

# Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*

The EU *Arabidopsis* Genome Project: M. Bevan<sup>1</sup>, I. Bancroft<sup>1</sup>, E. Bent<sup>1</sup>, K. Love<sup>1</sup>, H. Goodman<sup>2</sup>, C. Dean<sup>1</sup>, R. Bergkamp<sup>3</sup>, W. Dirkse<sup>3</sup>, M. Van Staveren<sup>3</sup>, W. Stiekema<sup>3</sup>, L. Drost<sup>1</sup>, P. Ridley<sup>1</sup>, S.-A. Hudson<sup>1</sup>, K. Patel<sup>1</sup>, G. Murphy<sup>1</sup>, P. Piffanelli<sup>1</sup>, H. Wedler<sup>4</sup>, E. Wedler<sup>4</sup>, R. Wambutt<sup>4</sup>, T. Weitzenegger<sup>5</sup>, T. M. Pohl<sup>5</sup>, N. Terryn<sup>6</sup>, J. Gielen<sup>6</sup>, R. Villarroel<sup>6</sup>, R. De Clerck<sup>6</sup>, M. Van Montagu<sup>6</sup>, A. Lechary<sup>7</sup>, S. Auborg<sup>7</sup>, I. Gy<sup>7</sup>, M. Kreis<sup>7</sup>, N. Lao<sup>8</sup>, T. Kavanagh<sup>8</sup>, S. Hempel<sup>9</sup>, P. Kotter<sup>9</sup>, K.-D. Entian<sup>9</sup>, M. Rieger<sup>10</sup>, M. Schaeffer<sup>10</sup>, B. Funk<sup>10</sup>, S. Mueller-Auer<sup>10</sup>, M. Silvey<sup>11</sup>, R. James<sup>11</sup>, A. Montfort<sup>12</sup>, A. Pons<sup>12</sup>, P. Puigdomenech<sup>12</sup>, A. Douka<sup>13</sup>, E. Voukelatou<sup>13</sup>, D. Milioni<sup>13</sup>, P. Hatzopoulos<sup>13</sup>, E. Piravandi<sup>14</sup>, B. Obermaier<sup>14</sup>, H. Hilbert<sup>15</sup>, A. Düsterhöft<sup>15</sup>, T. Moores<sup>16</sup>, J. D. G. Jones<sup>16</sup>, T. Eneva<sup>17</sup>, K. Palme<sup>17</sup>, V. Benes<sup>18</sup>, S. Rechman<sup>18</sup>, W. Ansorge<sup>18</sup>, R. Cooke<sup>19</sup>, C. Berger<sup>19</sup>, M. Delseny<sup>19</sup>, M. Voet<sup>20</sup>, G. Volckaert<sup>20</sup>, H.-W. Mewes<sup>21</sup>, S. Klosterman<sup>21</sup>, C. Schueller<sup>21</sup> & N. Chalwatzis<sup>21</sup>

<sup>1</sup> Department of Molecular Genetics, John Innes Centre, Colney, Norwich NR4 7UJ, UK

<sup>2</sup> Department of Genetics, Harvard Medical School, Boston, Massachusetts 02144, USA

<sup>3</sup> Department of Molecular Biology, CPRO-DLO, NL 6700 AA Wageningen, The Netherlands

<sup>4</sup> AGOWA GmbH, D 12489 Berlin, Germany

<sup>5</sup> GATC GmbH, D 78467 Konstanz, Germany

<sup>6</sup> Department Genetics, Vlaams Interuniversitair Instituut voor Biotechnologie, Universiteit Gent, B 9000 Gent, Belgium

<sup>7</sup> Institut de Biotechnologie des Plantes, Université de Paris-Sud, ERS/CNRS 569, F-91405 Orsay, France

<sup>8</sup> Department of Genetics, Trinity College, Dublin 2, Eire

<sup>9</sup> SRD GmbH, D 61440 Oberursel, Germany

<sup>10</sup> Genotype GmbH, D 69259 Wilhelmsfeld, Germany

<sup>11</sup> School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK

<sup>12</sup> CSIC, CID, 08034 Barcelona, Spain

<sup>13</sup> Agricultural University of Athens, Athens 118 55, Greece

<sup>14</sup> MediGene AG, D 82152 Planegg/Martinsried, Germany

<sup>15</sup> QLAGEN GmbH, D 4072 Hilden, Germany

<sup>16</sup> The Sainsbury Laboratory, John Innes Centre, Colney, Norwich NR4 7UJ, UK

<sup>17</sup> Max-Delbrück-Laboratorium in der Max-Planck Gesellschaft, 50829 Köln, Germany

<sup>18</sup> European Molecular Biology Laboratory, D 69012 Heidelberg, Germany

<sup>19</sup> UMR CNRS 5545, Université Perpignan, 66860 Perpignan Cedex, France

<sup>20</sup> Laboratory of Gene Technology, Katholieke Universiteit Leuven, B 3001 Leuven, Belgium

<sup>21</sup> Martinsrieder Institut für Protein Sequenzen, Max-Planck Institut für Biochemie, D 82152 Martinsried, Germany

The plant *Arabidopsis thaliana* (*Arabidopsis*) has become an important model species for the study of many aspects of plant biology<sup>1</sup>. The relatively small size of the nuclear genome and the availability of extensive physical maps of the five chromosomes<sup>2-4</sup> provide a feasible basis for initiating sequencing of the five chromosomes. The YAC (yeast artificial chromosome)-based physical map of chromosome 4 was used to construct a sequence-ready map of cosmid and BAC (bacterial artificial chromosome) clones covering a 1.9-megabase (Mb) contiguous region<sup>5</sup>, and the sequence of this region is reported here. Analysis of the sequence revealed an average gene density of one gene every 4.8 kilobases

(kb), and 54% of the predicted genes had significant similarity to known genes. Other interesting features were found, such as the sequence of a disease-resistance gene locus, the distribution of retroelements, the frequent occurrence of clustered gene families, and the sequence of several classes of genes not previously encountered in plants.

A region between markers COP9 and G3845 on the long arm of chromosome 4 from the ecotype Columbia was selected for sequencing. Figure 1 is a map of the position of 389 genes, predicted genes, retroelements and other features in the 1,874,503-base-pair (bp) sequenced region. The mean gene density of one gene every 4,806 bp is consistent with that determined in smaller regions sequenced on other chromosomes<sup>6</sup> (*Arabidopsis thaliana* Database (AtDB) URL: <http://genome-www.stanford.edu/>). Assuming the coding regions of the *Arabidopsis* genome are 100 Mb, the total complement of *Arabidopsis* protein-coding genes can be calculated as ~21,000. Two hundred and seventeen (56%) of the predicted genes are similar to expressed sequence tags (ESTs) at the >95% sequence similarity level. Assuming there are ~12,000 unique *Arabidopsis* ESTs (<http://www.tigr.org/>), an independent assessment of the number of genes can be calculated as  $389/217 \times 12,000 = 21,000$ . This is consistent with the assessment based on gene density and predicted genome size.

