-23 79 C Unknown 11.4868-114.58 2.48535 unknown 1.00-22-108.49 81 C Unknown 11.4868-114.58 2.48535 unknown 1.00-62-10.88 82 W Unknown 1.148-11-14.49 2.48535 unknown 1.00-62-10.88 83 C Unknown 1.20-14.74 2.00-62-21 1.00-62-21 84 W Unknown 1.20-14.74 2.00-62 1.00-68 85 C Unknown 1.43 11-14.74 2.00-64 1.00-69 85 C Unknown 1.43 11-14.74 2.00-64 2.00-64 85 C Unknown 1.43 11-14.74 2.00-74<	21036	
74 W. Dikknown 930923-1097227 Cinic finger Presidente of the control of domain of the control of domain	R65425 1.0E-120 3.0E-74	5 13
75 W Unknown 99353-100 709 (anne finger formaling) Procession of the control of t		6 13
76 W. Duktown 102341-0109 (18381) Diskolase Pseudintentana 2.7E-17 76 W. MADS box protein 10424-16947 PORSS LEA sobbean 2.7E-17 78 V. MADS box protein 10422-16847 PORSS LEA sobbean 2.7E-17 80 C. Urknown 11486-11456 24853 unknown 2.0E-33 2.0E-33 81 C. Urknown 11400-12218 Post of the control of the		3 4.19.01
77 C. Putative LEA protein 10454-10547 P20075 LEA soybeam 2.7E-17 78 W. MADS box protein 10452-106470 19207-16840 2208-33 0.0E-33 79 C. Ushkown 114586-11458 2488.83 unknown 1.0E-83 81 C. Ushkown 11600-2-116950 2488-1495 1.0E-83 1.0E-83 82 C. Ushkown 12190-1-2438 HLH motif hpothetical protein 5/mechacystis 1.0E-55 83 C. Ushkown 12488-13498 Q07283 richobyalin hmman 3.0E-08 85 W. Ushkown 14411-14449 P24859 Sec14 (PI/PC) transport Altabidopsis 4.0E-47 85 C. Ushkown 14311-14440 No.25441 trichobyalin mouse 2.0E-93 86 C. Ushkown 144810-147488 P24859 Sec14 (PI/PC) transport Altabidopsis 4.0E-47 89 C. Putatiwe acyltransferase 148-494-151 020 X95641 transferase 1.0E-193 80 C. Putatown		3 13
Name MADS box protein 10424-10487 V 12770 unknown Preca abies 2.0E-33 PARS OF Unknown 119 568-11485 248583 unknown 1.08-33 1.0E-33 RO C Unknown 114 568-11485 248583 unknown 1.0E-83 1.0E-83 RO Unknown 121 900-124218 1.0E-12483-12480 0.07283 trichohyalin human 3.0E-08 RO Unknown 121 818-12489 0.07283 trichohyalin human 3.0E-08 RO C Unknown 143 110-14488 P24859 Seel4 (PI/PC) transport Klayreromyces lacris 4.0E-47 RO Unknown 143 110-14488 P24859 Seel4 (PI/PC) transport Klayreromyces lacris 4.0E-47 RO Unknown 153 143-1440 serine C-palmitoyl mouse 2.0E-93 PO W Unknown 155 149-1450 X9884 peroxidase 165 140-16497 X9884 peroxidase PO W Unknown 160 18c 176 18 160 18c 176 18 112 ribosomal protein Arabidopsis 1.0E-104 PO W Unknown 160 18c 172 18 176 1743 P46681 <td>Z29854 1.0E+131</td> <td>3 12</td>	Z29854 1.0E+131	3 12
90 C Unknown 1198953-111956 St. 248833 unknown 10P-83 81 C Unknown 114508-114956 St. 248833 unknown 114508-114956 114508-11495 114508-114956 114508-114966 114508-114966 114508-114966 114508-114966 114508-114966 114508-114966 114508-114966		7
80 C. Unknown 116 0821-116 995 81 C. Unknown 116 090-12431 HLH motif Appropriate protein 5ynechocyssis 7.0E-55 82 W. Unknown 127 848-129 498 D90900 HLH motif bypothetical protein 5ynechocyssis 7.0E-55 83 W. Unknown 127 848-129 498 D90900 Trichchyalin human 3.0E-08 86 C. Unknown 143 111-143 479 Sec14 (PI/PC) transport Kluyveromyces lacts 4.0E-47 87 W. Unknown 143 111-143 479 Sec14 (PI/PC) transport Kluyveromyces lacts 4.0E-47 89 C. Unknown 153 143-154 40 September of Signal Society Arabidopsis 2.0E-93 90 W. Unknown 156 150-165 695 September of Signal Society Arabidopsis 2.0E-93 91 W. Unknown 160 186 or 162 618 Arabidopsis 1.0E-109 92 W. Unknown 160 186 or 162 618 Patzit Arabidopsis 2.0E-93 93 W. Unknown 160 186 or 162 618 Patzit Patzit Arabidopsis	7	13
Color Colo		
8.2 W Unknown 1217848-12438 B Nood-124218 8.3 C Unknown 127848-12498 D9006 hypothetical protein 5)mechocystis 7.0E-55 8.5 W Unknown 144110-14479 207283 trichohyalin human 3.0E-08 8.7 W Unknown 144110-14479 Secl4 (PI/PC) transport Klayveromyces lacis 4.0E-47 8.8 C Unknown 14810-14488 P24859 Secl4 (PI/PC) transport Klayveromyces lacis 4.0E-47 8.8 C Unknown 153143-15490 serine C-palmitoyl mouse 2.0E-93 90 W Unknown 153143-15490 serine C-palmitoyl mouse 2.0E-09 91 W Unknown 156310-16590 P36212 L12 ribosomal protein Arabidopsis 2.0E-10 92 W MAP kinase 16530-16590 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 94 W Peroxidase 16370-16490 P36212 L12 ribosomal protein Arabidopsis 7.0E-68 95 W Potein Protein<		
