

## Structural Analysis of *Arabidopsis thaliana* Chromosome 5. V. Sequence Features of the Regions of 1,381,565 bp Covered by twenty one Physically assigned P1 and TAC Clones

Takakazu KANEKO, Hirokazu KOTANI, Yasukazu NAKAMURA, Shusei SATO, Erika ASAMIZU, Nobuyuki MIYAJIMA, and Satoshi TABATA\*

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan

(Received 21 March 1998)

### Abstract

The nucleotide sequences of 21 P1 and TAC clones which have been precisely localized to the fine physical map of the *Arabidopsis thaliana* chromosome 5, were determined, and their sequence features were analyzed. The total length of the regions sequenced in this study were 1,381,565 bp, bringing the total length of the chromosome 5 sequences determined so far to 6,691,670 bp together with the regions of the 69 clones previously reported. By computer-aided analyses including similarity search against protein and EST databases and gene modeling with computer programs, a total of 337 potential protein-coding genes and/or gene segments were identified on the basis of similarity to the reported gene sequences. An average density of the genes and/or gene segments thus assigned was 1 gene / 4,100 bp. Introns were identified in 76.7% of the potential protein genes for which the entire gene structure were predicted, and the average number per gene and the average length of the introns were 3.9 and 176 bp, respectively. These sequence features are essentially identical to those in the previously reported sequences. The numbers of the *Arabidopsis* ESTs matched to each of the predicted genes have been counted to monitor the transcription level. The sequence data and gene information are available on the World Wide Web database KAOS (*Kazusa Arabidopsis* data Opening Site) at <http://www.kazusa.or.jp/arabi/>.

**Key words:** *Arabidopsis thaliana* chromosome 5; genomic sequence; P1 genomic library; TAC genomic library; gene prediction.

To reveal the whole genetic information in a dicot model plant, *Arabidopsis thaliana*, we initiated large-scale structural analysis of the genome. Of five chromosomes which comprise the *A. thaliana* genome of approximately 130 Mb, we chose chromosome 5 as the initial target. For precise localization of the clones for DNA sequencing, we first constructed the fine physical map of the chromosome 5<sup>1</sup> using three genomic libraries of *A. thaliana* Columbia, CICYAC,<sup>2</sup> P1<sup>3</sup> and TAC (Transformation-competent Artificial Chromosome, Mitsui Plant Biotechnology Research Institute, Japan). The template clones were selected from the P1 and TAC libraries based on the DNA markers on the fine physical map by means of polymerase chain reaction (PCR) using marker-specific primers. The isolated P1 and TAC clones were then subjected to shotgun-based sequence analysis. We previously reported the sequences of a total of 5.31 Mb which were obtained by analysis of 69 P1 and TAC clones.<sup>4–7</sup> In this paper, we newly determined

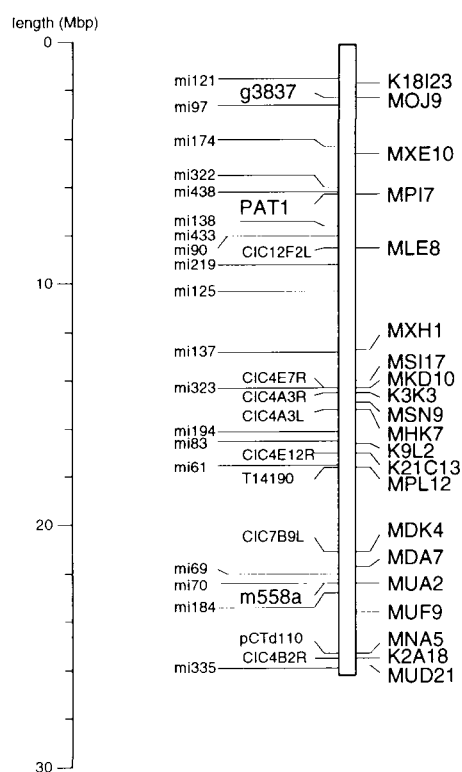
the sequences of 21 additional P1 and TAC clones. We describe gene organization and structural and functional information of the genes assigned in the sequenced regions which were deduced by computer-aided analysis.

### 1. Isolation and Sequencing of P1 and TAC Clones

Two types of genomic clones of *A. thaliana* Columbia, P1 and TAC, which are respectively represented by adding "M" and "K" to the first letters of the clone names, were used for sequence analysis. The average insert length of the P1 and TAC clones was approximately 80 kb. The P1 and TAC clones, containing the DNA segments which cover 21 different regions of chromosome 5, were isolated by screening the Mitsui P1 and TAC libraries by PCR with the primers designed from the sequence information of DNA markers of the defined positions, as described previously.<sup>4–7</sup> The DNA markers and the selected clones (in parentheses) are: MOP10\_Left (K18I23), g3837 (MOJ9), ends of MAC12 and MSH12 (MXE10), PAT1 (MPI7),

Communicated by Mituru Takanami

\* To whom correspondence should be addressed. Tel. +81-438-52-3933, Fax. +81-438-52-3934, E-mail: tabata@kazusa.or.jp



**Figure 1.** Relative locations of the P1 and TAC clones sequenced in this study and the associated markers on the physical map of chromosome 5. The positions of DNA markers used for P1 and TAC isolation and of other major DNA markers were localized on the map on the basis of the YAC tiling path and map information in *ref. 1*. The vertical open bar represents the entire length of chromosome 5. The names of P1 and TAC clones are given at the right, and those of markers at the left. The distance (Mbp) from the telomeric site of the top arm is given on the vertical scale.

CIC12F2L (MLE8), MWP19\_Left (MXH1), ends of MXA21 and MXI10 (MSH17), CIC4E7R (MKD10), CIC4A3R (K3K3), MPO12\_Left (MSN9), CIC4A3L (MHK7), MLN1\_Right (K9L2), CIC4E12R (K21C13), T14190 (MPL12), CIC7B9L (MDK4), MWA11\_Left (MDA7), m558a (MUA2), MUP24\_Right (MUF9), pCITd110 (MNA5), CIC4B2R (K2A18), and MSN2\_Right (MUD21). Relative positions of the markers and the sequenced clones on chromosome 5 are shown in Fig. 1. The orientation of each clone on the fine physical map of chromosome 5<sup>1</sup> has been confirmed by anchoring both ends of the clone to those at the corresponding positions of the map.

The nucleotide sequence of each P1 or TAC insert was determined according to the bridging shotgun method described previously.<sup>4-7</sup> The length of the nucleotide sequence of each P1 or TAC insert finally confirmed is given in parentheses after the clone name at the top of Fig. 2. The total of the DNA regions sequenced in this study

was 1,381,565 bp, bringing the total length of the regions of chromosome 5 sequenced so far to 6,691,670 bp.

## 2. Assignment of Potential Coding Regions

Potential protein coding regions were assigned by similarity search, and gene modeling was performed using prediction programs, as described in the previous papers.<sup>4-7</sup> In brief, similarity search against the non-redundant protein sequence database, owl (release 29), was carried out using the BLASTP<sup>8</sup> program, and information obtained were integrated into the gene models constructed with the aid of the following computer programs: Grail,<sup>9</sup> FEEXA in GeneFinder,<sup>10</sup> ER (Murakami, K., personal communication), ASPL in GeneFinder,<sup>10</sup> GENSCAN<sup>11</sup> and NetPlantGene programs.<sup>12</sup> The transcribed regions were assigned based on a comparison of the nucleotide sequences with *Arabidopsis* ESTs<sup>13,14</sup> in the non-redundant library of GenBank (release 104) and EMBL (release 52) databases using the BLASTN program.<sup>8</sup>

The potential protein-coding regions assigned were divided into three categories. A single exon or a region containing consecutive multiple exons showing similarity to a single reported gene throughout the alignment was designated as a potential protein gene; they were denoted by numbers with the clone names followed by sequential numbers from one end to another of the insert. A region which matched only to portions of a reported gene and only to *Arabidopsis* ESTs were assigned as a potential exon(s) and a transcribed region, respectively. These regions were denoted by adding "p" and "t" between the clone names and the sequential numbers in the identifiers, respectively. All the genes and gene segments assigned in each P1 and TAC clone according to the above procedure were schematically represented in Fig. 2, and the assignment data were listed in the table below each figure. In total, 257 potential protein genes, 23 potential exons, and 57 transcribed regions were assigned in the 1,381,565 bp regions. The number of genes and gene segments assigned so far in the total of 6,691,670 bp (including the previously reported sequences) is 1,573, and an average density of the genes in the three categories is estimated to be 1 gene per 4,254 bp. However, the possibility remains that additional genes may be discovered in the future among the genes and gene segments assigned so far, because our prediction is principally based on similarity to the registered sequences.

In addition to the protein coding regions, RNA coding regions were assigned on the basis of sequence similarity to the reported structural RNAs, and of prediction by the tRNAscan-SE program<sup>15</sup> in the case of tRNA genes. As indicated in Fig. 2, 4 tRNA genes corresponding to 4 amino acid species were identified in the 1,381,565 bp regions. These genes were denoted with the clone names followed by "r" and sequential numbers.

**Table 1.** Structural features of potential protein genes in *A. thaliana* chromosome 5

Features	257 genes <sup>a)</sup>	1,001 genes <sup>b)</sup>
Gene length (bp) including introns	194-10,730 (2,056)	164-11,377 (2,037)
Product length (amino acids)	41-2,756 (458)	41-2,756 (442)
Genes with introns	197	777
Number of introns/genes	0-25 (3.9)	0-42 (3.9)
Exon length (bp)	3-4,287 (283)	2-4,287 (270)
Intron length (bp)	24-2,110 (176)	20-5,405 (181)
GC content of exon	44%	44%
GC content of introns	32%	32%

Structural features of the 257 potential protein genes assigned in this study<sup>a)</sup> and the 1,001 genes assigned so far<sup>b)</sup> are listed. Average values are shown in parentheses.

### 3. Structural Features of Potential Protein Genes

In this report, the complete structures of 257 potential protein-coding genes were predicted by a combination of similarity search and computer prediction. Structural features of these genes, as well as those of 1,001 genes including those previously deduced which correspond to approximately 5.0% of the total gene constituents (20,000 genes) assumed for *A. thaliana*, are listed in Table 1. The average gene length including introns was approximately 2 kb, and 78% of these genes contained introns at an average number of 3.9 per gene. The average lengths of exons and introns were 270 bp and 181 bp, respectively.