Although the region sequenced here may be too small to detect chromosome-wide periodicity in features such as gene density, this feature varied between 7 genes in a 65-kb region (3400c to 3430w) to 12 genes in an adjacent 25-kb region (3435c to 3490c). Eighty-five per cent of predicted and experimentally determined genes contained multiple introns (from 1 to 29) which had no obvious distinguishing features apart from consensus donor and acceptor sites. Introns are 66.48% A+T compared to 55.96% A+T in exons, whereas intergenic regions are 67.77% A+T. There are no regions where experimentally determined genes overlap, but there are three instances of hypothetical genes encoded on the opposite strand to an experimentally determined gene. Putative promoter regions were often smaller than 200 bp, indicating that regulatory sequences may be found frequently in coding regions of other genes.

The degree of similarity between sequenced genes and those in the databases, assessed by their FASTA scores<sup>7</sup>, was used to classify genes. These classifications are shown in Table 1. Class 1 comprises 19 genes which had been sequenced previously; class 2 matches contain 73 genes that are highly similar (>1/3 FASTA self-score) to other genes, mostly from plants. Class 3 matches (242 predicted genes, 65%) comprise genes encoding proteins with a range of similarities to proteins of known function. The putative cellular

**Table 1** Classes of similarities to genes

Class	FASTA score	Type of matching protein	Number	Predicted function
1	Identical	Same protein	19	18
2	>1/3 self score	Known protein	73	70
3a	<1/3 self score-150	Known protein	124	108
3b	150-80	Known protein	118	13
4a	>1/3 self score-150	Hypothetical protein	26	0
4b	150-80	Hypothetical protein	22	0
5		None but has EST match	0	0
6		None	7	
Total			389	209

Predicted genes were ordered into six classes, based on their degree of similarity to known or hypothetical proteins. The degree of similarity was assessed by comparing the self-FASTA score of the predicted protein sequence to the FASTA score resulting from a comparison of the predicted protein and its closest homologue. The value of the self-score relative to the comparative score compensates for differing lengths of the peptide sequences. The number of predicted genes in each class, and the number of predicted genes with assigned cellular roles, are also shown.

**Table 2 Functional catalogue of plant genes**

01 Metabolism	02 Energy	03 Cell growth/division	04 Transcription
01.01 Amino acid	02.01 Glycolysis	03.01 Cell growth	04.01 rRNA synthesis
01.02 Nitrogen and sulphur	02.02 Gluconeogenesis	03.13 Meiosis	04.10 tRNA synthesis
01.03 Nucleotides	02.07 Pentose phosphate	03.16 DNA synth/replication	04.19 mRNA synthesis
01.04 Phosphate	2.10 TCA pathway	03.19 Recombination/repair	04.1901 General TFs
01.05 Sugars and polysaccharides	2.13 Respiration	0.322 Cell cycle	0.41904 Specific TFs
01.06 Lipid and sterol	2.16 Fermentation	03.25 Cytokinesis	04.1907 Chromatin modification
01.07 Cofactors	2.20 E-transport	03.26 Growth regulators	04.22 mRNA processing
	2.30 Photosynthesis	03.99 Other	04.31 RNA transport
05 Protein synthesis	06 Protein destination and storage	07 Transporters	08 Intracellular traffic
05.01 Ribosomal proteins	06.01 Folding and stability	07.01 Ions	08.01 Nuclear
05.04 Translation factors	06.04 Targetting	07.07 Sugars	08.02 Chloroplast
05.07 Translation control	06.07 Modification	07.10 Amino acids	08.04 Mitochondrial
05.10 tRNA synthases	06.10 Complex assembly	07.13 Lipids	08.07 Vesicular
05.99 Others	06.13 Proteolysis	07.16 Purine/pyrimidines	08.10 Peroxisomal
	06.20 Storage proteins	07.22 Transport ATPases	08.13 Vacuolar
		07.25 ABC-type	08.16 Extracellular
		07.99 Others	08.19 Import
			08.99 Others
09 Cell structure	10 Signal transduction	11 Disease/defence	20 Secondary metabolism
09.01 Cell wall	10.01 Receptors	11.01 Resistance genes	20.1 Phenylpropanoids/ phenolics
09.04 Cytoskeleton	10.04 Mediators	11.02 Defence-reglated	20.2 Terpenoids
09.07 ER/Golgi	10.0404 Kinases	11.03 Cell death	20.3 Alkaloids
09.10 Nucleus	10.0407 Phosphatases	11.04 Cell rescue	20.4 Non-protein amino acids
09.13 Chromosomes	10.0410 G proteins	11.05 Stress responses	20.5 Amines
09.16 Mitochondria	10.99 Others	11.06 Detoxification	20.6 Glucosinolates
09.19 Peroxisome		11.99 Others	20.99 Others
09.25 Vacuole			
09.26 Chloroplast			
09.99 Others			
12 Unclear classification	13 Unclassified	Transposons	
		14.01 LTR retroelements	
		14.02 Non-LTR retroelements	
		14.99 Other	

Fifteen putative cellular roles for genes in plants are shown that are broadly based on the yeast functional catalogue<sup>9</sup>. New categories, adapted for plant-specific roles, have been made for secondary product metabolism and disease, defence and stress responses, and several new subcategories of genes involved in photosynthesis, chloroplast structure, storage processes and polysaccharide metabolism have been established. Categories between 14–19 have been reserved for gene categories in other organisms.

roles of most of the *Arabidopsis* genes encoding proteins with significant amino-acid similarity (FASTA > 150) over their entire length, or a high degree of conserved sequences over functional domains, were determined. Class 4 matches comprise 49 predicted genes (12%) similar to proteins of unknown function. No predicted genes were found in class 5 matches, representing predicted genes with cognate ESTs having no significant matches to any protein. Finally, class 6 matches were found for seven predicted genes that had no EST matches and no matches to sequenced genes. These genes are questionable because they are supported by neither experimental data nor similarity to other proteins. In total, the putative cellular roles of 54% of the predicted genes were established by sequence similarity to genes of known function in plants and other organisms. Plant genes accounted for 65% of the significant similarities, whereas cross-phylum matches with vertebrates (12%), bacteria and Archaea (10%) and yeasts (8%) accounted for most of the remainder. This distribution of matches emphasizes the extremely distant evolutionary relationship between the plant kingdom and other phyla.

Genes with predicted or known functions were classified into 15 putative cellular roles described in Table 2, which are based on the functional catalogues established for *Escherichia coli*<sup>8</sup> and yeast<sup>9</sup>. The proportion of genes in each role category is shown in Fig. 2. Of the 206 genes analysed, the largest number were involved in primary and secondary metabolism (32%), reflecting the complex photoautotrophic metabolism of plants. The 14% of genes involved in disease and defence responses may not be representative, because of the cluster of eight putative resistance genes at the *CHPR* gene cluster. The high proportion of genes involved in information processing (transcription 15%, and signal transduction 8%) are typical of complex multicellular organisms<sup>9</sup>. Table 3 shows the

predicted genes, their putative cellular role, the most closely related homologue, FASTA scores and cellular role category.