8.3 C. Unknown 12426.34 L 26.304 l HLH mottl HLH mottl hypothetical protein Synechocystis 7.0E-55 8.4 W. Unknown 13418-136.498 l 209006 107283 trichohyalin human 3.0E-08 8.6 C. Unknown 144180-147488 P24859 Secl 4 (PI/PC) transport Kluyveromyces (act is protein) 4.0E-47 8.8 C. Unknown 145110-143490 X95641 serine C-palmitoyl mouse 2.0E-93 9.0 W. Unknown 155143-154400 X95641 serine C-palmitoyl mouse 2.0E-93 9.0 W. Unknown 156512-156956 Arabidopsis 0.0E+00 0.0E+00 9.0 W. Unknown 166312-156976 Arabidopsis 1.0E-10 0.0E+00 9.1 W. Dutative ribosomal 16638-16590 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 9.2 W. MaP kinase 16638-16590 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 9.1 W. Unknown 16638-16590 P36212 L12 ribosomal protein Ara	;	
84 W Unknown 13784-132498 D90006 hypothetical protein Synechocystis 7.0E-55 85 W Unknown 14718-136498 Q07283 trichohyalin human 3.0E-08 86 C Unknown 140167-140841 P24859 Sec14 (PHPC) transport Klayveromyces lactis 4.0E-47 88 C Unknown 15314-15490 X95641 transferase 1.0E-093 90 W Unknown 15519-156956 XP5641 transferase 1.0E-09 91 W Unknown 156519-156975 APR kinase 7.0E-06 1.0E-00 91 W Unknown 1665120-168375 Q39027 MAP kinase 7.0E-06 91 W Unknown 1665120-168375 Q39027 MAP kinase 7.0E-06 92 W Unknown 1665120-16840 P0000 P0000 P0000 P0000 93 W Unknown 1728-167649 P42743 ubiquitin-conjugating Arabidopsis 1.0E-10 94 W Unknown 1728-1174872 Q09316 P42743	AA395050 1.0E-138	
85 W Cuknown 1341821345498 Q07283 trichohyalin human 3.0E-08 86 C Unknown 143111-143491 Potein 2.40E-07 4.0E-47 87 C Unknown 143111-143491 Protein Sec14 (PI/PC) transport Kluyveromyces lactis 4.0E-47 89 C Putative acyltransferase 148494-151 020 X95641 serine C-palmitoyl mouse 2.0E-93 90 W Unknown 155 19-156 956 MAP kinase 185 20-156 956 0.0E+00 91 W Unknown 165 19-156 956 Arabidopxis 0.0E+00 92 W MAP kinase 165 19-16 970 X98804 peroxidase 1.0E-109 93 W Unknown 165 16-16 970 X98804 peroxidase 1.0E-109 94 W Patative ribosomal 165 36-165 900 P36212 L12 ribosomal protein Arabidopxis 2.0E-10 95 W Putative ribosomal 165 36-165 90 P36212 L12 ribosomal protein Arabidopxis 1.0E-109 95 W Unknown 169 146-172 037<	7	4 13
86 C Unknown 140167–143479 Sec14 (PI/PC) transport Klayveromyces lactis 4.0E-47 87 C Unknown 143111–143479 Sec14 (PI/PC) transport Klayveromyces lactis 4.0E-47 89 C Unknown 153143–154490 Septembritos Condend 2.0E-93 90 W Unknown 15619–16596 MAP kinase 158 205–13937 Q39027 MAP kinase 2.0E-93 91 W Unknown 160 S120-116813 MAP kinase 160 S120-116813 0.0E+00 0.0E+00 92 W Unknown 160 S120-165906 Pactoridase 163 S01-165900 Pactoridase 1.0E-109 94 W Unknown 160 S120-165900 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 95 W Putative ribosomal 165 S61-165900 P36212 L12 ribosomal protein Arabidopsis 2.0E-19 95 W Potein Potein 160 H66-172037 P4681 AAPP protein Arabidopsis 2.0E-19 95 W Unknown 160 H66-172037 P4681 AL022141 Arabidopsis 2.0E-93 101 W Galactosidase 182 484-192778 P4582		
87 W Unknown 143111–143479 Sec14 (PI/PC) transport Kluyveromyces lactis 4.0E-47 88 C Unknown 144810–147488 P24859 Sec14 (PI/PC) transport Kluyveromyces lactis 4.0E-47 89 C Putative acyltransferase 148494–151020 X95641 serine C-palmitoyl mouse 2.0E-93 90 W Unknown 15519–166 956 MAP kinase 158205–158 37 Q39027 MAP kinase 7.0E+00 91 W Unknown 160 5120-168 183 Deported Arabidopsis 0.0E+00 95 W Putative ribosomal 165 301-165 90 P36212 L12 ribosomal protein Arabidopsis 1.0E-109 95 W Putative ribosomal 165 361-165 90 P36212 L12 ribosomal protein Arabidopsis 7.0E-68 95 W Putative ribosomal 165 361-165 90 P36212 L12 ribosomal protein Arabidopsis 7.0E-18 96 C Ubiquitin-conjugating 167 16-1737 P46681 A1P2 protein Parabidopsis 7.0E-14 97 C Actin-interacting 169 16-1		
88 C Unknown 144810-147488 P24859 Secl4 (PI/PC) transport Kluyveromyces lactis 4.0E-47 89 C Putative acyltransferase 148494-151 020 X95641 protein 2.0E-93 90 W Unknown 153143-154490 2.0E-93 91 W Unknown 156310-156956 3.00 92 W MAP kinase 158205-159373 Q39027 MAP kinase 7.0E-00 92 W Unknown 160186 or 162181 1.0E-109 1.0E-109 94 W Putative ribosomal 165361-165900 P36212 L12 ribosomal protein 4.0E-109 95 W Putative ribosomal 165361-165900 P36212 L12 ribosomal protein 4.0E-109 95 W Putative ribosomal 165361-165900 P36212 L12 ribosomal protein 4.0E-109 96 C Ubiquitin-conjugating 166728-167649 P46881 A1822 A1822 97 C Actin-interacting 169146-172037 P46881 A1	0.0E + 00	