### 4. Expression Level of Potential Protein Genes and Gene Segments

The nucleotide sequence of each of the potential protein-coding genes was compared with those in the *Arabidopsis* EST database, and the number of matched *Arabidopsis* ESTs was counted. Of the 337 genes and gene segments that we assigned in chromosome 5 in this study, 166 carried matched ESTs, and 770 out of 1,573 genes and gene segments assigned so far matched ESTs suggesting that the current EST database represents 49% of the gene complement in *A. thaliana*. The putative products of the genes which hit 10 times or more to the EST files include those showing sequence similarity to chlorophyll *a/b*-binding protein in *Pisum sativum* (MDK4.7), methionine synthase in *Mesembryanthemum crystallinum* (MPI7.9), cysteine protease in *P. sativum* (MUF9.1), elongation factor 1- $\alpha$  in *A. thaliana* (MUF9.2), 14-3-3-like protein gf14 kappa in *A. thaliana* (MNA5.10), tubulin beta-4 chain in *A. thaliana* (K9L2.9), reticuline oxidase precursor in *Eschscholtzia californica* (K9L2.12), CLC-*a* chloride channel protein in *A. thaliana* (MHK7.8), ketoconazole resistance protein in *A. thaliana* (K2A18.6), alpha subunit of proteasome in *A. thaliana* (K2A18.13), and ferredoxin-NADP reductase precursor in *Vicia faba* (K2A18.15). These genes are suggested to be a class of highly expressed genes. The

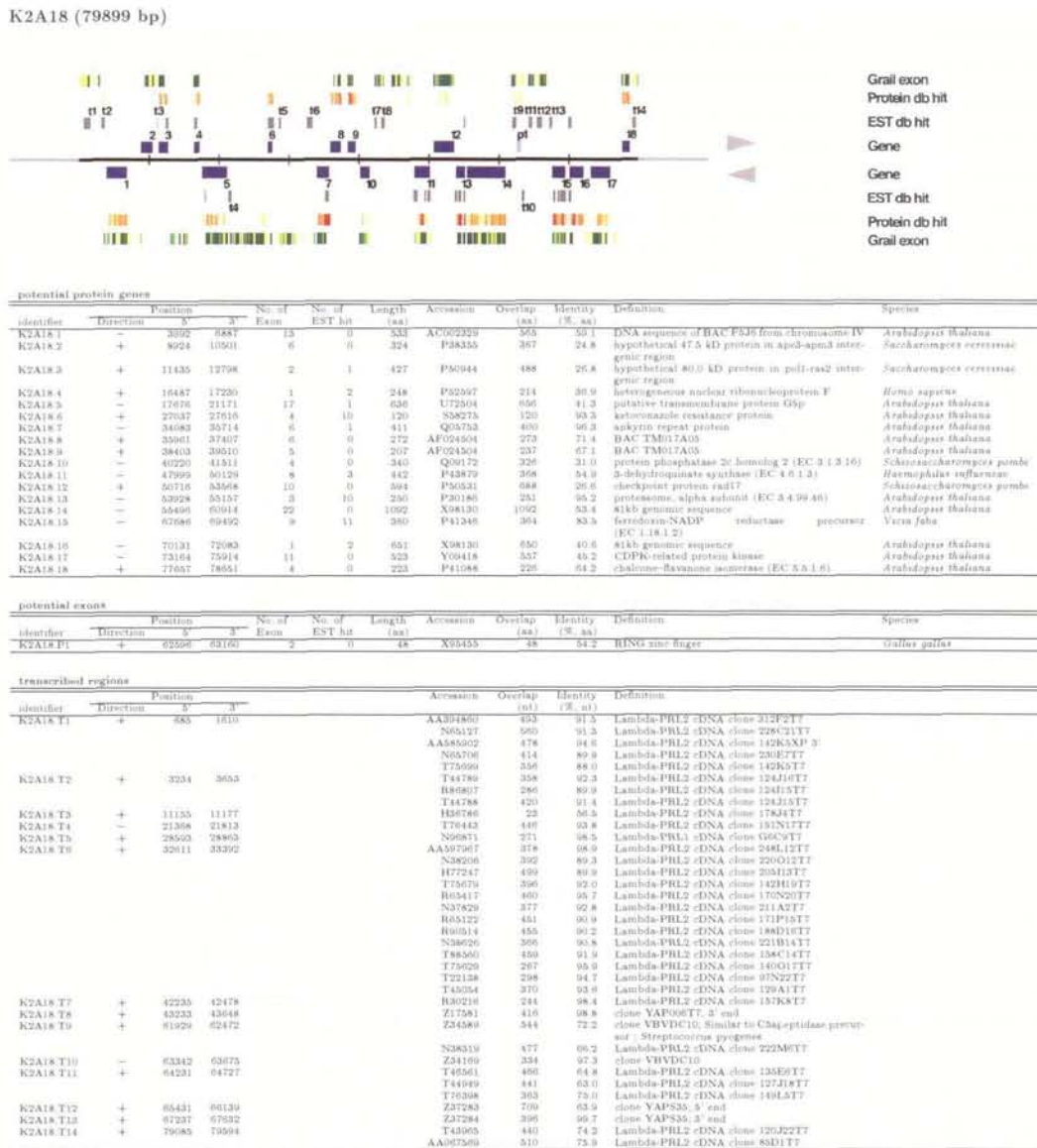
sequence data as well as the gene information shown in this paper are available through the World Wide Web at <http://www.kazusa.or.jp/arabi/>.

**Acknowledgments:** We thank S. Sasamoto and K. Naruo for excellent technical assistance, A. Tanaka for technical advice and the members of DNA Sequencing Laboratory (T. Kimura, T. Hosouchi, K. Kawashima, M. Matsumoto, A. Matsuno, E. Mitsui, A. Muraki, N. Nakazaki, S. Okumura, S. Shinpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and M. Yatabe) for their excellent team work. Thanks are also due to the Mitsui Plant Biotechnology Research Institute, the Research Institute of Innovative Research for the Earth and Arabidopsis Biological Resource Center at Ohio State University for providing the DNA libraries and the DNA markers. This work was supported by the Kazusa DNA Research Institute Foundation.

### References

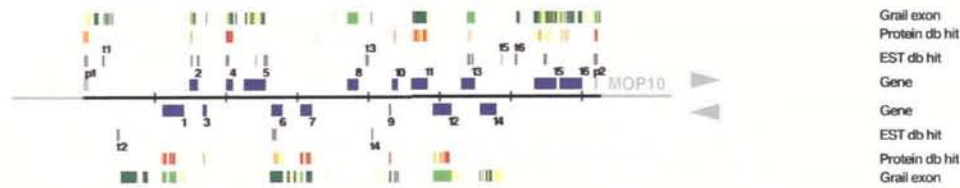
- Kotani, H., Sato, S., Liu, Y-G et al. 1997, A fine physical map of *Arabidopsis thaliana* chromosome 5: Construction of a sequence-ready contig map, *DNA Res.*, **4**, 371-378.
- Creusot, F., Fouilloux, E., Dron, M. et al. 1995, The CIC library: a large insert YAC library for genome mapping in *Arabidopsis thaliana*, *Plant J.*, **8**, 763-770.
- Liu, Y.-G., Mitsukawa, N., Vazquez-Tello, A., and Whittier, R. F. 1995, Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking, *Plant J.*, **7**, 351-358.
- Sato, S., Kotani, H., Nakamura, Y. et al. 1997, Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones, *DNA Res.*, **4**, 215-230.
- Kotani, H., Nakamura, Y., Sato, S. et al. 1997, Structural analysis of *Arabidopsis thaliana* chromosome 5. II. Sequence features of the regions of 1,044,062 bp covered by thirteen physically assigned P1 clones, *DNA Res.*, **4**, 291-300.
- Nakamura, Y., Sato, S., Kaneko, T. et al. 1997, Structural analysis of *Arabidopsis thaliana* chromosome 5. III. Sequence features of the regions of 1,191,918 bp covered by seventeen physically assigned P1 clones, *DNA Res.*, **4**, 401-414.

7. Sato, S., Kaneko, T., Kotani, H. et al. 1998, Structural analysis of *Arabidopsis thaliana* chromosome 5. IV. Sequence features of the regions of 1,456,315 bp covered by nineteen physically assigned P1 and TAC clones, *DNA Res.*, **5**, 41–54.
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–410.
9. Uberbacher, E. C. and Mural, R. J. 1991, Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach, *Proc. Natl. Acad. Sci. USA*, **88**, 11261–11265.
10. Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. 1994, Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, *Nucl. Acids Res.*, **22**, 5156–5163.
11. Burge, C. and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, **268**, 78–94.
12. Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouze, P., and Brunak, S. 1996, Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information, *Nucl. Acids Res.*, **24**, 3439–3452.
13. Newman, T., Bruijn, F. J., Green, P. 1994, Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones, *Plant Physiol.*, **106**, 1241–1255.
14. Cooke, R., Raynal, M., Laudié, M. et al. 1996, Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs, *Plant J.*, **9**, 101–124.
15. Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucl. Acids Res.*, **25**, 955–964.



**Figure 2.** Gene organization in the 21 P1 and TAC clones. The orientation of each clone in this figure (left to right) corresponds to that in Fig. 1 (top to bottom). Positions of the assigned or predicted genes and gene segments in each insert of the P1 and TAC clones were schematically presented by color-coded boxes above (rightward) and below (leftward) the wide line in the middle which represents the entire sequenced region. The insert length is given in parentheses together with the clone name at the top. Arrowheads indicate the directions of the DNA strands (5' to 3'). Dark and faint blue boxes with numbers represent the positions of the assigned potential protein genes and potential exons, respectively, and red bars represent the positions of structural RNA genes. Gray boxes with numbers indicate the positions of the transcribed regions. The regions which showed similarity to the sequences in the protein database are shown by yellow, orange and red bars, corresponding to BLASTP scores of 70–100, 100–250, and 250 or more, respectively. The green bars indicate the positions of the potential exons predicted by the GraIL program. The three different shades of green correspond to the region with GraIL scores of less than 70, 70–90, and 90 or more, respectively. The potential protein genes, the gene segments and the potential RNA genes assigned as described in the text are listed below each of the figures. The accession numbers are as follows: AB011474 (K2A18), AB010694 (K3K3), AB011475 (K9L2), AB010692 (K18I23), AB010693 (K21C13), AB011476 (MDA7), AB010695 (MDK4), AB011477 (MHK7), AB011478 (MKD10), AB010696 (MLE8), AB011479 (MNA5), AB010697 (MOJ9), AB011480 (MPI7), AB010698 (MPL12), AB011481 (MS17), AB010699 (MSN9), AB011482 (MUA2), AB010700 (MUD21), AB011483 (MUF9), AB011484 (MXE10), and AB011485 (MXH1).