Five classes of repeat were encountered in the 1.87-Mb region: DNA sequence repeats in non-coding regions, retroelements, chloroplast DNA fragments, dispersed members of protein-coding families, and clustered repeats of related protein-coding genes. Together they comprise ~19% of the sequence. Several types of retroelement are found, in three configurations, in six places. First, two examples of a single copy of long terminal repeat (LTR) element, most similar to maize *hopscotch*<sup>10</sup> (4045w and 4760c), were found distant from protein-coding genes, although 5045w is

**Figure 1** This map shows the positions of genes, predicted genes and other features. A more detailed interactive representation is available via URL <http://www.mips.biochem.mpg.de/mips/athaliana/>. The annotated sequence has accession numbers Z97335–Z97344 inclusive. Predicted genes are numbered, starting at ATDL3000w on the proximal end of the contig, according to their position relative to the centromeric repeats which begin 2.3 Mb north of ATDL3000w. This allows for a minimum of 500 genes in this as-yet unsequenced region, with five digits to represent each predicted gene and any different versions. Genes are named according to standard conventions: AT, *Arabidopsis thaliana*; D, chromosome 4; L, long arm; w or c refers to the strand that encodes each protein. Genes are represented by rectangles pointing in the direction of transcription, starting at the ORF at the beginning of the predicted gene. The classes of matches of the predicted genes described in Table 1 are shown as different colours: class 1 genes are green, class 2 are turquoise, class 3a are pink, class 3b are red, class 4a and 4b are blue, and class 6 are dark blue. LTR repeats are blue pointed lines, non-genic sequence repeats are green pointed lines, tRNAs are red pointed lines, chloroplast DNA homologies are black pointed lines, and cognate cDNAs are turquoise pointed lines. The scale unit is 10 kb.

adjacent to a fragment of a retroelement, 4050w. Second, retroelements are found as adjacent pairs, such as 3275w, a TA1-3-like LTR retroelement<sup>11</sup>, and 3275w, a TA11-1-like non-LTR retroelement<sup>11</sup>, which are also distant from other genes. Finally, there are four examples of retroelement insertions in protein-coding genes. A TA11-1-like non-LTR retroelement (3835c) is inserted into the 5' end of the splicing factor homologue 3830c, and a remnant retroelement with homology over the finger protein (3840c) is adjacent to this element. A TA1-2-like non-LTR element, 3970c, is inserted close to the 3' end of a sesquiterpene cyclase gene (3975c). There are two insertions in *CHPR* putative disease-resistance genes described below. This pattern of dispersed single and pairs of retroelements in the low copy region of an *Arabidopsis* chromosome contrasts with the pattern of retroelements found in the larger genomes of plants such as maize, where retroelements of LTR- and non-LTR type form nested structures of multiple elements of different types that comprise 50% of the 2,400 Mb maize genome<sup>12</sup>, in contrast to the total interspersed middle repetitive DNA that comprised 10% of the *Arabidopsis* genome<sup>11</sup>. The expansion of retroelement numbers is proposed as the principal mechanism for increasing genome size in higher plants<sup>13</sup>.

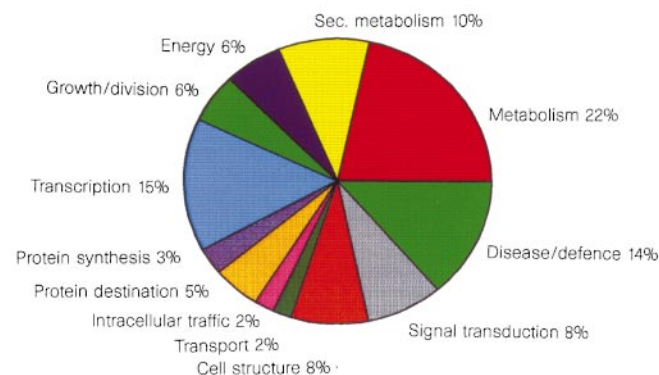
Nearly 20% of the predicted protein-coding genes are members of gene families. The largest families include five predicted indole acetic acid glycosyl transferase genes, seven cytochrome P-450 proteins clustered in an 80-kb region, and a family of five closely related glutaredoxin genes clustered in 15 kb on the same DNA strand. Eight of the nine families comprise pairs of genes that are also adjacent on the same strand, and members of seven of these families were highly related. The frequent observation of close similarities among adjacent members of gene families on the same strand indicates that simple duplication and subsequent divergence may be a common mechanism for expanding gene families in *Arabidopsis*.

A 90-kb region encodes eight genes (4460c–4510c, 4525w) similar to the *Arabidopsis* ecotype Landsberg *erecta* (*Ler*) disease-resistance gene *RPP5* that confers resistance to the oomycete fungus *Peronospora parasitica* race Noco2 (ref. 14). The ecotype Columbia does not contain a functional *RPP5* gene; therefore the genes are named Columbia homologues of *Peronospora* resistance (*CHPR*). Two genes, 4490c and 4510c, contain frameshift mutations that would encode truncated, non-functional proteins, and 4500c lacks an in-frame ATG initiator codon. Two other *CHPR* genes contain retrotransposon insertions with terminal repeats in their coding regions. The Columbia-0 *RPP4* fungus-resistance gene maps to this locus<sup>14</sup>; therefore the two remaining genes, 4470c and 4505c (which has a cognate EST) could encode proteins conferring this resistance. A truncated kinase gene (4515c) is adjacent to intact putative resistance genes. It has a precise deletion of the first exon, compared to the intact *Ler* kinase orthologue (T.M. and J.D.G.J., unpublished), and fragments of the truncated kinase are found at the 3' end of all the *CHPR* genes, except for the truncated 4525w gene on the opposite strand. These fragments may be generated and amplified by recombination events within the putative resistance gene cluster. Putative resistance genes are found as single copies elsewhere in the sequenced region: 3600w is closely related to the *Solanum pimpinellifolium* Cf2.2 fungus-resistance gene, 3225c is related to the tobacco *N* TMV-resistance gene, and 3345c is closely related to the *Arabidopsis* *RPS2* fungus-resistance gene. The precise disease resistance specificities of these genes, if any, are not known.

Five regions similar to the chloroplast genome were found: four were between 85–268 bp and were 70–80% similar, and a 2-kb tract, found between gene models 3690c and 3695c, has 80–94% similarity to chloroplast ribosomal protein L12 and tRNA<sup>Pro</sup>. Finally, 53 mono-, di- and trinucleotide repeats between 20–50 bp were found, most commonly in introns.

A variety of genes were found that have either not yet been encountered in plants, or for which additional new examples will be of interest because they reveal conservation of additional cellular functions among sequenced organisms. Two genes have significant similarity to the cotton *celA1* (ref. 15) gene and may encode putative cellulose synthases. Two hydroxynitrile lyase homologues, 3595w and 4370c, were found that probably catalyse the first step in HCN production from cyanogenic glycosides. Many plants exhibit cyanogenesis as a deterrent to herbivores, but it was not known that *Arabidopsis* was among this group. Terpene cyclases catalyse the cyclization of allylic diphosphate substrates, and participate in a key regulatory step in the complex pathway of isoprenoid synthesis<sup>16</sup>. The adjacent genes 4390w and 4395w encode homologues of the monoterpene cyclase limonene cyclase, 3975c encodes a sesquiterpene cyclase homologue, and the adjacent genes 3715c and 3730c encode homologues of the triterpene cyclase lupeol cyclase. Gene 3390w encodes a protein with high homology to the carboxy-terminal half of HsORC1, the largest component of the DNA replication complex. An *Arabidopsis* complementary DNA encoding ORC2, another component of the putative initiator complex, has previously been characterized<sup>17</sup>. The presence of two of the possible six components of a complex that recognizes origins of DNA replication adds to the evidence for a replicon model of chromosome replication in *Arabidopsis*. 4725w encodes a protein that is 31% similar to of the *Drosophila* SPE1 protein, a DNA-mismatch repair enzyme related to the bacterial *mutS* and human *MSH2* mismatch repair superfamily<sup>18</sup>.