89 C Putative acyltransferase 148 494–151 020 X95641 protein Protein 2.0E-93 90 W Unknown 155 194–156 956 4.0 1.0E-109 0.0E+00 91 W Unknown 156 519–156 956 3.73 4.0 Arabidopsis 0.0E+00 92 W Unknown 160 186 or 162 18 3.33 4.0 Arabidopsis 0.0E+00 93 W Unknown 160 186 or 162 18 163 701–164 970 8.8804 peroxidase 4.0 Arabidopsis 1.0E-109 95 W Putative ribosomal 165 361–165 900 P36212 L12 ribosomal protein 4.0 Arabidopsis 1.0E-109 95 W Putative ribosomal 165 361–165 90 P36212 L12 ribosomal protein 4.0 <t< td=""><td>AA395164 1.0E-08</td><td></td></t<>	AA395164 1.0E-08	
89 C Putative acyltransferase 148 494-151 020 X95641 serine C-palmitoyl mouse 2.0E-93 90 W Unknown 153 143-154490 Conception 153 143-154490 Conception Co		
90 W Unknown 153143-154490 Condition 91 W Unknown 156519-156956 MAP kinase 15820-156956 0.0E+00 92 W MAP kinase 15820-156956 0.0E+00 0.0E+00 93 W Unknown 1605120r 161813 0.0E+00 0.0E+00 94 W Peroxidase 163701-164970 X98804 peroxidase 1.0E-109 95 W Protein 160186 or 162618 P42743 ubiquiin-conjugating 166728-167649 P42743 ubiquiin-conjugating 166728-167649 P42743 ubiquiin-conjugating 166728-167649 P42743 ubiquiin-conjugating 16728-167649 P42743 ubiquiin-conjugating 166728-167649 P42743 ubiquiin-conjugating 16728-167649 P42743 ubiquiin-conjugating 16728-167649 P42743 ubiquiin-conjugating Arabidopsis 1.0E-19 97 C Actin-interacting 169146-172037 P4681 P4743 ubiquiin-conjugating Arabidopsis 1.0E-14 98	R90586 8.0E-81	3 1,06
MAP kinase 188 205-158 556 W. MAP kinase 158 205-158 556 W. MAP kinase 158 205-158 576 W. Unknown 165 150-1618 W. Unknown 165 301-164 970 W. Peroxidase 163 701-164 970 W. Putative ribosomal 165 301-164 970 W. Chidinin-conjugating 166 728-167 649 W. Chidinown 172 841-174 872 W. Chidinown 172 841-174 872 W. Chidinown 182 008-182 368 W. Chidinown 182 008-182 368 W. Chidinown 182 008-182 368 W. Chidinown 184 844-192 778 W. Chidinown 194 311 > 198 459 W. Chidinown 194 311 W. Chidinown		13
92 W. MAP kinase 158 205 159 373 Q39027 MAP kinase 7 Arabidopsis 0.0E+00 93 W. Unknown 160 512 or 161 813 160 186 or 162 618 160 512 or 161 813 0.0E+00 0.0E+00 94 W. Peroxidase 163 701-164 970 X98804 peroxidase Arabidopsis 1.0E-109 95 W. Putative ribosomal 165 361-165 900 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 96 C. Ubiquitin-conjugating 166 728-167 649 P42743 ubiquitin-conjugating 166 128-167 649 P42743 ubiquitin-conjugating 1.0E-144 97 C. Actin-interacting 169 146-172 037 P46681 AIP2 protein yeast 1.0E-144 98 W. Unknown 172 841-174 872 Q09316 unknown Caenorhabditits elegans 1.0E-144 100 C. Unknown 182 008-182 368 B-galactosidase 194 44-192 778 P45582 B-galactosidase 10 Arabidopsis 1.0E-120 101 W. Acid phosphatase 195 361-195 881 AL022141 acid phosphatase		C1 12
94 W Peroxidase 106 512 or 161 813 Control 94 W Peroxidase 160 512 or 162 618 1.0E-109 95 W Putative ribosomal 165 361-165 900 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 95 W Putative ribosomal 165 361-165 900 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 96 C Ubiquitin-conjugating 166 728-167 649 P42743 ubiquitin-conjugating Arabidopsis 7.0E-68 97 C Arabidopsis Arabidopsis 7.0E-144 97 C Arabidopsis 1.0E-144 98 W Unknown 172 841-174872 Q09316 unknown 7.0E-244 99 W Unknown 182 008-182368 55379 cytochrome P450 4rabidopsis 2.0E+91 101 W β-Galactosidase 188 44-192778 P45582 β-galactosidase precursor 7.0E-256 102 W Arcid phosphatase 196 431 > 198	TA6344 1 OF 64	
94 W Peroxidase 160 186 or 162 618 Peroxidase 160 186 or 162 618 10E-109 95 W Putative ribosomal 163 701–164 970 X98804 peroxidase 1.0E-109 95 W Putative ribosomal 165 361–165 900 P36212 L12 ribosomal protein 7.0E-11 96 C Ubiquitin-conjugating 166 728-167 649 P42743 ubiquitin-conjugating 4 rabidopsis 7.0E-68 97 C Actin-interacting 169 146–172.037 P4681 AIP2 protein yeast 1.0E-144 98 W Unknown 172 841–174.872 Q09316 unknown Caenorhabditis elegans 1.0E-121 99 W Cytochrome P450 177 617–181 645 S53379 cytochrome P450 Arabidopsis 2.0E+91 100 C Unknown 182 008–182 368 ACC Pecalactosidase 184 84–192 778 P4582 Pecalactosidase procensor 1.0E-121 101 W Acid phosphatase 196 431 > 198 459 ACC Acid ph	1.0E-64	13.04.04