K18I23 (72691 bp)

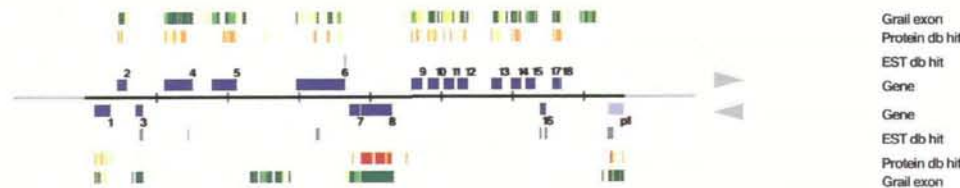


potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K18I23.1	-	11191-14186	3	0	499	S58203	307	31.9	cytochrome P450hr	<i>Sorghum bicolor</i>
K18I23.2	+	14951-16130	4	2	265	P11650	217	32.3	chalcone-flavanone isomerase a (EC 5.5.1.6)	<i>Pisum hybrid</i>
K18I23.3	-	16818-17345	1	0	176	AC002320	182	43.4	DNA sequence of BAC F3J6 from chromosome IV	<i>Arabidopsis thaliana</i>
K18I23.4	+	20058-21015	3	1	255	U30481	255	99.6	expansin At-EXP2	<i>Arabidopsis thaliana</i>
K18I23.5	+	22577-23644	12	3	448	P44843	395	27.8	trk system potassium uptake protein trkh	<i>Haemophilus influenzae</i>
K18I23.6	-	25422-27885	5	1	245	Z57329	410	35.6	chromosome 4, ESSA 1 contig fragment No. 4	<i>Arabidopsis thaliana</i>
K18I23.7	-	30431-32125	4	0	324	X91172	324	85.8	prxK1 gene	<i>Raphanus sativus</i>
K18I23.8	+	37063-38640	1	0	526	P18199	377	25.7	tyrosine-specific transport protein	<i>Escherichia coli</i>
K18I23.9	-	42981-43175	1	0	65	P46269	62	83.9	ubiquitin-cytochrome c reductase complex ubiquinone-binding protein qpc (EC 1.10.2.2)	<i>Solanum tuberosum</i>
K18I23.10	+	43367-44170	1	0	268	AC000104	175	37.7	Sequence of BAC F19P19 from chromosome 1	<i>Arabidopsis thaliana</i>
K18I23.11	+	46059-48345	6	0	565	AC002338	576	64.1	chromosome II BAC T09D09 genomic sequence	<i>Arabidopsis thaliana</i>
K18I23.12	-	49054-51676	1	0	874	AC002294	917	42.1	chromosome I BAC F11P17 genomic sequence	<i>Arabidopsis thaliana</i>
K18I23.13	+	53076-54960	3	2	411	U92527	249	32.1	DNA-binding protein Pit6	<i>Lycopersicon esculentum</i>
K18I23.14	-	55763-58009	6	0	241	P94397	263	28.5	39 kD FK506-binding nuclear protein	<i>Drosophila melanogaster</i>
K18I23.15	+	63883-65498	11	1	558	P25808	640	36.3	ATP-dependent tRNA helicase spb4	<i>Saccharomyces cerevisiae</i>
K18I23.16	+	66921-70674	10	0	658	U13070	428	33.9	cosmid F01F	<i>Caenorhabditis elegans</i>

potential exons										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K18I23.P1	+	1-829	1	2	276	U29580	321	29.0	cosmid ZK546	<i>Caenorhabditis elegans</i>
K18I23.P2	+	71885-72268	1	2	128	P05198	118	65.3	translational initiation factor 2 alpha subunit	<i>Homo sapiens</i>

transcribed regions										
identifier	Direction	Position	Accession	Overlap (nt)	Identity (% nt)	Definition				
K18I23.T1	+	2748-3018	X34592	358	99.3	clone VBVD59				
K18I23.T2	-	4722-5111	Z34174	396	97.7	clone VBVD59				
K18I23.T3	+	39621-40054	N65537	431	95.3	Lambda-PRL2 cDNA clone 231H9T7				
			R64922	336	91.2	Lambda-PRL2 cDNA clone 168K4T7				
			H77054	390	94.7	Lambda-PRL2 cDNA clone 222MT7				
			H77229	303	93.4	Lambda-PRL2 cDNA clone 205P7T				
			W43107	280	94.6	Lambda-PRL2 cDNA clone 246O10T7				
K18I23.T4	-	40438-40668	P20045	231	100.0	clone TAP0273.5 end				
K18I23.T5	+	58719-58837	T45702	89	77.5	Lambda-PRL2 cDNA clone 131D1T7				
K18I23.T6	+	60623-60986	H76598	364	93.7	Lambda-PRL2 cDNA clone 201F23T7				
			T75759	309	94.8	Lambda-PRL2 cDNA clone 145K18T7				

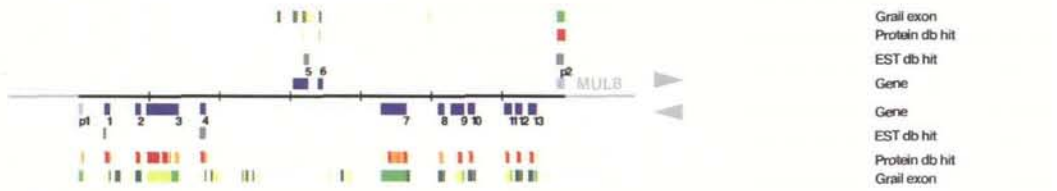
K21C13 (75709 bp)



potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K21C13.1	-	1359-3632	10	0	391	L06594	375	42.4	Avocado polygalacturonase	<i>Persea americana</i>
K21C13.2	+	4099-5902	4	0	357	Z97337	332	39.5	chromosome 4, ESSA 1 contig fragment No. 2	<i>Arabidopsis thaliana</i>
K21C13.3	-	7196-8158	1	1	321	Z97342	287	28.9	chromosome 4, ESSA 1 contig fragment No. 7	<i>Arabidopsis thaliana</i>
K21C13.4	+	11171-15110	5	1	1170	AC002354	1121	39.3	genomic sequence of BAC F06P23 from chromosome IV	<i>Arabidopsis thaliana</i>
K21C13.5	+	17836-21296	5	0	915	U93215	834	31.2	chromosome II BAC T06B20 genomic sequence	<i>Arabidopsis thaliana</i>
K21C13.6	+	29650-36490	9	2	781	Z97342	787	26.8	chromosome 4, ESSA 1 contig fragment No. 7	<i>Arabidopsis thaliana</i>
K21C13.7	+	37134-38649	3	0	443	Z97341	405	28.1	chromosome 4, ESSA 1 contig fragment No. 6	<i>Arabidopsis thaliana</i>
K21C13.8	-	38791-43077	1	0	1429	Z97342	1437	58.2	chromosome 4, ESSA 1 contig fragment No. 7	<i>Arabidopsis thaliana</i>
K21C13.9	+	45874-47394	4	0	338	Z97335	407	31.7	chromosome 4, ESSA 1 contig fragment No. 0	<i>Arabidopsis thaliana</i>
K21C13.10	+	48170-49666	4	0	438	Z97335	410	39.5	chromosome 4, ESSA 1 contig fragment No. 0	<i>Arabidopsis thaliana</i>
K21C13.11	+	50357-51803	4	0	388	Z97335	396	37.6	chromosome 4, ESSA 1 contig fragment No. 0	<i>Arabidopsis thaliana</i>
K21C13.12	+	52342-53842	4	0	389	Z97335	430	33.0	chromosome 4, ESSA 1 contig fragment No. 0	<i>Arabidopsis thaliana</i>
K21C13.13	+	57105-58560	3	0	435	Z97335	454	33.7	chromosome 4, ESSA 1 contig fragment No. 0	<i>Arabidopsis thaliana</i>
K21C13.14	+	59875-61270	5	0	330	P42620	316	48.7	hypothetical 37.4 kD protein in exuI-tdeC intergenic region	<i>Escherichia coli</i>
K21C13.15	+	61932-63202	3	0	371	Z97342	397	29.0	chromosome 4, ESSA 1 contig fragment No. 7	<i>Arabidopsis thaliana</i>
K21C13.16	-	63999-64726	2	5	75	G02284	74	50.0	Human deleted in split hand/split foot 1	<i>Homo sapiens</i>
K21C13.17	+	65711-66679	3	0	173	P42620	237	33.3	hypothetical 37.4 kD protein in exuI-tdeC intergenic region	<i>Escherichia coli</i>
K21C13.18	+	66760-66990	1	0	77	P42620	76	51.3	hypothetical 37.4 kD protein in exuI-tdeC intergenic region	<i>Escherichia coli</i>

potential exons										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K21C13.P1	-	73524-75706	4	2	647	D88748	217	50.7	AR411	<i>Arabidopsis thaliana</i>

K3K3 (68889 bp)



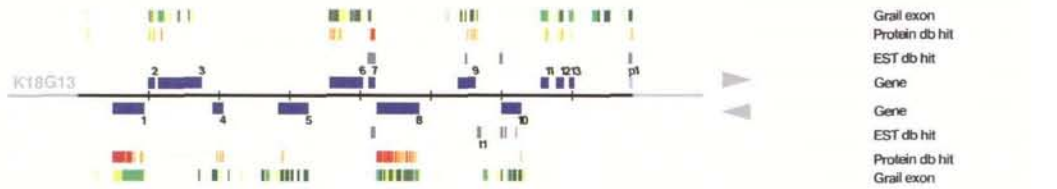
potential protein genes

identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K3K3.1	-	3712 4484	2	2	222	U75192	222	99.5	germin-like protein	<i>Arabidopsis thaliana</i>
K3K3.2	-	8089 8867	2	0	271	U75194	272	92.8	germin-like protein	<i>Arabidopsis thaliana</i>
K3K3.3	-	9676 14219	2	0	1466	AC062962	1453	38.9	Sequence of BAC F20P6 from chromosome 1	<i>Arabidopsis thaliana</i>
K3K3.4	-	17288 18063	2	2	222	U75192	222	100.0	germin-like protein	<i>Arabidopsis thaliana</i>
K3K3.5	+	30427 32350	6	1	303	S77427	327	24.8	hypothetical protein	<i>Synechocystis sp</i>
K3K3.6	+	33088 34966	2	0	111	P48512	109	32.1	transcription initiation factor IIb	<i>Arabidopsis thaliana</i>
K3K3.7	-	42876 45444	1	0	1223	X88130	916	41.2	81kb genomic sequence	<i>Arabidopsis thaliana</i>
K3K3.8	-	50596 51882	2	0	246	U30481	229	54.6	expansin At-EXP2	<i>Arabidopsis thaliana</i>
K3K3.9	-	52756 54964	3	0	320	S53082	250	57.2	hypothetical protein, pollen allergen homolog	<i>Thermus aquaticus</i>
K3K3.10	-	55230 56199	2	0	252	U64893	328	56.1	expansin	<i>Pinus taeda</i>
K3K3.11	-	60388 61422	2	0	263	S53082	248	57.7	hypothetical protein, pollen allergen homolog	<i>Thermus aquaticus</i>
K3K3.12	-	61944 62880	2	0	253	U64893	232	55.6	expansin	<i>Pinus taeda</i>
K3K3.13	-	63797 64994	2	0	284	U30481	220	57.7	expansin At-EXP2	<i>Arabidopsis thaliana</i>

potential exons

identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K3K3.P1	-	1 643	1	0	214	Y12321	211	33.2	SLG-Sc and SLA-Sc genes and Melmoth retro-transposon sequence	<i>Brassica oleracea</i>
K3K3.P2	+	67768 68889	1	5	374	U53418	374	91.4	UDP-glucose dehydrogenase	<i>Glycine max</i>