Among the genes encoding proteins involved in information processing, only 15% were found as cognate ESTs, reflecting the relatively low abundance of transcripts and possible differential expression of these genes. In this class, 4235c is notable as it encodes a 434-amino-acid protein 40% identical to yeast ADA2. ADA2 is required for the function of acidic activation domains of transcription factors in yeast, where it forms a complex with GCN5, a histone *N*-acetyltransferase<sup>19</sup>. The *Arabidopsis* HOOKLESS protein, involved in the late steps of ethylene signal transduction, is a member of the GCN5-related *N*-acetyltransferase (GNAT) superfamily<sup>20</sup>, therefore it is possible that a protein complex analogous to the yeast ADA2–GCN5 transcription adaptor–chromatin modification complex may be found in *Arabidopsis*. Seven genes highly similar to genes involved in RNA processing were found. The



**Figure 2** The pie chart shows the proportion of predicted genes with assigned cellular roles in each of the functional categories described in Table 2. A total of 209 genes were assigned to functional categories.

**Table 3** A summary of genes with assigned cellular roles, based on significant similarities to genes of known function. The gene identifications are the same as in Fig. 1. The putative gene identity, closest homologue, their FASTA scores, and the functional category (described in Table 2), are shown. A more extensive list of homologies and scores is available through the PEDANT database at URL <http://www.mips.biochem.mpg.de/mips/athaliana/>.

high degree of similarity of the *Arabidopsis* homologues to the human RNA helicase 1 gene (4365c) and to the yeast *SUV3* ATP-dependent RNA helicase (3435c) may indicate significant functional conservation of RNA processing mechanisms among eukaryotes. Two homologues of kinesins from *Caenorhabditis elegans* and *Xenopus*, 3115c and 3205c, show sequence conservation over the three functional domains, indicating that mechanisms of intracellular motility may be conserved between *C. elegans*, vertebrates and plants. Mutation of a putative Ca<sup>2+</sup>-binding kinesin called ZWI-CHEL caused morphological defects in trichomes<sup>21</sup>, demonstrating that this class of motor molecule contributes to cell shape in plants. Twelve per cent of the predicted genes found in the FCA region encode potential components of signal transduction pathways. For example, a homologue of a *C. elegans* calcium-sensor protein related to *Drosophila* frequenin, a calcium-binding protein that modulates neuronal efficiency<sup>22</sup>, is encoded by 4205c. Another *Arabidopsis* homologue of the yeast *SNF-1* kinase, which regulates cellular sugar metabolism, is encoded by 3330c. 3280c encodes a protein kinase closely related to a developmentally regulated yeast kinase, *SPS1*, required for spore wall formation<sup>23</sup>.

Four highly significant findings have been made in this pilot-scale sequencing project. First, there is a consistently high gene density over an extended contiguous region, with 389 predicted genes in 1.87 Mb. The relatively high gene density encountered in this region is also found in other sequenced regions. On the basis of this work and the 13 Mb of available sequence from four chromosomes, and the size of YAC contigs covering most of the low-copy regions of the five chromosomes, the total number of protein-coding genes is probably about 21,000. Second, the genome sequence has a high information content: 54% of the predicted genes can be assigned cellular roles on the basis of enzymatic, structural or other functions inferred by sequence similarity to proteins of known function. Nevertheless, the specific functions of most of these genes in plants requires further analysis. The remaining 46% of predicted genes, which either have no significant similarities to other genes or are similar to genes of unknown function, require extensive systematic experimentation to determine their cellular roles. Third, nearly 20% of the predicted genes are members of gene families that may have arisen by gene duplication and divergence. In other sequenced eukaryotes, such as *C. elegans* and yeast, the proportion is not as high. If the number of gene families in *Arabidopsis* is found to be ~15,000 after more comprehensive sequencing, *Arabidopsis* will have a similar-sized genome complement to the model metazoans, *Drosophila*<sup>24</sup> and *C. elegans*<sup>25</sup>. This may represent a minimal number of genes required for the function of complex multicellular organisms with highly diverged mechanisms of development and environmental interactions<sup>26</sup>. Finally, it is now clear that a straightforward shotgun-sequencing strategy can generate contiguous sequence from nearly all of the low-copy regions of the *Arabidopsis* genome. □

## Methods

**Contig assembly.** The physical map of *Arabidopsis* chromosome 4 is represented by 4 YAC contigs covering 17 Mb (ref. 2). Sequencing was initiated at the FCA locus on the 13.5-Mb lower arm. The YAC map was used to assemble two cosmid contigs containing 1.2 Mb, using subcloning of YACs into cosmids in order to complete regions not represented in genomic cosmid libraries<sup>5</sup>. BAC clones were identified by hybridization to YAC clones and were assembled into the contigs to complete coverage<sup>27</sup>.

**Sequencing strategy.** Sequence was determined from an overlapping contiguous set of 6 BAC clones, 15 Lorist cosmids, 16 CC cosmids, 5 CAT cosmids, and 45 YAC subclones in the binary cosmid vector pCLD 04541. All libraries were prepared using DNA isolated from the Columbia ecotype. Clones were distributed in a network of 17 EU labs for sequencing, and the resulting sequence was assembled at the Martinsrieder Institut für Protein Sequenzen (MIPS). The quality controls used during sequence assembly involved comparison of 220,134 bp of overlap sequences, resequencing of selected

regions (90,566 bp), and resequencing of suspected low accuracy regions (8,684 bp), such as those harbouring potential frameshifts revealed by gene modelling. The total sequence produced was 2,094,637 bp, and the total non-redundant sequence was 1,874,503 bp. Shotgun sequencing of cosmid and BAC clones was the most common sequencing strategy.