94 W Peroxidase 163 701–164 970 X98804 peroxidase Arabidopsis 1.0E-109 95 W Putative ribosomal 165 361–165 900 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 96 C Ubiquitin-conjugating 166 728–167 649 P42743 ubiquitin-conjugating Arabidopsis 7.0E-68 97 C Actin-interacting 169 146–172 037 P46681 AIP2 protein yeast 1.0E-144 99 W Unknown 172 841–174 872 Q09316 unknown Caenorhabditis elegans 1.0E-121 100 C Unknown 182 008–182 368 P45582 β-galactosidase tomato 7.6E-256 101 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 196431 > 198 459 7.5E-120 102 W Unknown 196431 > 198 459 Acid phosphatase 196431 > 198 459 Arabidopsis 7.5E-120 103 W Unknown 196431 > 198 459 Arabidopsis 7.6E-2	10.70:1	
95 W Putative ribosomal 165 361–165 900 P36212 L12 ribosomal protein Arabidopsis 2.0E-11 96 C Ubiquitin-conjugating protein 166 728–167 649 P42743 ubiquitin-conjugating Arabidopsis 7.0E-68 97 C Actin-interacting protein 169 146–172 037 P46681 AIP2 protein yeast 1.0E-144 98 W Unknown 172 841–174 872 Q09316 unknown Caenorhabditis elegans 1.0E-121 99 W Cytochrome P450 177 617–181 645 S55379 cytochrome P450 Arabidopsis 2.0E+91 100 C Unknown 182 484–192 778 P45582 β-galactosidase tomato 7.6E-256 102 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 196 431 > 198 459 7.5E-120 103 W Unknown 196 431 > 198 459 Arabidopsis 7.5E-120 PThe homology classification can be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last colu	F13569 1 0E+113	2 12
96 C Ubiquitin-conjugating protein 166 728–167 649 P42743 ubiquitin-conjugating protein Arabidopsis 7.0E-68 97 C Actin-interacting protein 169 146–172 037 P46681 AIP2 protein yeast 1.0E-144 98 W Unknown 172 841–174 872 Q09316 unknown 1.0E-121 99 W Cytochrome P450 177 617–181 645 S55379 cytochrome P450 Arabidopsis 2.0E+91 100 C Unknown 182 008–182 368 P45582 β-galactosidase tomato 7.6E-256 101 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase Arabidopsis 7.5E-120 103 W Unknown 196 431 > 198 459 Acid phosphatase Arabidopsis 7.5E-120 nonclinal category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last colu The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last colu	1.0E-107	3 5 01
96 C Ubiquitin-conjugating protein 166 728–167 649 P42743 ubiquitin-conjugating protein 4rabidopsis 7.0E-68 97 C Actin-interacting protein 169 146–172 037 P46681 AIP2 protein 1.0E-144 98 W Unknown 172 841–174 872 Q09316 unknown 1.0E-121 99 W Cytochrome P450 177 617–181 645 S55379 cytochrome P450 4rabidopsis 2.0E+91 100 C Unknown 188 484–192 778 P45582 β-galactosidase tomato 7.6E-256 101 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 4rabidopsis 7.5E-120 103 W Unknown 196 431 > 198 459 Acid phosphatase 4rabidopsis 7.5E-120 8.5. strand. b The homology classification can be found in Table 3. chromational category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columbe columbe.		
96 C Ubiquitin-conjugating protein 166 728–167 649 P42743 protein ubiquitin-conjugating protein 4rabidopsis 7.0E-68 97 C Actin-interacting protein 169 146–172 037 P46681 AIP2 protein 1.0E-144 98 W Unknown 172 841–174 872 Q09316 unknown 1.0E-121 99 W Cytochrome P450 177 617–181 645 \$55379 cytochrome P450 4rabidopsis 2.0E+91 100 C Unknown 182 008–182 368 P45582 β-galactosidase 10mato 7.6E-256 101 W β-Galactosidase 198 484–192 778 P45582 P-galactosidase 4rabidopsis 7.5E-120 102 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 4rabidopsis 7.5E-120 103 W Unknown 196 431 > 198 459 Acid phosphatase 196 431 > 198 459 Acid phosphatase 7.5E-120 - S. strand. b-The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columb	AA042150 0.0E+00	
97 C Actin-interacting 169146–172 037 P46681 AIP2 protein yeast 1.0E-144 98 W Unknown 172 841–174 872 Q09316 unknown Caenorhabditis elegans 1.0E-121 99 W Cytochrome P450 177 617–181 645 \$55379 cytochrome P450 4rabidopsis 2.0E+91 100 C Unknown 188 484–192 778 P45582 β-galactosidase 100 4rabidopsis 7.6E-256 102 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 4rabidopsis 7.5E-120 103 W Unknown 196 431 > 198 459 Acid phosphatase 196 431 > 198 459 7.5E-120 *S, strand. *S, strand. **Arabidopsis **Arabidopsis 7.5E-120 **DThe homology classification can be found in Table 3. **The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columnum in the MIPs web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html).	AA728692 6.0E-85	3 2,13