MDK4 (78596 bp)



potential protein genes

identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MDK4.1	-	4970 9402	2	0	1424	Z97343	1528	37.0	chromosome 4, ESSA 1 contig fragment No. 8	<i>Arabidopsis thaliana</i>
MDK4.2	+	9968 10921	6	0	177	D83970	271	42.8	CPR8 protein	<i>Pinus sarguensis</i>
MDK4.3	+	11374 17552	6	0	454	AC092352	271	32.5	chromosome II BAC F4P9 genomic sequence	<i>Arabidopsis thaliana</i>
MDK4.4	+	19122 20607	3	0	319	S58293	362	48.9	myb-related protein M4	<i>Arabidopsis thaliana</i>
MDK4.5	-	28449 32733	8	0	685	U65390	208	35.0	CaMB-channel protein	<i>Nicotiana glauca</i>
MDK4.6	+	35692 40370	17	0	664	U58087	742	38.7	putative endo/exonuclease MmMre11a	<i>Mus musculus</i>
MDK4.7	+	41186 42142	3	32	265	S33775	265	88.4	chlorophyll a/b-binding protein	<i>Pinus salinus</i>
MDK4.8	-	42371 48351	21	0	1219	S51823	1099	99.1	myosin heavy chain ATM2	<i>Arabidopsis thaliana</i>
MDK4.9	+	53859 56389	10	2	321	S76174	210	53.3	hypothetical protein	<i>Homo sapiens</i>
MDK4.10	-	60027 62849	9	5	492	P40929	297	31.0	hypothetical 32.6 kD protein in met30 chr5 intergenic region	<i>Saccharomyces cerevisiae</i>
MDK4.11	+	65567 66673	1	0	369	Z97340	394	36.0	chromosome 4, ESSA 1 contig fragment No. 5	<i>Arabidopsis thaliana</i>
MDK4.12	+	67755 68822	1	0	336	Z97340	406	35.2	chromosome 4, ESSA 1 contig fragment No. 5	<i>Arabidopsis thaliana</i>
MDK4.13	+	69953 70324	1	0	244	P18160	294	26.1	non-receptor tyrosine kinase spore lysis A	<i>Dictpostelium discoideum</i> (EC 2.7.1.112)

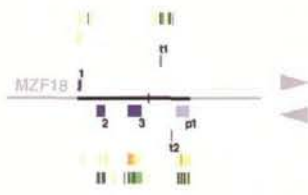
potential exons

identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MDK4.P1	+	78096 78596	2	3	129	L47117	140	45.9	late embryogenesis abundant protein (EMBP)	<i>Picea glauca</i>

transcribed regions

identifier	Direction	Position 5' 3'	Accession	Overlap (nt)	Identity (% nt)	Definition
MDK4.T1	-	66606 67194	H17023 Z18083	423 176	90.3 98.3	Lambda-PRL2 cDNA clone 202E21T7 clone VB12-2032

MLE8 (15871 bp)

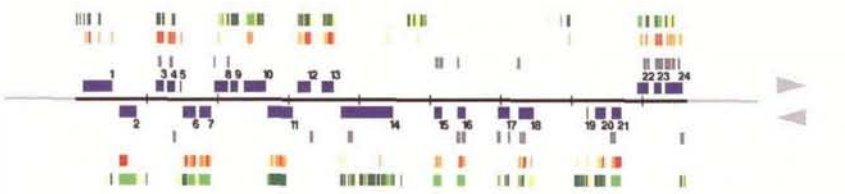


potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MLE8.1	+	1 501	1	0	167	P33543	179	37.4	putative kinase-like protein tmk11 precursor	<i>Arabidopsis thaliana</i>
MLE8.2	-	2732 3846	3	0	303	JC6263	293	27.3	SP6 binding protein homolog	<i>Cucumis sativus</i>
MLE8.3	-	7110 9059	5	0	517	D10973	423	33.8	<i>Synechococcus</i> sp. rpoD1 gene for major sigma factor	<i>Synechococcus</i> sp.

potential exons										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MLE8.P1	-	13863 15797	6	0	474	D78129	355	34.4	Human adult Male liver squalene epoxidase	<i>Homo sapiens</i>

transcribed regions										
identifier	Direction	Position	Accession	Overlap (nt)	Identity (% nt)	Definition				
MLE8.T1	+	11372 11813	F15348	242	98.3	clone YAY1047.5 end				
MLE8.T2	-	13164 13384	F15347	221	97.3	clone YAY1047.3 end				

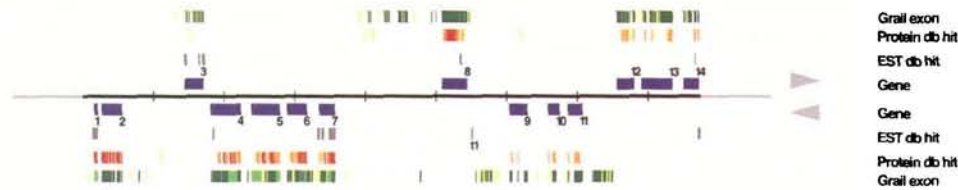
MOJ9 (86380 bp)



potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MOJ9.1	+	1077 5190	7	0	321	S33227	327	56.0	transcription factor OBF5 2, oct element-binding	<i>Zea mays</i>
MOJ9.2	-	6179 8560	1	0	794	D86180	374	61.0	Pea seedling phosphoribosyltransferase	<i>Pisum sativum</i>
MOJ9.3	+	11349 12407	2	1	330	Y08618	333	67.6	polygalacturonase-inhibiting protein	<i>Citrus sinensis</i>
MOJ9.4	+	12917 13989	2	4	330	Y08618	330	62.4	polygalacturonase-inhibiting protein	<i>Citrus sinensis</i>
MOJ9.5	+	14719 14816	1	0	66	Z87339	72	61.1	chromosome 4, ESSA Ycontig fragment No. 4	<i>Arabidopsis thaliana</i>
MOJ9.6	-	15139 15601	2	0	507	S62899	502	50.8	cytochrome P450	<i>Glycine max</i>
MOJ9.7	-	17417 19075	2	0	528	S62899	530	37.4	cytochrome P450	<i>Glycine max</i>
MOJ9.8	+	19574 21498	8	2	284	Z54216	252	34.1	cosmid T24H10	<i>Caenorhabditis elegans</i>
MOJ9.9	+	21834 22892	1	0	353	A57223	216	24.5	pheromone receptor VNS3	<i>Dattya saccipes</i>
MOJ9.10	+	23721 25822	9	0	707	P32583	348	27.3	suppressor protein srp40	<i>Saccharomyces cerevisiae</i>
MOJ9.11	-	27057 30634	3	0	851	P47735	987	35.8	receptor-like protein kinase 5 precursor (EC 2.7.1.-)	<i>Arabidopsis thaliana</i>
MOJ9.12	+	31299 33147	8	1	330	P43273	330	100.0	transcription factor Hbp-1b	<i>Arabidopsis thaliana</i>
MOJ9.13	+	34718 36373	8	0	330	S48122	324	99.4	transcription factor OBF5	<i>Arabidopsis thaliana</i>
MOJ9.14	-	37410 44762	24	1	1036	AC002332	1083	30.2	chromosome II BAC F4P9 genomic sequence	<i>Arabidopsis thaliana</i>
MOJ9.15	-	50630 51570	1	2	347	P32839	394	46.6	flavonol sulfotransferase-like (EC 2.8.2.-)	<i>Arabidopsis thaliana</i>
MOJ9.16	-	53939 55018	1	0	356	P32839	305	47.2	flavonol sulfotransferase-like (EC 2.8.2.-)	<i>Arabidopsis thaliana</i>
MOJ9.17	-	59648 61260	3	2	235	P43393	183	28.4	fruit protein pKIW1501	<i>Actinidia chinensis</i>
MOJ9.18	-	62679 64693	3	4	487	AC000132	486	46.9	Sequence of BAC F21M12 from chromosome I	<i>Arabidopsis thaliana</i>
MOJ9.19	-	72150 72392	1	0	81	AC002062	63	34.9	Sequence of BAC F20P5 from chromosome I	<i>Arabidopsis thaliana</i>
MOJ9.20	-	73373 74848	4	0	363	Z66384	329	43.5	cosmid T11G6	<i>Caenorhabditis elegans</i>
MOJ9.21	-	75722 77089	1	2	456	D30922	439	60.1	novel serine/threonine protein kinase	<i>Arabidopsis thaliana</i>
MOJ9.22	+	79329 80948	2	2	450	Z70521	463	30.0	C. melo mRNA product	<i>Cucumis melo</i>
MOJ9.23	+	81702 82778	4	20	253	AC002329	252	98.9	DNA sequence of BAC F3J6 from chromosome IV	<i>Arabidopsis thaliana</i>
MOJ9.24	+	83289 85744	6	5	521	S51529	510	45.1	SPF1 protein	<i>Ipomoea batatas</i>