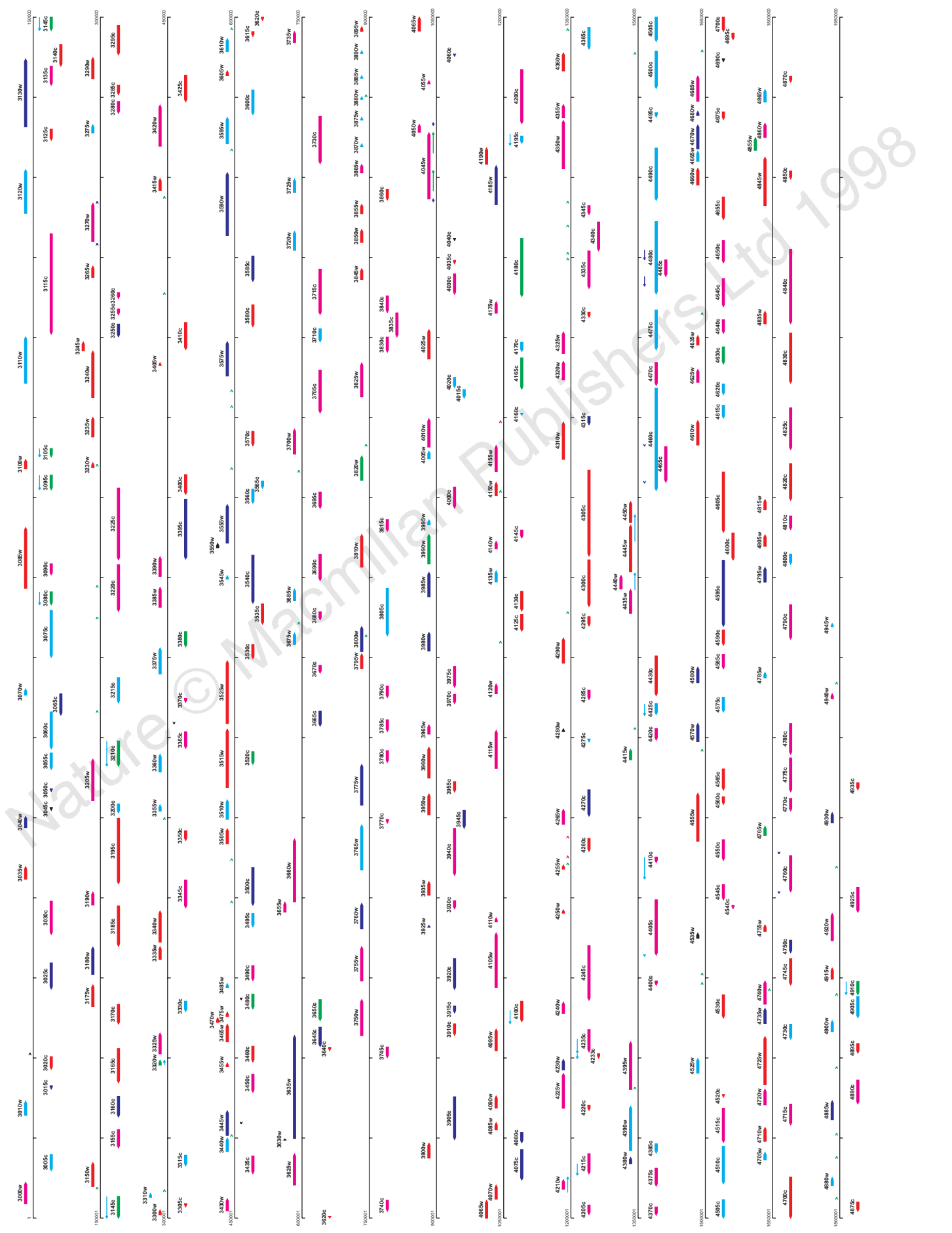
**Sequence analysis.** An initial BLASTX analysis<sup>28</sup> was used to compare all reading frames with all protein sequences and separately with the translations of *Arabidopsis* and other plant ESTs. A search for tRNAs and repeats was also made. Genefinder, Genmark and XGRAIL were modified using published *Arabidopsis* sequence and used for the identification of *Arabidopsis* genes. NetPlantGene was used for recognition of splice sites<sup>29</sup>. Where possible, the predictions were checked for consistency with known protein sequences or cognate cDNA sequences, and the gene models adjusted accordingly. A total of 21,031 bp of cognate cDNA sequences were produced to aid modelling. The resulting protein sequences were extracted using FINDORFS for further analysis.

Received 8 August; accepted 31 October 1997.












- Meyerowitz, E. M. & Somerville, C. R. (eds) *Arabidopsis* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1994).
- Schmidt, R. *et al.* Physical map and organization of *Arabidopsis* chromosome 4. *Science* **270**, 480–483 (1995).
- Zachgo, E. A. *et al.* A physical map of chromosome 2 of *Arabidopsis thaliana*. *Genome Res.* **6**, 19–25 (1996).
- Schmidt, R., Love, K., West, J., Lenehan, Z. & Dean, C. Description of 31 YAC contigs spanning the majority of *Arabidopsis thaliana* chromosome 5. *Plant J.* **11**, 563–572 (1997).
- Bancroft, I. *et al.* A strategy involving the use of high-redundancy YAC subclone libraries facilitates the contiguous representation in cosmid and BAC clones of 1.7 Mb of the genome of *Arabidopsis thaliana*. *Weeds World* **4**, 1–9 (1997).
- Sato, S. *et al.* 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 (1997).
- Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988).
- Riley, M. Functions of gene products in *E. coli*. *Microbiol. Rev.* **57**, 862–952 (1993).
- Mewes, H.-W. *et al.* Overview of the yeast genome. *Nature* **387** (suppl.) 7–84 (1997).
- White, S. E., Habera, L. F. & Wessler, S. R. Retrotransposons in the flanking regions of normal plant genes: A role for copia-like elements in the evolution of gene structure and expression. *Proc. Natl Acad. Sci. USA* **91**, 11792–11796 (1994).
- Konieczny, A., Voytas, D. F., Cummings, M. P. & Ausubel, F. M. A superfamily of *Arabidopsis thaliana* retrotransposons. *Genetics* **127**, 801–809 (1991).
- SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768 (1996).
- Wessler, S. R., Bureau, T. E. & White, S. E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5**, 814–821 (1995).
- Parker, J. E. *et al.* The *Arabidopsis* downy mildew resistance gene *RPP5* shares similarity to the Toll and interleukin-1 receptors with *N* and *L6*. *The Plant Cell* **9**, 879–894 (1997).
- Pear, J. R., Kawagoe, Y., Sreckengost, W. E., Delmer, D. P. & Stalker, D. M. Higher plants contain homologues of the bacterial *celA* genes encoding the catalytic subunit of cellulose synthase. *Proc. Natl Acad. Sci. USA* **93**, 12637–12642 (1996).
- Back, K. & Chappell, J. Cloning and bacterial expression of a sesquiterpene cyclase from *Hyoscamus muticus* and its molecular comparison to related terpene cyclases. *J. Biol. Chem.* **270**, 7375–7381 (1995).
- Gavin, K. A., Hidaka, M. & Stillman, B. Conserved initiator proteins in eukaryotes. *Science* **270**, 1667–1671 (1995).
- Fishel, R. *et al.* The human mutator gene homologue *MHS2* and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–1038 (1993).
- Marcus, G. A., Silverman, N., Berger, S. L., Horiuchi, J. & Guarente, L. Functional similarity and physical association between GCN5 and ADA2: putative transcriptional adaptors. *EMBO J.* **13**, 4807–4815 (1994).
- Neuwald, A. F. & Landsman, D. GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein. *Trends Biochem. Sci.* **22**, 154–155 (1997).
- Oppenheimer, D. G. *et al.* Essential role for a kinesin-like protein in *Arabidopsis* trichome morphogenesis. *Proc. Natl Acad. Sci. USA* **94**, 6261–6266 (1997).
- Pongs, O. *et al.* Frequenin—a novel calcium-binding protein that modulates synaptic efficacy in the *Drosophila* nervous system. *Neuron* **11**, 15–28 (1993).
- Friesen, H., Lunz, R., Doyle, S. & Segall, J. Mutation in the *SPS1*-encoded protein kinase of *Saccharomyces cerevisiae* leads to defects in transcription and morphology during spore formation. *Genes Dev.* **8**, 2162–2175 (1994).
- Gabor Miklos, G. & Rubin, G. M. The role of the genome project in determining gene function: Insights from model organisms. *Cell* **86**, 521–529 (1996).
- Wilson, R. *et al.* 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368**, 32–38 (1994).
- Meyerowitz, E. M. Plants and the logic of development. *Genetics* **145**, 5–9 (1997).
- Bent, E., Johnson, S., Bancroft, I. BAC representation of two low-copy regions of the genome of *Arabidopsis thaliana*. *The Plant J.* (in the press).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24**, 3439–3452 (1996).