97 C Actin-interacting 169146–172037 P46681 AIP2 protein yeast 1.0E-144 98 W Unknown 172841–174872 Q09316 unknown Caenorhabditis elegans 1.0E-121 99 W Cytochrome P450 177617–181645 \$55379 cytochrome P450 Arabidopsis 2.0E+91 100 C Unknown 188484–192778 P45582 p-galactosidase tomato 7.6E-256 101 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase Arabidopsis 7.5E-120 103 W Unknown 196431>198 459 Acid phosphatase Arabidopsis 7.5E-120 *S, strand. *S, strand. **Arabidopsis **Arabidopsis **Arabidopsis **Arabidopsis *** Arand.		
98 W Unknown Caenorhabditis elegans 1.0E-121 99 W Cytochrome P450 177 617–181 645 S55379 cytochrome P450 4rabidopsis 2.0E+91 100 C Unknown 182 008–182 368 P45582 β-galactosidase 7.6E-256 101 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 7.5E-120 102 W Acid phosphatase 196 431 > 198 459 AL022141 acid phosphatase 7.5E-120 103 W Unknown 196 431 > 198 459 Acid phosphatase 7.5E-120 *S, strand. *S, strand. *Acid phosphatase Found in Table 3. *Acid phosphatase *Arabidopsis 7.5E-120	AA042437 1.0E-70	2 9,04
98 W Unknown 172841–174872 Q09316 Unknown Catenorhabditis elegans 1.0E-121 99 W Cytochrome P450 177617–181645 \$55379 cytochrome P450 4rabidopsis 2.0E+91 100 C Unknown 182 008–182 368 P45582 p-galactosidase tomato 7.6E-256 101 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 4rabidopsis 7.5E-120 103 W Unknown 196 431 > 198 459 7.5E-120 7.5E-120 aS, strand. bThe homology classification can be found in Table 3. chttp://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columnum category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columnum can be columnum can be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columnum can be columnum can can be columnum can can be columnum can		
99 W Cytochrome P450 177617–181 645 S55379 cytochrome P450 Arabidopsis 2.0E+91 100 C Unknown 182 008–182 368 β-galactosidase tomato 7.6E-256 101 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase Arabidopsis 7.5E-120 102 W Acid phosphatase 196 431 > 198 459 Acid phosphatase 7.5E-120 103 W Unknown 196 431 > 198 459 Acid phosphatase 7.5E-120 *S. strand. *S. strand. *PThe homology classification can be found in Table 3. *CThe functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columnum of the material columnum of		
100 C Unknown 182008–182368 101 W β-Galactosidase 188484–192778 P45582 β-galactosidase tomato 7.6E-256 precursor 102 W Acid phosphatase 193761–195881 AL022141 acid phosphatase 7.5E-120 103 W Unknown 196431 > 198459	N96214 0.0E+00	2 20,1
precursor precursor 102 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase 7.5E-120 precursor 103 W Unknown 196431 > 198 459	000	6 13
102 W Acid phosphatase 193 761–195 881 AL022141 acid phosphatase Arabidopsis 7.5E-120 103 W Unknown 196431 > 198 459 S. strand. Table 3. cThe functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last columnum precursor acid phosphatase Arabidopsis 7.5E-120 7.5	23 /2/3 1.0E-1/9	
103 W Unknown 196431 > 198459 S, strand. The homology classification can be found in Table 3. The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last colu		
aS, strand. ^b The homology classification can be found in Table 3. The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last colu		13.57.57
"s, strand. ¹ The homology classification can be found in Table 3. The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last colu		
The nonloogy cassurcation can be found in Table 3. The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last colu		
The functional category to be found at the MIPS web site (http://mundac.mips.blocnem.de/arabilica/gene/funcat_table.num). In the fast cold	TOT of the state o	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1
* wars fully commond and submitted to the GanBank as compate cDNAs with accession numbers A 1007506. A 1007508. The numbering is don	putative cognate ENIS are indicate	ed ($>$ 95% identified $>$ 700707 (∞
To September 1921 The Company of the Company as cognate CD/NAS with accession numbers AJ0022596. The numbering is done according to the EMIDL/Company entires 2597/07 (genes 7 20079) The numbering is done according to the EMIDL/Company entires 2597/07 (genes 250.).	ecording to the EMBL/Genbank e	nunes 2.99 / 0 / (ge
1-50) and 299/08 (genes 51-105).		
e GenBank as cognate cDNAs with accession numbers AJ002596-AJ002598. The numbering is dot	ccording to tr	ie EMBL/Genbank (