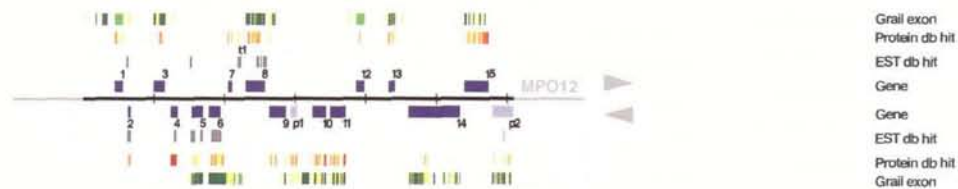
MPL12 (87434 bp)



potential protein genes										
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MPL12.1	-	1023 2051	1	2	143	AC002343	137	41.6	BAC T19F06 genomic sequence	<i>Arabidopsis thaliana</i>
MPL12.2	-	2982 5588	9	0	668	S32816	677	98.5	potassium channel protein KAT1	<i>Arabidopsis thaliana</i>
MPL12.3	+	14448 17080	11	5	419	A53781	412	24.3	ribonucleoprotein La	<i>Drosophila melanogaster</i>
MPL12.4	-	18132 22285	8	3	1205	Z97242	1226	42.7	chromosome 4, ESSA 1 contig fragment No. 7	<i>Arabidopsis thaliana</i>
MPL12.5	-	23863 27868	6	0	1122	U97106	1324	41.1	downy mildew resistance protein RPP5	<i>Arabidopsis thaliana</i>
MPL12.6	-	28932 31739	11	0	636	S52247	620	70.2	probable replication origin activator protein ROA	<i>Zea mays</i>
MPL12.7	-	33469 35656	7	6	473	F52410	473	99.4	3-oxoacyl-[acyl-carrier-protein] synthase I precursor (EC 2.3.1.41)	<i>Arabidopsis thaliana</i>
MPL12.8	+	50829 54431	2	1	1173	D829257	1033	35.7	receptor protein kinase	<i>Arabidopsis thaliana</i>
MPL12.9	-	60433 62928	3	0	326	AC002337	365	37.3	chromosome II BAC T08I13 genomic sequence	<i>Arabidopsis thaliana</i>
MPL12.10	-	65888 67461	2	0	303	X57323	349	30.0	outward rectifying potassium channel KCO1	<i>Arabidopsis thaliana</i>
MPL12.11	-	68658 70747	2	0	443	X97323	375	34.9	outward rectifying potassium channel KCO1	<i>Arabidopsis thaliana</i>
MPL12.12	+	75639 78925	10	0	499	X00929	364	63.2	peptidase of D1 protein	<i>Hordeum vulgare</i>
MPL12.13	+	79059 83416	11	0	1022	U64842	722	25.6	cosmid F25B4	<i>Caenorhabditis elegans</i>
MPL12.14	+	85049 87180	6	2	383	U29536	244	28.7	cosmid F45E12	<i>Caenorhabditis elegans</i>

transcribed regions						
identifier	Direction	Position 5' 3'	Accession	Overlap (nt)	Identity (% nt)	Definition
MPL12.T1	-	54953 55220	AA597897	274	99.3	Lambda-PRL2 cDNA clone 201C2217

MSN9 (61001 bp)

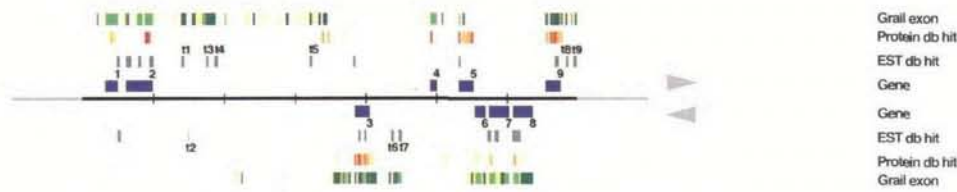


potential protein genes										
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MSN9.1	+	4482 5636	1	0	385	AF007270	290	36.2	BAC IG502P16	<i>Arabidopsis thaliana</i>
MSN9.2	-	6303 6659	2	5	88	Z17697	77	92.2	similar to Chick 60S ribosomal protein L5	<i>Arabidopsis thaliana</i>
MSN9.3	+	9956 11599	1	0	560	AF000657	518	31.1	BAC F19G10	<i>Arabidopsis thaliana</i>
MSN9.4	-	12356 13330	1	1	328	Y11791	276	100.0	peroxidase ATP26a, partial	<i>Arabidopsis thaliana</i>
MSN9.5	-	15332 16904	7	5	315	Q05783	312	36.9	ankyrin repeat protein	<i>Arabidopsis thaliana</i>
MSN9.6	-	17785 19437	1	3	351	Z97335	356	45.9	chromosome 4, ESSA 1 contig fragment No. 0	<i>Arabidopsis thaliana</i>
MSN9.7	+	20526 21078	2	0	69	AF007269	79	39.2	BAC IG502N01	<i>Arabidopsis thaliana</i>
MSN9.8	+	23005 25701	9	3	592	AC001229	552	37.1	Sequence of BAC F5J14 from chromosome 1	<i>Arabidopsis thaliana</i>
MSN9.9	-	26331 28691	5	0	245	Z97339	289	40.1	chromosome 4, ESSA 1 contig fragment No. 4	<i>Arabidopsis thaliana</i>
MSN9.10	-	32420 34338	6	0	326	Z97339	318	64.2	chromosome 4, ESSA 1 contig fragment No. 4	<i>Arabidopsis thaliana</i>
MSN9.11	-	34931 37099	6	0	346	Z97339	346	55.5	chromosome 4, ESSA 1 contig fragment No. 4	<i>Arabidopsis thaliana</i>
MSN9.12	+	38659 39786	1	0	376	AC000104	407	29.7	Sequence of BAC F19P19 from chromosome 1	<i>Arabidopsis thaliana</i>
MSN9.13	+	43294 44127	4	0	188	D21813	193	44.0	Lily ORF	<i>Lilium longiflorum</i>
MSN9.14	-	45030 53261	26	0	709	I49127	613	33.3	intracellular protein	Unknown
MSN9.15	+	53994 57383	14	0	483	U45574	408	97.3	protein farnesyl transferase beta subunit (ERA1) gene	<i>Arabidopsis thaliana</i>

potential exons										
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MSN9.P1	-	29310 30227	3	0	144	Z97339	155	67.9	chromosome 4, ESSA 1 contig fragment No. 4	<i>Arabidopsis thaliana</i>
MSN9.P2	-	37869 60841	15	1	345	I49127	456	31.1	intracellular protein	Unknown

transcribed regions						
identifier	Direction	Position 5' 3'	Accession	Overlap (nt)	Identity (% nt)	Definition
MSN9.T1	+	21815 22332	AA042237	518	91.2	CD4-13 cDNA clone EtG617

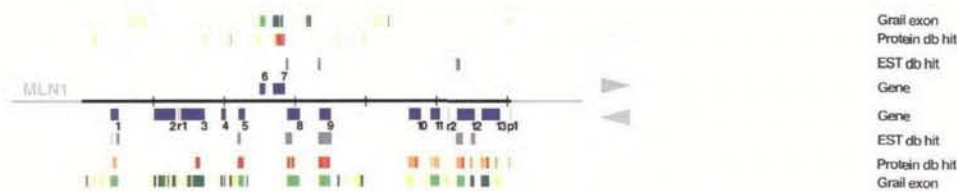
MUD21 (69850 bp)



potential protein genes										
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MUD21_1	+	3251 4952	1	2	584	U62798	628	32.6	SCARECROW	<i>Arabidopsis thaliana</i>
MUD21_2	+	6105 9862	4	4	485	A10373	309	35.2	serine/threonine kinase homolog PRO25	unidentified
MUD21_3	-	38446 49228	9	3	303	Y14316	373	80.2	MAPK gamma protein kinase	<i>Arabidopsis thaliana</i>
MUD21_4	+	49029 50037	1	0	313	AC001229	212	33.3	Sequence of BAC F5114 from chromosome 1	<i>Arabidopsis thaliana</i>
MUD21_5	+	53175 55207	9	1	361	S71172	361	100.0	protein kinase 41K (EC 2.7.1.-)	<i>Arabidopsis thaliana</i>
MUD21_6	-	55425 56861	3	0	415	AC000348	455	29.0	Genomic sequence for BAC T7N9	<i>Arabidopsis thaliana</i>
MUD21_7	+	57412 60235	5	2	809	AC000348	914	26.3	Genomic sequence for BAC T7N9	<i>Arabidopsis thaliana</i>
MUD21_8	-	60819 63611	5	3	815	AF004879	916	27.5	resistance complex protein I2C-2	<i>Lycopersicon esculentum</i>
MUD21_9	+	65441 67516	6	1	479	Z98707	478	60.3	chromosome 4, ESSA 1 AP2 contig fragment No 1	<i>Arabidopsis thaliana</i>

transcribed regions										
identifier	Direction	Position 5' 3'	Accession	Overlap (nt)	Identity (% nt)	Definition				
MUD21_T1	+	13913 14303	Z57613	391	95.3	clone VBVFH11				
MUD21_T2	-	14905 14980	F14147	76	93.4	clone VBVFH11				
MUD21_T3	+	17414 17719	T78536	305	90.0	Lambda-PRL2 cDNA clone 158C20T7				
MUD21_T4	+	18553 19092	AA394643	540	71.8	Lambda-PRL2 cDNA clone 158C20XP_3				
MUD21_T5	+	32096 32387	T13363	292	89.3	Lambda-PRL2 cDNA clone 118D18T7				
MUD21_T6	-	43570 43942	AA586123	373	95.8	Lambda-PRL2 cDNA clone 77B11XP_3				
MUD21_T7	-	44818 45031	T45137	414	95.5	clone 77B11T7				
MUD21_T8	+	68356 68615	T42846	260	77.0	Lambda-PRL2 cDNA clone 117C7T7				
MUD21_T9	+	69382 69804	T76188	423	71.0	Lambda-PRL2 cDNA clone 147F21T7				

K9L2 (60583 bp)