**Acknowledgements.** This work was initiated and sponsored by the European Commission, DG-XII Life Sciences. Additional support from the BBSRC Plant and Animal Genome Analysis Programme, GREG (Groupe de Recherche et d'Etude des Genomes), BioResearch Ireland, and Plan Nacional de Investigación Científica y Técnica is gratefully acknowledged.

Correspondence and requests for materials should be addressed to M.B. (e-mail: bevan@bbsrc.ac.uk).



Key:

-  class 1: proteins of known function
-  class 2: strong similarity to proteins of known function (more than one third of self-score)
-  class 3a: weak similarity to proteins of known function (FASTA score > 150)
-  class 3b: weak similarity to proteins of known function (FASTA score ≤ 150)
-  class 4a/b: weak similarity to proteins of unknown function (FASTA score > 150/≤ 150)
-  class 6: questionable ORF
  
-  non-genic sequence repeat
-  cognate cDNA
-  tRNA
-  LTR repeat
-  similarity to chloroplast genome

Nature © Macmillan Publishers Ltd 1998

**Table 3. A summary of genes with assigned cellular roles based on significant similarities to genes of known function**

GENE	IDENTITY	FUN.CAT.	HOMOLOG	FASTA	SELF FASTA
ATD13005C	SERINE HYDROXYMETHYL TRANSFERASE -LIKE	1.01	A42906 PEA SERINE HYDROXYMETHYL TRANSFERASE	1355	2339
ATD13010W	ADENOSYLOHOMOXYSTEINASE	1.01	SAHH_MEDSA ADENOSYLOHOMOXYSTEINASE	2290	2411
ATD13480C	CYSTEINE SYNTHASE	1.01	S65533 ARATH. CYSTEINE SYNTHASE	1473	1473
ATD13495C	HIS7	1.01	HIS7_ARATH.IMPIDAZOLEGLYCEROL-PHOSPHATE DEHYDRATASE	1001	1183
ATD14945W	ACETYLORNITHINE DEACETYLASE -LIKE	1.01	ARGE_DICDI DICTYOSTELIUM ACETYLORNITHINASE	212	602
ATD13380C	ATP SULPHURYLASE	1.02	S68201 ARATH. ATP SULPHURYLASE	2292	2299
ATD13110W	DNA CYTOSINE 5 ME TRANSFERASE -LIKE	1.03	S59804 ARATH. DNA CYTOSINE 5 METHYL TRANSFERASE	5516	7550
ATD14715C	PUR U -LIKE	1.03	S74352 SYNTRICHOCYSTIS PUR U	422	1497
ATD13080C	BETA 1,3 GLUCANASE	1.05	S31906 ARATH. BETA 1,3 GLUCANASE	2401	2401
ATD13105C	XYLOGLUCAN ENDOTRANSGLYCOSYLASE	1.05	S71226 ARATH. XTR7 XYLOGLUCAN ENDOTRANSGLYCOSYLASE	1515	1520
ATD13690C	BETA AMYLASE	1.05	S36094 ARATH. BETA AMYLASE	2533	2549
ATD13675W	UTP-GLUCOSE GLUCOSYL TRANSFERASE -LIKE	1.05	S49150 CASSAVA UTP-GLUCOSE GLUCOSYL TRANSFERASE	1005	2241
ATD13680C	UTP-GLUCOSE GLUCOSYL TRANSFERASE -LIKE	1.05	S41950 CASSAVA UTP-GLUCOSE GLUCOSYL TRANSFERASE	389	1553
ATD13685W	UTP-GLUCOSE GLUCOSYL TRANSFERASE -LIKE	1.05	S49150 CASSAVA UTP-GLUCOSE GLUCOSYL TRANSFERASE	1116	2390
ATD13690C	CELLULOSE SYNTHASE -LIKE	1.05	GHU58283_1 COTTON CELA1 CELLULOSE SYNTHASE	307	3774
ATD13795C	CELLULOSE SYNTHASE -LIKE	1.05	GHU58283_1 COTTON CELA1 CELLULOSE SYNTHASE	342	3402
ATD14030C	PECTIN METHYLESTERASE -LIKE	1.05	PC4168 ARATH. PECTIN METHYLESTERASE	980	3333
ATD14105W	GALACTOKINASE -LIKE	1.05	S27988 HAEMOPHILUS GALACTOKINASE	330	5102
ATD14170C	BETA 1,3 GLUCANASE -LIKE	1.05	S65077 PARA RUBBER BETA 1,3 GLUCANASE	910	1784
ATD14320W	GLUCOSYL TRANSFERASE -LIKE	1.05	CF10810_8 C.ELEGANS GLUCOXYL GLUCOSYL TRANSFERASE	288	4627
ATD14325W	GLYCOPENIN -LIKE	1.05	Z72514_A_C.ELEGANS T10B10.8	289	2341
ATD14575C	BETA AMYLASE -LIKE	1.05	D50866 SOYBEAN BETA AMYLASE (EC 3.2.1.2)	1178	2492
ATD14625W	BETA 1,3, GLUCANASE -LIKE	1.05	S20668 ARATH. BETA 1,3 GLUCANASE	554	2998
ATD14720W	INOSITOL 2 DEHYDROGENASE -LIKE	1.05	H0511_B.SUTILLUS INOSITOL 2 DEHASE	175	1818
ATD14920W	TREHALOSE 6P SYNTHASE SUBUNIT -LIKE	1.05	U07184 ASPERGILLUS NIGER TREHALOSE 6P SYNTHASE	966	4392
ATD13255C	CARNITINE RACEMASE -LIKE	1.06	I41014 E.COLI CARNITINE RACEMASE	164	1139
ATD13260C	CARNITINE RACEMASE -LIKE	1.06	I41014 E.COLI CARNITINE RACEMASE	181	1124
ATD13355W	ACYLAMINOACYL PEPTIDASE -LIKE	1.06	IU0135_PORCINE ACYLAMINOACYL PEPTIDASE	288	2104
ATD13610W	CTP-PHOSPHOCOLINE CYTIDYLYLTRANS. -LIKE	1.06	U50451 ARATH. CTP PHOSPHOCOLINE CYT TRANSFERASE	479	1394