Table 2 Classes of similarities to genes

Class	BLAST E value	Type of matching protein	Number	Predicted function
1	identical	same protein	4	4
2	$< 10^{-50}$	known protein	23	23
3	between 10^{-10} and 10^{-50}	known protein	33	28
4	$< 10^{-10}$	hypothetical protein	8	_
5	$> 10^{-10}$	none, but has cognate EST match	11	_
6	< 150	none, no cognate EST match	24	_
		Total	103	55

five. In addition, the re-annotation pointed to a few frame-shifts caused by sequence errors that have been corrected this way. Similar issues have been raised for annotation of bacterial genomes [18], but to a lesser extent. As the task of annotation of a higher eukaryote genome is much more difficult because of the larger size of the genome, the lesser gene content, the split gene structure, and less homologies to known database entries, such a result is not surprising. It is clearly a warning for caution when using the present-day annotations, and it indicates that complete re-annotation of the genome will be needed.

3.3. Statistical analysis of the contig

As can be deduced from Table 1, 59 genes can be found on the Watson and 44 on the Crick strand. Gene density is quite high (3.85 kb/gene) compared to that of the FCA region [3]

(4.8 kb/gene) and the mean density of 4.1 kb/gene reported for 6.7 Mb sequenced on chromosome 5 [19]. Nevertheless, regions of 1.2 Mb on chromosome 5 have been reported with a similarly high gene density of 3.84 kb/gene [20]. In total, 42% of the genes are highly similar to *Arabidopsis* EST sequences (>95% similarity).

Table 3 represents a statistical analysis done on both strands of the 400-kb contig in terms of occurrence and size of genes, introns, and exons. Intergenic regions cover approximately 53% of the contig, whereas introns represent 16% and coding sequences 31%. As a mean, the first intron is longer than the others and the first and last exons longer than the middle ones. It is noteworthy that local as well as strand heterogeneities were observed. For example, genes in the first half of the sequence have smaller introns than those in the second half of the contig, whereas exons have a similar aver-

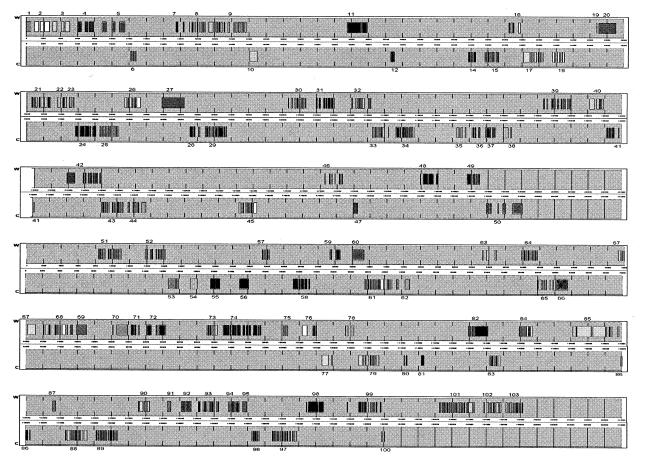


Fig. 2. Schematic overview of the exons of the 103 genes present in the 379-kb contig. Genes are marked by numbers; each number represents a new gene. Gene 13 is not indicated.

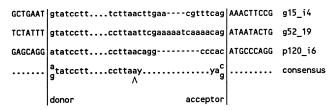


Fig. 3. U12 class introns. g15_i4, U12-type intron #4 of gene 15 from this contig; g52_i9, U12-type intron #9 of gene 52 from this contig; p120_i6, U12-type intron #6 from the human p120 gene. The intron sequence is given in lowercase. \land marks the branch point.

age size (data not shown). In addition, both exons and introns on the Watson strand are larger than those on the Crick strand.

3.4. Finding of two U12-type introns

The 102 protein-encoding genes were found or predicted to contain 429 introns. All of these introns, except two, have the consensus sequences of classical U2-type introns. The remaining two are the last intron (i4) from gene 15, which is distantly related to the histone-binding protein from Xenopus, and the last intron (i9) from gene 52, which is clearly similar to a small yeast gene family encoding membrane proteins of unknown function. Interestingly, a paralogue of gene 52 was found in another contig from chromosome 4 (al022224/al021637), but this gene does not seem to have any intron. These introns, although having GT-AG borders, display the distinctive features of U12-type introns with their very conserved donor and branch sites and the typical short distance between the potential branch point and the acceptor site with no polypyrimidine tract, as shown in Fig. 3. Such introns were initially found in animals where they have been shown to need specific snRNAs for their splicing. More recently, they have been found in plants too [21,22] and are considered to be of rare occurrence. The finding of two members out of a total 429 may give a first estimation of their actual frequency in the Arabidopsis genome. For the time being, these introns are not correctly predicted by any gene prediction program.

3.5. Gene clustering

There are three clusters of tandem gene repeats in the contig. A cluster of five *P450* genes is found at its 5' extremity. These genes have the same genomic structure (three exons) and their coding sequences are very similar to each other (74–83% similarity between copies 1–4, 60% between them and copy 5), their best homologue being an elicitor-induced *P450* from licorice. There is another *P450* gene (gene 99) at the other end of the contig that differs from the genes in the cluster (low homology, 10 exons). The clustering of these five *P450* genes suggests that they might originate from a common ancestor late in evolution and could probably be involved in different steps of the same pathway (secondary metabolism, defence, etc.).

Three patatin genes (genes 30, 31, and 32) are also found clustered. Patatin is the major storage protein of potato and homologues have already been found in other species, but not yet in *Arabidopsis*. The patatins found in this cluster are similar to each other, the first and second copy being very close (90%). The third copy is more distant (67–69% similarity to the others) and differs specifically at the N-terminus, suggesting a different subcellular localization of this member.