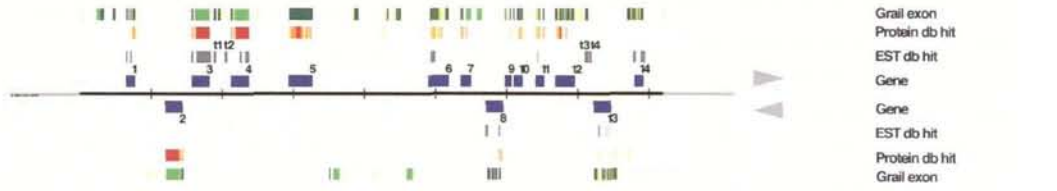


potential protein genes										
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K9L2_1	-	4053 5147	1	3	361	U50439	426	38.0	chromosome II BAC T06D20 genomic sequence	<i>Arabidopsis thaliana</i>
K9L2_2	-	10146 13142	9	0	481	Y10571	373	31.6	dsG gene	<i>Homo sapiens</i>
K9L2_3	+	13948 17218	7	0	624	X27314	474	43.7	cd2 kinase homologue, cd2McC	<i>Medicago sativa</i>
K9L2_4	-	19634 20207	2	0	91	U78721	99	38.6	chromosome II BAC T01B08 genomic sequence	<i>Arabidopsis thaliana</i>
K9L2_5	-	22020 22913	1	1	295	Z37260	104	88.5	<i>A. thaliana</i> transcribed sequence, clone PAP548, complete; Similar to LEA D7 - <i>Gossypium hirsutum</i>	<i>Arabidopsis thaliana</i>
K9L2_6	+	25009 25759	1	0	248	P08723	207	26.6	prostatic spermine-binding protein precursor	<i>Rattus norvegicus</i>
K9L2_7	+	26939 28546	4	0	396	P48260	496	42.9	hypothetical 54.3 kD protein yc724	<i>Cyanophora paradoxa</i>
K9L2_8	-	28932 30965	1	4	388	U34558	587	42.4	Human translation initiation factor eIF3 p66 subunit	<i>Homo sapiens</i>
K9L2_9	-	33473 35022	3	25	444	P24636	444	99.8	tubulin beta-4 chain	<i>Arabidopsis thaliana</i>
K9L2_10	-	46131 47726	1	0	332	P30986	348	41.2	reticuline oxidase precursor (EC 1.5.3.9)	<i>Eichscholtzia californica</i>
K9L2_11	-	49178 50473	1	0	432	AF013253	440	33.2	BAC IG003H0	<i>Arabidopsis thaliana</i>
K9L2_12	-	52904 55397	2	12	541	P30986	533	43.5	reticuline oxidase precursor (EC 1.5.3.9)	<i>Eichscholtzia californica</i>
K9L2_13	-	56432 58934	2	0	518	P30986	518	41.7	reticuline oxidase precursor (EC 1.5.3.9)	<i>Eichscholtzia californica</i>

potential exons										
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
K9L2_P1	-	60394 60581	1	0	62	U59232	67	47.8	berberine bridge enzyme	<i>Fagopyrum esculentum</i>

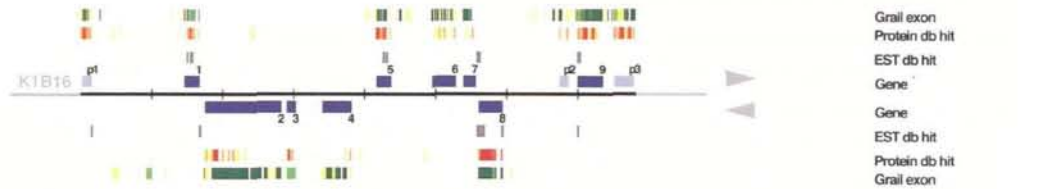
potential RNA genes										
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (nt)	Accession	Overlap (nt)	Identity (% nt)	Definition	Species
K9L2_R1	-	13813 13684	1	0	72	U29948	57	89.0	(RNA-Lys-CCU)	<i>Leishmania tarentolae</i>
K9L2_R2	-	51708 51780	1	0	73	X17513	73	82.0	(RNA-Val-TAC)	<i>Homo sapiens</i>

MDA7 (82033 bp)



potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MDA7.1	+	6458-7708	7	2	225	Z97341	95	77.9	chromosome 4, ESSA 1 contig fragment No. 6	<i>Arabidopsis thaliana</i>
MDA7.2	-	11925-14387	3	0	599	Y11828	599	100.0	hap81.4 gene	<i>Arabidopsis thaliana</i>
MDA7.3	+	15720-18228	3	9	599	Y11827	599	99.4	hap81.2 gene	<i>Arabidopsis thaliana</i>
MDA7.4	+	21243-23750	3	3	599	Y11827	599	98.5	hap81.2 gene	<i>Arabidopsis thaliana</i>
MDA7.5	+	29370-32727	3	0	1064	U77888	1123	36.4	receptor-like protein kinase	<i>Ipomoea nil</i>
MDA7.6	+	48954-51890	8	2	519	Z49130	432	42.8	cosmid T6D98	<i>Caenorhabditis elegans</i>
MDA7.7	+	53511-54981	3	0	320	Y11815	376	39.9	myb factor, 1355 bp	<i>Oryza sativa</i>
MDA7.8	-	57078-59474	10	4	371	P14197	362	36.2	AAC3 protein	<i>Dicotylethium discoidium</i>
MDA7.9	+	59782-60676	2	0	127	X95018	49	57.1	tyrosine/threonine phosphatase 1	<i>Mus musculus</i>
MDA7.10	+	61043-62249	3	0	185	AC090132	172	60.5	Sequence of BAC F21M12 from chromosome 1	<i>Arabidopsis thaliana</i>
MDA7.11	+	64076-65281	4	4	148	L29077	148	89.2	ubiquitin-conjugating enzyme	<i>Pinus edulis</i>
MDA7.12	+	66806-69601	11	0	398	Z59708	328	46.3	chromosome 4, ESSA 1 AP2 contig fragment No. 2	<i>Arabidopsis thaliana</i>
MDA7.13	-	72109-74610	8	2	340	U50478	347	30.3	Solanum lycopersicum actin gene	<i>Lycopersicon esculentum</i>
MDA7.14	+	77979-79226	5	3	275	Z78544	222	30.6	cosmid K04G11	<i>Caenorhabditis elegans</i>

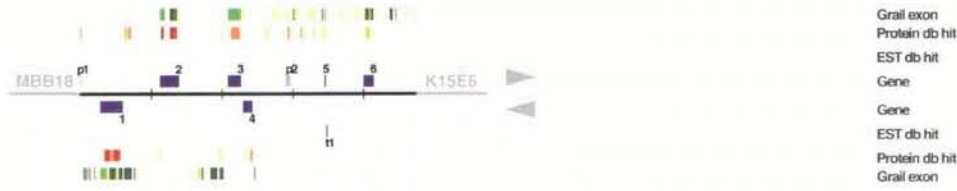
transcribed regions										
identifier	Direction	Position	Accession	Overlap (nt)	Identity (% nt)	Definition				
MDA7.T1	+	18804-19123	N96116 H56261	320 309	59.1 98.7	CD4-15 cDNA clone G1G6T7 Lambda-PRL2 cDNA clone 176O15T7				
MDA7.T2	+	20368-20689	AA051221	322	93.2	Lambda-PRL2 cDNA clone 176O15XP.3				
MDA7.T3	+	70920-71591	AA597540 Z35357	455 202	91.4 100.0	Lambda-PRL2 cDNA clone 200P21T7 clone RAT6A04, complete				
			T43117	215	91.2	Lambda-PRL2 cDNA clone 110K10T7				
			H76823	380	75.5	Lambda-PRL2 cDNA clone 20AATT				
			T44109	379	75.3	Lambda-PRL2 cDNA clone 122116T7				
MDA7.T4	+	71653-71932	Z30864	280	98.6	clone VBVRF09				



potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MHK7.1	+	14446-16770	7	3	307	P25076	313	82.1	cytochrome c1, home protein precursor	<i>Solanum tuberosum</i>
MHK7.2	-	17563-28293	18	0	2756	U49844	2902	30.6	Human FRAP-related protein	<i>Homo sapiens</i>
MHK7.3	-	29060-30301	1	0	414	AF007269	458	39.3	BAC IG002201	<i>Arabidopsis thaliana</i>
MHK7.4	+	34109-38108	11	0	881	P30776	701	25.0	double-strand-break repair protein rad51	<i>Schizosaccharomyces pombe</i>
MHK7.5	+	41651-43709	6	3	403	L47479	368	100.0	atrophophrasin III methylase	<i>Arabidopsis thaliana</i>
MHK7.6	+	49500-52927	14	0	475	Z69635	508	39.2	cosmid F10B6	<i>Caenorhabditis elegans</i>
MHK7.7	+	53948-55648	4	0	465	U39046	504	27.0	myosin heavy chain kinase B	<i>Dicotylethium discoidium</i>
MHK7.8	-	56594-59445	6	15	775	Z71445	776	99.9	CLC-a chloride channel protein	<i>Arabidopsis thaliana</i>
MHK7.9	+	69953-73572	4	3	1029	Z67336	1066	44.5	chromosome 4, ESSA 1 contig fragment No. 1	<i>Arabidopsis thaliana</i>

potential exons										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MHK7.P1	+	116-1325	6	1	284	U31932	284	72.9	amino acid permease 1	<i>Vicia faba</i>
MHK7.P2	+	67357-68714	2	0	217	Z97336	229	33.6	chromosome 4, ESSA 1 contig fragment No. 1	<i>Arabidopsis thaliana</i>
MHK7.P3	+	75129-77946	5	0	812	Z97336	816	51.3	chromosome 4, ESSA 1 contig fragment No. 1	<i>Arabidopsis thaliana</i>

MKD10 (47460 bp)

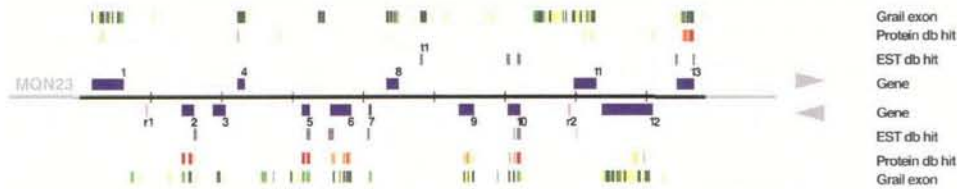


potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MKD10.1	-	2825 5979	6	0	927	L35053	1147	33.2	Transposon MAGGY gag and pol gene homologues	<i>Magnaporthe oryzae</i>
MKD10.2	+	11289 12900	4	0	437	X97075	453	74.8	proline oxidase	<i>Arabidopsis thaliana</i>
MKD10.3	+	20850 22637	1	0	596	AC00182	581	28.4	Sequence of BAC F21M12 from chromosome 1	<i>Arabidopsis thaliana</i>
MKD10.4	-	22977 24254	1	0	494	AF007270	369	33.3	BAC IG002P16	<i>Arabidopsis thaliana</i>
MKD10.5	+	34470 34689	2	0	41	U7566	41	78.0	pollen coat protein homolog	<i>Brassica rapa</i>
MKD10.6	+	40004 41499	3	0	361	Z99708	393	32.3	chromosome 4, ENSA 1 AP2 contig fragment No. 2	<i>Arabidopsis thaliana</i>

potential exons										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MKD10.P1	+	53 178	1	0	42	AF001308	42	81.0	chromosome IV BAC clone Y10M13	<i>Arabidopsis thaliana</i>
MKD10.P2	+	28977 29602	3	0	81	P19456	86	61.4	plasma membrane ATPase 2 (EC 3.6.1.35)	<i>Arabidopsis thaliana</i>

transcribed regions										
identifier	Direction	Position	Accession	Overlap (nt)	Identity (% nt)	Definition				
MKD10.11	-	34742 34925	Z18442	185	96.8	clone GBGy10.3 end				