ATD13765W	FATTY ACID HYDROPEROXIDASE -LIKE	1.06	U67996 ARATH. ALLENE OXIDE SYNTHASE	1571	3889
ATD14010W	2-HYDROXYHEPTA-2,7 DIONATE ISOMERASE -LIKE	1.06	FA6506 2-HYDROXYHEPTA-2,7 DIONATE ISOMERASE	362	1782
ATD14015C	EPOXIDE HYDROLASE -LIKE	1.06	D16268 ARATH. EPOXIDE HYDROLASE	704	1924
ATD14020C	EPOXIDE HYDROLASE -LIKE	1.06	U02495 SCLARIA TURBERGSIUM EPOXIDE HYDROLASE	103	2669
ATD14145C	3-OH BUTYRYL COA DEHYDRATASE -LIKE	1.06	CRT_CLOB CLOSTRIIDIUM CROTONASE	362	1109
ATD14375C	PHOSPHATIDYL SERINE DECARBOXYLASE -LIKE	1.06	A38732 HAMSTER PHOSPHATIDYL SERINE DEC ASE	260	2218
ATD14405C	ACTY- COA OXIDASE -LIKE	1.06	IJ2066 HUMAN ACTY- COA OXIDASE PEROXISOMAL	1462	4508
ATD14425C	ENOLY COA HYDRATASE -LIKE	1.06	S17595 HUMAN ENOLY COA HYDRATASE	150	2181
ATD14435W	TRIACYL GLYCEROL LIPASE -LIKE	1.06	IQ1390 RHIZOPUS TRIACYLGLYCEROL LIPASE	157	2935
ATD14630C	FARNESYL DIPHOSPHOSPHATE SYNTHASE -LIKE	1.06	S71182 ARATH. FARNESYL DIPHOSPHOSPHATE SYNTHASE	1694	1804
ATD14770C	PALMITOYL-PROTEIN THIOESTERASE -LIKE	1.06	U50315_D_C.ELEGANS PALMITOYL COA THIOESTERASE	214	1471
ATD14775C	PALMITOYL-PROTEIN THIOESTERASE -LIKE	1.06	U50315_D_C.ELEGANS PALMITOYL COA THIOESTERASE	469	2810
ATD13145C	PHYTONE DESATURASE	1.07	CRJ1_ARATH. PHYTONE DEHYDROGENASE	1451	2488
ATD14665W	LACTATE DEHYDROGENASE -LIKE	2.01	D13817 RICE LACTATE DEHYDROGENASE	1153	1615
ATD13355W	GERMIN PRECURSOR OXALATE OXIDASE -LIKE	2.13	S71254 ARABIDOPSIS GERMIN TYPE 2	715	1009
ATD13485W	FERRIDOXIN [2S-2S] -LIKE	2.2	FEH1006 NOSTOC FERRIDOXIN	235	730
ATD13870W	GLUTAREDOXIN -LIKE	2.2	S54825 CASTOR BEAN GLUTAREDOXIN	185	514
ATD13875W	GLUTAREDOXIN -LIKE	2.2	S54825 CASTOR BEAN GLUTAREDOXIN	187	507
ATD13880W	GLUTAREDOXIN -LIKE	2.2	S54825 CASTOR BEAN GLUTAREDOXIN	191	508
ATD13885W	GLUTAREDOXIN -LIKE	2.2	S54825 CASTOR BEAN GLUTAREDOXIN	186	504
ATD13890W	GLUTAREDOXIN -LIKE	2.2	S54825 CASTOR BEAN GLUTAREDOXIN	192	503
ATD14000C	UBIQUINOL:CYT. C OXIDOREDUCTASE -LIKE	2.2	Y087296 ALFALFA UBIQUINOL:CYT. C OXIDOREDUCTASE	299	1525
ATD14250W	NADH DEHYDROGENASE -LIKE	2.2	S27171 NEUROSPORA NADH DEHYDROGENASE	122	523
ATD13895C	PYRVATE PHOSPHATE DIKINASE -LIKE	2.3	U02529 ENTAMOEBA PYRVATE PHOSPHATE DIKINASE	1846	4631
ATD13820W	REP PROTEIN	2.3	U27999 ARATH. REP CLONING VECTOR PROTEIN	342	1342
ATD13390W	REPLICATION ORIGIN CONTROL PROTEIN -LIKE	3.16	U43416 HUMAN REPLICATION CONTROL PROTEIN 1	855	3829
ATD13420W	CENTROMERE PROTEIN -LIKE	3.16	S28261 HUMAN CENTROMERE PROTEIN	465	7588
ATD14035C	RECA -LIKE	3.19	IQ0738 BACTEROIDES RECA PROTEIN	92	677
ATD14725W	DNA MISMATCH REPAIR PROTEIN -LIKE	3.19	SPE1_DROME DROSOPHILA SPE1 CHECKER 1	49	2649
ATD13090C	INDOLE 3 ACETATE GLUCOSYLTRANSE -LIKE	3.26	A54793 MAIZE INDOLE 3 ACETATE GLUCOSYL TRANSE	519	2259
ATD13780C	INDOLE 3 ACETATE GLUCOSYL TRF. -LIKE	3.26	A54739 MAIZE INDOLE 3 ACETATE GLUCOSYL TRANSFERASE	681	2484
ATD13785C	INDOLE 3 ACETATE GLUCOSYL TRF. -LIKE	3.26	A54739 MAIZE INDOLE 3 ACETATE GLUCOSYL TRANSFERASE	637	2459
ATD13790C	INDOLE 3 ACETATE GLUCOSYL TRF. -LIKE	3.26	A54739 MAIZE INDOLE 3 ACETATE GLUCOSYL TRANSFERASE	716	2484
ATD13815C	INDOLE 3 ACETATE GLUCOSYL TRF. -LIKE	3.26	A54739 MAIZE INDOLE 3 ACETATE GLUCOSYL TRANSFERASE	457	2262
ATD14410C	GA-20 OXIDASE -LIKE	3.26	U20873 ARATH. GIBBERELLIN 20-OXIDASE GA-5	153	657
ATD14350W	GROWTH REGULATOR -LIKE	3.99	A44236 ARATH. AUXIN-INDEPENDENT GROWTH PROMOTER	1407	4464
ATD13835W	RNA POLYMERASE ENDONUCLEASE -LIKE	4.1	S53416 YEAST SEN1 PROTEIN	266	2466
ATD14685W	ASPARAGINE TRNA LIGASE -LIKE	4.1	B64115 HAEMOPHILUS ASPARAGINE-TRNA LIGASE	404	2147
ATD13300W	RNA POL 5TH SUBUNIT -LIKE	4.19	B44457 SOYBEAN RNAPOL2 5TH SUBUNIT	138	988
ATD13370C	RNA POL 5TH SUBUNIT -LIKE	4.19	B44457 SOYBEAN RNAPOL2 5TH LARGEST SUBUNIT	237	821
ATD14850C	HEAT SHOCK TRANSCRIPTION FACTOR -LIKE	4.1904	U07595 HUMAN TRANSCRIPTION FACTOR TH1 SUBUNIT	197	777
ATD13030C	HEAT SHOCK TRANSCRIPTION FACTOR -LIKE	4.1904	S25480 PERUVIAN TOMATO HEAT SHOCK TRANSCRIPTION FACTOR	556	6013