The third cluster is an imperfect tandem repeat of a gene encoding a protein with strong similarity to hydroxynitrile lyase, an enzyme that produces cyanide by hydrolyzing cyanogen glucosides, which are secondary metabolites produced in a narrow range of plants. Only the first repeat contains a complete and potentially functional gene with three exons. The gene in the second repeat seems truncated after the first exon, and would thus be a pseudo-gene. Paralogues of this gene have been found in the *FCA* region [3]; the finding of such a gene is unexpected, because *Arabidopsis* has not been reported to produce cyanogen glucosides. It would be interesting to check whether these genes are functional and induced by pathogens and/or predator attack, and to examine which substrate their products would hydrolyze.

3.6. Genome duplication

During the process of gene search, several genes at the 5' extremity of the contig were observed to have homologues located in the AC002391 contig from chromosome 2. To

Table 3				
Statistical	analysis	of the	AP2	contig

Total	Number	Mean per gene	Total size (bp)	Mean size (bp)	%	W strand ^c (+)	C strand ^c (-)
Genic regions							
RNA CDS	1		71				
Protein CDS ^a	100		121 289	1214	31.2	56*	44*
Without introns	15		13 707	914		7*	8*
With introns	85		107 692	1270		49*	36*
Introns ^{a,b}							
Total	429	5.05	63 530	148	16.2	160	133
First	85		16 785	197		248	129
Rest	344		46 745	136		137	134
Exons ^{a,b}							
Total	526	6.19	107 692	205		230	180
First	85		25 961	305		314	293
Rest	346		55 625	161		176	141
Last	85		26 106	307		364	229
Intergenic regions							
	102		208 556	2045	52.6		

^aStatistics on full-length genes.

^bMeans refer to intron-containing genes; for all genes, the mean numbers of introns and exons per gene are 4.29 and 5.26, respectively.

^eW and C strand columns refer to sizes, except for figures followed by an asterisk, where they represent numbers. The region analyzed is the whole 396-kb contig, except for gene 13, for which the prediction is uncertain (a pseudo-gene).

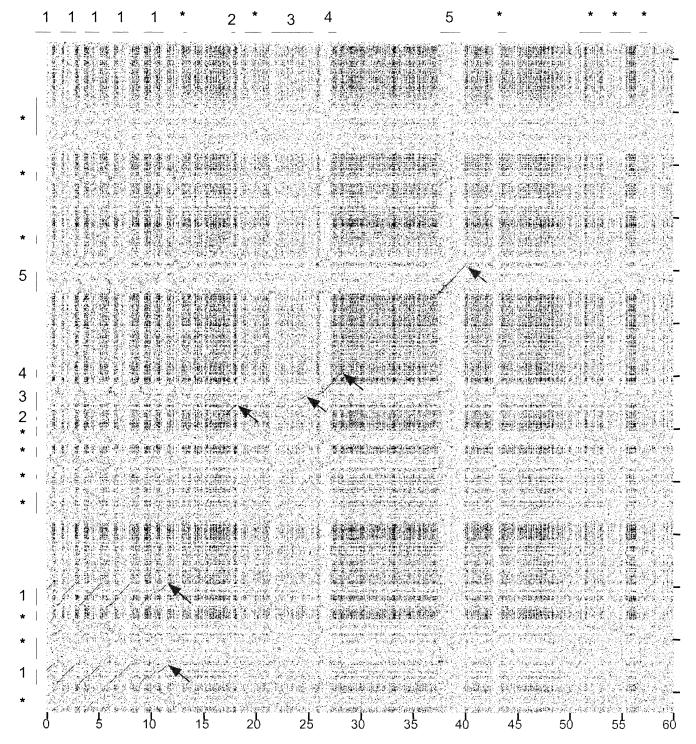


Fig. 4. Dot-plot comparison of a region on chromosome 2 (AC002391, bp 15000–75000) with the first 60 kb of the sequence described here (Z99707). Genes are numbered as in Table 1. Asterisks indicate non-homologous genes, the arrows inside the plot mark the regions of homologous genes.

check the gene arrangement, a dot-plot comparison of both contigs was performed using the GCG compare/dot-plot and the dotter [23] programs. As seen in Fig. 4, the first 45 kb (Watson strand) of the contig show similarities to a region of comparable size of the AC002391 contig (Crick strand, 77 000–32 000). The extent of the duplication is perhaps larger, because no sequence is available yet on the 5' side of the AP2 contig described here. The homology is patchy, being mostly

restricted to the coding sequence of the homologous genes. Among the 12 genes found in this region on the contig, nine have a counterpart in the chromosome 2 contig, the order and the orientation of the genes being conserved. This situation strongly suggests an ancient duplication of this region in the *Arabidopsis* genome. This duplication must indeed be ancient, because most non-coding sequences (introns, intergenic sequences) are not conserved and because the two copies

differ by several insertions/deletions of various genes. One such deletion is found in the chromosome 2 counterpart of gene 9 for which homologies are found with several metal (Cu, Cd) transporters. On chromosome 4, gene 9 is predicted with 13 exons, encoding a 819-amino acid protein. On chromosome 2, the corresponding gene is truncated on its 5' side and is predicted to have three exons encoding a 221-amino acid protein. The P450 cluster described above is the 5'-most part of the duplicated region. Whereas the chromosome 4 copy contains five P450 members in tandem repeat, only two P450 members are found in the chromosome 2 counterpart, separated by an insertion of two foreign genes and flanked by a P450 of different origin. The genetic distances between the different P450 members suggests that the duplication of the P450 genes on chromosome 2, and at least the duplications responsible for the four first copies on chromosome 4, occurred after the duplication of the 45-kb region itself. Although the Arabidopsis genome is of small size, duplication of individual genes appears to be frequent. Duplication on a larger scale has been suggested from mapping data [24], but, to our knowledge, this is the first demonstration of a duplication of a large region in Arabidopsis. The completion of the Arabidopsis genome sequence will tell us to which extent these duplications occur. A comparative analysis of the repeats in various ecotypes and neighbor species will inform us when this happened in the Arabidopsis evolutionary his-