MNA5 (88356 bp)

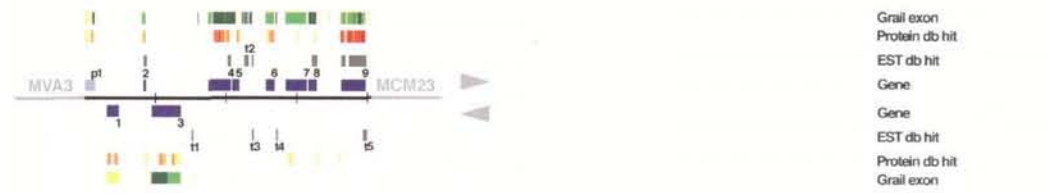


potential protein genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MNA5.1	+	1713 6233	14	0	733	Z49129	678	26.7	cosmid C47G2	<i>Caenorhabditis elegans</i>
MNA5.2	-	14339 16116	3	4	312	P46667	312	100.0	homeobox-leucine zipper protein Athb-5	<i>Arabidopsis thaliana</i>
MNA5.3	-	18730 20536	3	0	324	S72495	129	36.0	transcriptional activator Rb homolog	<i>Rhizobium galegae</i>
MNA5.4	+	22225 23247	1	0	341	AF007270	330	31.8	BAC IG002P16	<i>Arabidopsis thaliana</i>
MNA5.5	-	31308 32408	2	2	290	S99592	132	90.2	histone H3	<i>Schizosaccharomyces pombe</i>
MNA5.6	-	36302 38251	7	2	438	X98130	467	42.4	81kb genomic sequence	<i>Arabidopsis thaliana</i>
MNA5.7	+	40786 42174	1	1	130	U04957	118	44.9	Sea Island polio-rich cell wall protein gene	<i>Corynephorus barbatus</i>
MNA5.8	+	43293 44958	6	0	284	Z10680	246	28.0	cosmid C26G4	<i>Caenorhabditis elegans</i>
MNA5.9	-	53494 55649	6	0	308	Y10162	305	48.5	cyclin-D like protein	<i>Oenopodium rubrum</i>
MNA5.10	-	60417 62194	4	14	245	P48848	245	97.1	14-3-3-like protein gfi4 kappa	<i>Arabidopsis thaliana</i>
MNA5.11	+	69702 72992	11	1	607	JC5133	541	31.2	deubiquitinating enzyme	<i>Mus musculus</i>
MNA5.12	-	73723 80900	22	0	1219	P46875	504	33.3	kinesin-like protein c	<i>Arabidopsis thaliana</i>
MNA5.13	+	84331 86786	8	3	518	U81288	402	75.9	PcRT17.1	<i>Ficus sativum</i>

transcribed regions										
identifier	Direction	Position	Accession	Overlap (nt)	Identity (% nt)	Definition				
MNA5.T1	+	48018 48441	T46302	424	90.2	Lambda-PRL2 cDNA clone 139M2177				

potential RNA genes										
identifier	Direction	Position	No. of Exon	No. of EST hit	Length (nt)	Accession	Overlap (nt)	Identity (% nt)	Definition	Species
MNA5.R1	-	9382 9468	2	0	74	D50933	74	100.0	tRNA-Met(CAU)	<i>Arabidopsis thaliana</i>
MNA5.R2	-	69088 69109	1	0	82	Z67990	55	84.0	tRNA-Ser(GCU)	<i>Caenorhabditis elegans</i>

MPI7 (40548 bp)



potential protein genes											
identifier	Direction	Position	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species	
MPI7.1	-	3190-4899	1	0	570	AC000106	595	31.0	Sequence of BAC F7G19 from chromosome 1	<i>Arabidopsis thaliana</i>	
MPI7.2	+	8366-8583	1	3	106	P11892	106	52.8	616 ribosomal protein r25, chloroplast precursor	<i>Pisum sativum</i>	
MPI7.3	-	9598-13645	5	0	1133	AC002354	1138	38.5	genomic sequence of BAC F06P23 from chromosome IV	<i>Arabidopsis thaliana</i>	
MPI7.4	+	17864-20763	4	1	861	AC002354	769	45.0	genomic sequence of BAC F06P23 from chromosome IV	<i>Arabidopsis thaliana</i>	
MPI7.5	+	20831-21895	1	0	355	U53154	347	23.3	osmid C39G8	<i>Caenorhabditis elegans</i>	
MPI7.6	+	25693-26888	1	0	398	P55081	424	44.8	microfibrillar-associated protein 1	<i>Homo sapiens</i>	
MPI7.7	+	28455-31427	1	0	991	U52064	997	21.4	ORF73 homolog	<i>Kaposi's sarcoma-associated herpes-like virus (Kaposi)</i>	
MPI7.8	+	31884-32838	1	2	385	U51116	444	26.4	NP_180	<i>Mesembryanthemum crystallinum</i>	
MPI7.9	+	36320-39741	11	21	765	U84880	765	89.7	methionine synthase		

potential exons											
identifier	Direction	Position	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species	
MPI7.P1	+	3-1478	1	0	492	AC000106	527	35.7	Sequence of BAC F7G19 from chromosome 1	<i>Arabidopsis thaliana</i>	

transcribed regions											
identifier	Direction	Position	Accession	Overlap (nt)	Identity (% nt)	Definition					
MPI7.T1	-	15171-15395	F03977	225	55.8	AT-NMC cDNA clone H49TP					
MPI7.T2	+	22591-23152	R29784	562	69.2	Lambda-PRL2 cDNA clone 159D4T7					
MPI7.T3	-	23672-23901	F15223	191	91.0	clone GBGD299.3 end					
MPI7.T4	+	27128-27376	F15224	216	86.8	clone GBGD299.5 end					
MPI7.T5	-	36742-36996	F14281	249	92.4	clone YAY871.3 end					
MPI7.T6	-	36742-36996	N96177	255	95.7	Lambda-PRL1 cDNA clone G10C1T7					

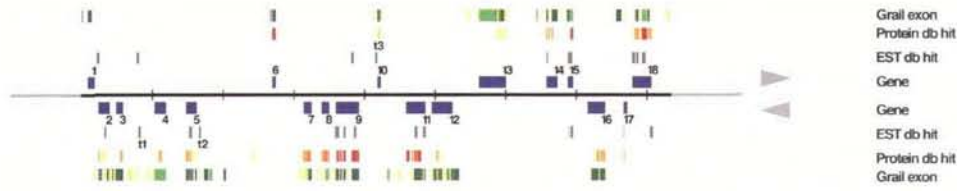
MSI17 (26624 bp)



potential protein genes											
identifier	Direction	Position	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species	
MSI17.1	+	3359-5134	5	0	264	Z92536	218	31.7	cosmid SCY10G2	<i>Mycobacterium tuberculosis</i>	

potential exons											
identifier	Direction	Position	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species	
MSI17.P1	-	1-411	1	0	137	L47183	137	82.8	reverse transcriptase	<i>Arabidopsis thaliana</i>	
MSI17.P2	+	20263-26344	7	0	1329	Z97342	1322	38.5	chromosome 4, ESSA I contig fragment No. 7	<i>Arabidopsis thaliana</i>	

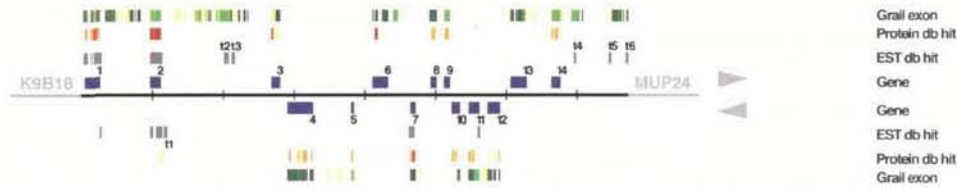
MUA2 (83478 bp)



potential protein genes											
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (nt)	Identity (% aa)	Definition	Species	
MUA2 1	+	985 1897	2	0	171	S58498	193	29.5	IAA12 protein	<i>Arabidopsis thaliana</i>	
MUA2 2	-	2462 3998	6	2	282	Q08623	215	48.8	GSI protein	<i>Homo sapiens</i>	
MUA2 3	-	4952 5863	1	0	304	U92074	326	32.2	Human RecA-like protein	<i>Homo sapiens</i>	
MUA2 4	-	10350 11909	1	0	520	D89136	437	28.6	Fission yeast cDNA product	<i>Schizosaccharomyces pombe</i>	
MUA2 5	-	14839 16276	6	1	274	S46959	277	47.7	poim 1	<i>Solanum tuberosum</i>	
MUA2 6	+	20960 27409	1	0	190	S58882	190	100.0	CCHH finger protein 2	<i>Arabidopsis thaliana</i>	
MUA2 7	-	31413 32490	3	0	285	P35594	273	63.3	brassinosteroid-regulated protein brat	<i>Glycine max</i>	
MUA2 8	-	33940 34990	3	0	284	S71225	284	62.3	xyloglucan endo-transglycosylase-related protein XTR-6	<i>Homo sapiens</i>	
MUA2 9	-	35999 39176	5	6	332	U27609	276	100.0	xyloglucan endo-transglycosylase	<i>Arabidopsis thaliana</i>	
MUA2 10	+	41813 42292	2	0	129	D35922	108	50.0	novel serine/threonine protein kinase	<i>Arabidopsis thaliana</i>	
MUA2 11	-	45938 48621	8	4	630	U58971	557	62.7	calmodulin-binding protein	<i>Nicotiana glauca</i>	
MUA2 12	-	49537 52480	12	0	512	P22805	507	30.2	adenosylmethionine-S- amino-7-oxononanoate aminotransferase (EC 2.6.1.62)	<i>Bacillus sphaericus</i>	
MUA2 13	+	66248 69037	8	0	1054	AC002376	1106	36.1	Sequence of BAC TIG11 from chromosome 1	<i>Arabidopsis thaliana</i>	
MUA2 14	+	65845 67326	3	1	333	S82938	340	40.6	myb-related protein M4	<i>Arabidopsis thaliana</i>	
MUA2 15	+	68806 69515	2	4	207	Z29548	103	95.1	similar to Pathogenesis-related protein 1	<i>Arabidopsis thaliana</i>	
MUA2 16	-	71688 74084	8	0	398	Z97336	413	39.2	chromosome 4, ESSA 1 contig fragment No. 1	<i>Arabidopsis thaliana</i>	
MUA2 17	-	76741 77208	2	2	48	S60256	42	59.6	TGF-beta receptor interacting protein 1 homolog	<i>Arabidopsis thaliana</i>	
MUA2 18	+	77971 80639	11	6	456	X92957	430	75.8	xylose isomerase	<i>Hordeum vulgare</i>	

transcribed regions						
identifier	Direction	Position 5' 3'	Accession	Overlap (nt)	Identity (% nt)	Definition
MUA2 T1	-	7801 8367	F19866	293	98.3	clone OBO259, 3' end
MUA2 T2	+	16640 16874	F19747	355	95.8	clone OBO259, 5' end
MUA2 T3	+	41516 41724	Z24545	235	96.6	clone RAR134, 5' end
MUA2 T3	+	41516 41724	N65448	239	94.6	Lambda-PRL2 cDNA clone 22946T7