Acknowledgements: The authors wish to acknowledge the ABRC Stock Center in Ohio for the distribution of ESTs, M. Bevan for coordinating the ESSA program, H.W. Mewes, S. Klostermann, and N. Chalwatzis from MIPS for computer data analysis, Peter Breyne and Koen Goethals for critical reading of the manuscript, M. De Cock for help preparing it, and Rebecca Verbanck for the figures. This work was supported by the EU (Biotech ERBB102-CT93-0075 and ERBB104-CT96-0338) and by the Flemish government (VLAB-COT).

References

- Goodman, H.M., Ecker, J.R. and Dean, C. (1995) Proc. Natl. Acad. Sci. USA 92, 10831–10835.
- [2] Arabidopsis Genome Initiative (1997) Plant Cell 9, 476–478.
- [3] Bevan, M., Bancroft, I., Bent, E., Love, K., Piffanelli, P., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema, W., Drost, L., Ridley, P., Hudson, S.-A., Patel, K., Murphy, G., Wedler, H., Wedler, E., Wanbutt, R., Weitzenegger, T., Pohl, T., Terryn, N., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Lecharny, A., Kreis, M., Lao, N., Kavanagh,

- T., Hempel, S., Kotter, P., Entian, K.-D., Rieger, M., Scholfer, M., Funk, N., Muller-Auer, S., Silvey, M., James, R., Montfort, A., Pons, A., Puigdomenech, P., Douka, A., Voukelatou, E., Milioni, D., Hatzopoulos, P., Piravandi, E., Obermaier, B., Hilbert, H., Duesterhoeft, A., Moores, T., Jones, J., Eneva, T., Palme, K., Benes, V., Rechman, S., Ansorge, W., Cooke, R., Berger, C., Delseny, M., Volckaert, G., Mewes, H.-W., Schueller, C. and Chalwatzis, N. (1998) Nature 391, 485–488.
- [4] Koornneef, M., van Eden, J., Hanhart, C.J., Stam, P., Braaksma, F.J. and Feenstra, W.J. (1983) J. Hered. 74, 265–272.
- [5] Bowman, J.L., Smyth, D.R. and Meyerowitz, E.M. (1989) Plant Cell 1, 37–52.
- [6] Jofuku, K.D., den Boer, B.G.W., Van Montagu, M. and Okamuro, J.K. (1994) Plant Cell 6, 1211–1225.
- [7] Dean, C. and Lister, C. (1996) in: Weeds World: The International Electronic Arabidopsis Newsletter (Anderson, M., Ed.), Vol. 3 (iii), pp. 1–6, Nottingham Arabidopsis Stock Centre, Nottingham.
- [8] Zabeau, M. and Vos, P. (1993) European Patent Application EP 534858A1.
- [9] Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. and Zabeau, M. (1995) Nucleic Acids Res. 23, 4407–4414.
- [10] Bancroft, I., Love, K., Bent, E., Sherson, S., Lister, C., Cobbett, C., Goodman, H.M. and Dean, C. (1997) in: Weeds World: The International Electronic Arabidopsis Newsletter (Anderson, M., ed.), Vol. 4 (ii), pp. 1–9, Nottingham Arabidopsis Stock Centre, Nottingham.
- [11] Déhais, P., Coppieters, J. and Van Montagu, M. (1996) Arch. Physiol. Biochem. 104, B38.
- [12] Staden, R. (1996) Mol. Biotechnol. 5, 233-241.
- [13] Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S. (1996) Nucleic Acids Res. 24, 3439– 3452.
- [14] Tolstrup, N., Rouzé, P. and Brunak, S. (1997) Nucleic Acids Res. 25, 3159–3163.
- [15] Borodovsky, M. and MacIninch, J.D. (1993) Comput. Chem. 17, 123–133.
- [16] Burge, C. and Karlin, S. (1997) J. Mol. Biol. 268, 78-94.
- [17] Pedersen, A.G. and Nielsen, H. (1997) in: Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. and Valencia, A., Eds.), pp. 226–233, American Association for Artificial Intelligence Press, Menlo Park, CA.
- [18] Galperin, M.Y. and Koonin, E.V. (1998) In Silico Biol. 1, 0007 (http://www.bioinfo.de/isb/1998/01/0007/).
- [19] Nakamura, Y., Sato, S., Kaneko, T., Kotani, H., Asamizu, E., Miyajima, N. and Tabata, S. (1997) DNA Res. 4, 401–414.
- [20] Kaneko, T., Kotani, H., Nakamura, Y., Sato, S., Asamizu, E., Miyajima, N. and Tabata, S. (1998) DNA Res. 5, 131–145.
- [21] Wu, H.-J., Gaubier-Comella, P., Delseny, M., Grellet, F., Van Montagu, M. and Rouzé, P. (1996) Nature Genet. 14, 383–384.
- [22] Sharp, P.A. and Burge, C.B. (1997) Cell 91, 875–879.
- [23] Sonnhammer, E.L.L. and Durbin, R. (1995) Gene 167, GC1-10.
- [24] Kowalski, S.P., Lan, T.-H., Feldmann, K.A. and Paterson, A.H. (1994) Genetics 138, 499–510.