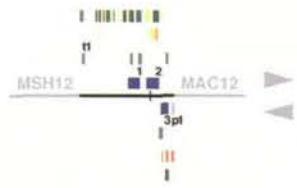
MUF9 (77298 bp)



potential protein genes											
identifier	Direction	Position 5' 3'	No. of Exon	No. of EST hit	Length (aa)	Accession	Overlap (nt)	Identity (% aa)	Definition	Species	
MUF9 1	+	544 2649	8	12	358	Z68291	350	75.1	cysteine protease	<i>Pinus radiata</i>	
MUF9 2	+	8726 11172	2	39	449	P139005	449	100.0	elongation factor 1-alpha	<i>Arabidopsis thaliana</i>	
MUF9 3	+	26829 28017	2	0	299	Z99708	245	48.2	chromosome 4, ESSA 1 AP2 contig fragment No. 2	<i>Arabidopsis thaliana</i>	
MUF9 4	-	29051 32687	11	0	771	U83245	759	39.5	auxin response factor 1	<i>Arabidopsis thaliana</i>	
MUF9 5	-	38093 38419	1	0	109	P38389	86	67.4	protein transport protein sec61 beta subunit	<i>Arabidopsis thaliana</i>	
MUF9 6	+	41083 43287	3	0	439	S48856	504	40.1	zinc finger protein pcpl	<i>Solanum tuberosum</i>	
MUF9 7	-	48419 47165	1	2	249	S29959	267	44.9	arabinogalactan-like protein	<i>Pinus taeda</i>	
MUF9 8	+	49315 50127	1	0	271	P35196	285	37.2	hypothetical 32.7 kD protein in nth2-coq1 intergenic region	<i>Saccharomyces cerevisiae</i>	
MUF9 9	+	31221 32036	1	0	272	P35196	258	37.6	hypothetical 32.7 kD protein in nth2-coq1 intergenic region	<i>Saccharomyces cerevisiae</i>	
MUF9 10	-	52290 53447	2	0	338	L47117	337	42.1	late embryogenesis abundant protein (EMBT)	<i>Picea glauca</i>	
MUF9 11	-	54700 56185	2	1	439	L47117	299	42.1	late embryogenesis abundant protein (EMBT)	<i>Picea glauca</i>	
MUF9 12	-	57377 59147	6	0	255	P37528	227	42.3	hypothetical 21.4 kD protein in dacta-sets intergenic region	<i>Bacillus subtilis</i>	
MUF9 13	+	60635 62853	11	0	405	U11029	323	40.2	cosmid C0088	<i>Caenorhabditis elegans</i>	
MUF9 14	+	66449 67669	1	0	407	Z97343	406	46.8	chromosome 4, ESSA 1 contig fragment No. 8	<i>Arabidopsis thaliana</i>	

transcribed regions						
identifier	Direction	Position 5' 3'	Accession	Overlap (nt)	Identity (% nt)	Definition
MUF9 T1	-	11687 12117	AA598152	431	92.7	Lambda-PRL2 cDNA clone 275D8T7
MUF9 T2	+	20125 20796	H37634	509	93.9	Lambda-PRL2 cDNA clone 184H161T7
MUF9 T3	+	21198 21647	U74576	342	96.3	clone GS154 mRNA sequence
MUF9 T4	+	69597 69871	AA651421	450	90.9	Lambda-PRL2 cDNA clone 184H16XP 3'
MUF9 T5	+	74530 74870	T45844	180	94.4	Lambda-PRL2 cDNA clone 132P11T7
MUF9 T6	+	76990 77298	R86818	215	96.3	Lambda-PRL2 cDNA clone 12G11T7
MUF9 T6	+	76990 77298	T48492	241	99.7	Lambda-PRL2 cDNA clone 117P13T7
MUF9 T6	+	76990 77298	T46505	227	86.1	Lambda-PRL2 cDNA clone 134L4T7
MUF9 T6	+	76990 77298	AA395214	309	97.8	Lambda-PRL2 cDNA clone 117P13XP 3'

MXE10 (13474 bp)



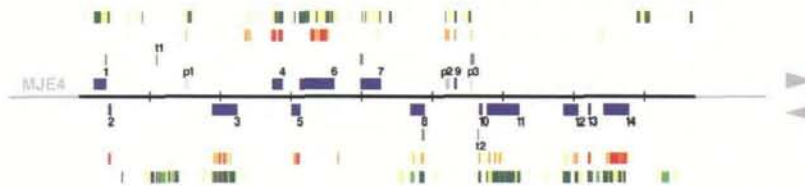
- Grail exon
- Protein db hit
- EST db hit
- Gene
- Gene
- EST db hit
- Protein db hit
- Grail exon

potential protein genes											
identifier	Direction	Position 5'	3'	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MXE10.1	+	6943	8509	3	2	352	S67549	404	20.2	probable membrane protein YOR18ic	<i>Saccharomyces cerevisiae</i>
MXE10.2	+	9545	11234	1	0	563	AC000132	548	26.1	Sequence of BAC F21M12 from chromosome 1	<i>Arabidopsis thaliana</i>
MXE10.3	-	11560	12550	2	4	192	P41227	189	57.1	N-terminal acetyltransferase complex arid1 subunit homolog	<i>Homo sapiens</i>

potential exons											
identifier	Direction	Position 5'	3'	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MXE10.P1	-	13078	13472	2	0	79	S71200	79	100.0	agamous-like protein 15	<i>Arabidopsis thaliana</i>

transcribed regions											
identifier	Direction	Position 5'	3'	Accession	Overlap (nt)	Identity (% nt)	Definition				
MXE10.T1	+	335	699	AA905388	364	99.2	Lambda PRL2 cDNA clone 186J1XP.3				

MXH1 (87210 bp)



- Grail exon
- Protein db hit
- EST db hit
- Gene
- Gene
- EST db hit
- Protein db hit
- Grail exon

potential protein genes											
identifier	Direction	Position 5'	3'	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MXH1.1	+	2073	3810	6	1	442	AC002341	282	30.5	chromosome II BAC T14G11 genomic sequence	<i>Arabidopsis thaliana</i>
MXH1.2	-	4058	4492	1	0	145	P47815	145	61.4	eukaryotic initiation factor 1a	<i>Triticum aestivum</i>
MXH1.3	-	18739	22335	13	0	682	U67717	439	73.6	limbrin/plastin-like	<i>Triticum aestivum</i>
MXH1.4	+	27318	28837	2	0	442	X97864	435	55.2	cytochrome P450	<i>Arabidopsis thaliana</i>
MXH1.5	-	29959	31234	1	0	422	X8130	288	40.3	81kb genomic sequence	<i>Arabidopsis thaliana</i>
MXH1.6	+	31233	36118	10	0	359	Z97336	380	47.0	chromosome 4, ESSA 1 contig fragment No. 1	<i>Arabidopsis thaliana</i>
MXH1.7	+	39844	42721	9	1	360	P16131	342	28.4	ERD1 protein	<i>Saccharomyces cerevisiae</i>
MXH1.8	-	46819	48837	2	2	404	Z97343	278	42.4	chromosome 4, ESSA 1 contig fragment No. 8	<i>Arabidopsis thaliana</i>
MXH1.9	+	58067	53433	1	0	123	AF007270	114	56.8	BAC IG02P16	<i>Arabidopsis thaliana</i>
MXH1.10	-	59551	56998	2	0	123	S91612	114	92.5	beta-1,3-glucanase homolog	<i>Brassica napus</i>
MXH1.11	-	57649	62240	11	0	1042	S75142	1122	29.2	sensory transduction histidine kinase	<i>Homo sapiens</i>
MXH1.12	-	68475	70551	4	0	408	AF907271	429	27.5	BAC TM021B04	<i>Arabidopsis thaliana</i>
MXH1.13	-	71979	72395	1	0	139	Z97339	141	60.3	chromosome 4, ESSA 1 contig fragment No. 4	<i>Arabidopsis thaliana</i>
MXH1.14	-	74148	77788	6	0	459	Z97340	561	25.7	chromosome 4, ESSA 1 contig fragment No. 5	<i>Arabidopsis thaliana</i>

potential exons											
identifier	Direction	Position 5'	3'	No of Exon	No of EST hit	Length (aa)	Accession	Overlap (aa)	Identity (% aa)	Definition	Species
MXH1.P1	+	19129	19272	1	0	48	AF907271	48	77.1	BAC TM021B04	<i>Arabidopsis thaliana</i>
MXH1.P2	+	51842	52390	2	0	149	S65812	196	40.3	RNA-directed DNA polymerase (EC 2.7.7.49)	<i>Halobeta rubescens</i>
MXH1.P3	+	55402	55575	1	3	57	X92976	57	71.9	ZAP1	<i>Arabidopsis thaliana</i>

transcribed regions											
identifier	Direction	Position 5'	3'	Accession	Overlap (nt)	Identity (% nt)	Definition				
MXH1.T1	+	10918	11094	N68876	177	90.1	CD4-16 cDNA clone H1A10T7				
MXH1.T2	-	55327	56532	AA394752	206	92.7	Lambda PRL2 cDNA clone 307CTT7.5				