ATD13245W	G-BOX BINDING TRANSCRIPTION FACTOR -LIKE	4.1904	U18349 BEAN G-BOX BINDING PROTEIN	140	1308
ATD13310W	CCAAT BOX BINDING FACTOR A SUBUNIT -LIKE	4.1904	CBA MAIZE CCAAT-BOX BINDING FACTOR	487	758
ATD13670C	CONSTANS -LIKE	4.1904	S51453 ARATH. CONSTANS FLOWERING TIME TXN FACTOR	218	1448
ATD14115W	TRANSCRIPTION FACTOR -LIKE	4.1904	S48041 PARSLEY GC-1 PROTEIN	108	977
ATD14200C	HOMEOBOX -LIKE	4.1904	U41543 C.ELEGANS LIM HOMEOBOX PROTEIN	226	5499
ATD14235C	TRANSCRIPTIONAL ADAPTOR -LIKE	4.1904	A43252 YEAST PROB. TRANSCRIPTION ADAPTOR ADA2	493	2418
ATD14240W	TRANSCRIPTION FACTOR -LIKE	4.1904	U18349 PHASOLUS PHASEOLIN G-BOX BINDING PROTEIN PG2	317	2253
ATD14400C	AP2-DOMAIN TINY -LIKE	4.1904	X04698 ARATH. TINY PROTEIN	304	489
ATD14415W	ATHR HOMEOIC PROTEIN HAT 4	4.1904	S31424 ARATH. HOMEOIC PROTEIN HAT4	1323	1323
ATD14650C	SCARECROW -LIKE	4.1904	U62978 ARATH. SCARECROW PUTATIVE TRANSCRIPTION FACTOR	404	1834
ATD14765W	HAT 1 HOMEOBOX TRANSCRIPTION FACTOR	4.1904	HAT1_ARATH. HOMEOBOX-LEUCINE ZIPPER HAT1	1331	1331
ATD14780C	EREBP -LIKE	4.1904	S38125 TOBACCO EREBP-4	427	876
ATD14785W	EREBP-2 -LIKE	4.1904	D38126 TOBACCO EREBP-2 TRANSCRIPTION FACTOR	440	876
ATD14890C	GLABROUS 2 -LIKE	4.1904	A53900 ARATH. HOMEOIC PROTEIN GL2	487	3104
ATD14910C	HEAT SHOCK TRANSCRIPTION FACTOR 1	4.1904	S52641 ARATH. HEAT SHOCK TRANSCRIPTION FACTOR	3162	3234
ATD14925C	MYB TRANSCRIPTION FACTOR -LIKE	4.1904	S58280 ARATH. MYB TRANSCRIPTION FACTOR	171	429
ATD14940W	ZN FINGER PROTEIN -LIKE	4.1904	S60325 ARATH. SUPERMAN PROTEIN	216	864
ATD13435C	ATP-DEP. RNA HELICASE -LIKE	4.22	S63453 YEAST SUV3 RNA HELICASE	710	2235
ATD13830C	SPLICING FACTOR -LIKE	4.22	S50096 C.ELEGANS PROB. SPLICING FACTOR	188	2818
ATD13965W	ATP-DEPENDENT RNA HELICASE -LIKE	4.22	A28464 RICE ORYZALIN DEAD BOX PROTEIN	543	2166
ATD14140W	SPLICING FACTOR -LIKE	4.22	S50096 C.ELEGANS PROB. SPLICING FACTOR	246	1411
ATD14180C	FCA FLOWERING TIME GENE	4.22	Z78992 ARATH. FCA GENE	3547	3547
ATD14340C	ATP-DEPENDENT RNA HELICASE -LIKE	4.22	S64750 ATP-DEPENDENT RNA HELICASE DR51	715	3239
ATD14850C	RNA HELICASE 1 -LIKE	4.22	A56236 HUMAN RNA HELICASE	4328	4328
ATD13165C	L2 60S RIBOSOMAL PROTEIN -LIKE	5.01	RI2_TOBACCO 60S RIBOSOMAL PROTEIN L2	137	3653
ATD13200C	RIBOSOMAL PROTEIN L41 -LIKE	5.01	M62396 CANDIDA MALTOZA RIBOSOMAL PROTEIN L41	483	753
ATD13545W	L27 RIBOSOMAL PROTEIN -LIKE	5.01	U10046 PEA RIBOSOMAL PROTEIN L27	613	909
ATD14055W	RIBOSOMAL PROTEIN L19 -LIKE	5.01	S58013 RAT RIBOSOMAL PROTEIN L19	164	544
ATD14355C	RIBOSOMAL PROTEIN L15 -LIKE	5.01	S4800 YEAST RIBOSOMAL PROTEIN L15.E.B	708	977
ATD14790C	RIBOSOMAL PROTEIN L15.E.A -LIKE	5.01	S48502 YEAST RIBOSOMAL PROTEIN L15.E.A	829	1125
ATD14905C	PSII D1 PROTEIN PROCESSING ENZYME -LIKE	6.01	S65416 SPINACH PSII D1 PROCESSING ENZYME	1409	2401
ATD14215C	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE -LIKE	6.07	S45495 S.POMBE ISPA PROTEIN	1114	3648
ATD13440W	SERINE PROTEASE CHAIN C7 -LIKE	6.13	S55400 HUMAN PROTEASOME CHAIN HSC7-1	601	1461
ATD13560C	SERINE PROTEASE -LIKE	6.13	A55800 MUSK MELON CUCUMISIN PRECURSOR	491	1969
ATD13565C	SERINE PROTEASE -LIKE	6.13	A55800 MUSK MELON CUCUMISIN PRECURSOR	515	1000
ATD13755W	UBIQUITIN DEGRADATION PATHWAY -LIKE	6.13	S59814 YEAST UFD1 PROTEIN	280	3762
ATD14275C	CYSTEINE PROTEINASE INHIBITOR -LIKE	6.13	A28464 RICE ORYZALIN CYSTEINE PROTEINASE INHIBITOR	240	574
ATD14345C	METALLOENDOPEPTINASE -LIKE	6.13	A41820 SOYBEAN METALLOPEPTINASE	466	1793
ATD14885C	SERINE PROTEASE -LIKE	6.13	A34614 HUMAN PLACENTAL SERINE PROTEASE	274	1135
ATD14790C	UBIQUITIN CARBOXY TERMINAL PROTEINASE -LIKE	6.13	A40085 HUMAN UBIQUITIN C-TERM PROTEASE	396	2193
ATD14850C	SUGAR TRANSPORTER -LIKE	7.07	Z46381_G_C.ELEGANS SUGAR TRANSPORT PROTEIN	797	2464
ATD14810C	GLYCEROL-3-PHOSPHATE PERMEASE -LIKE	7.07	HAEMOPHILUS GLYCEROL-3-PHOSPHATE PERMEASE	325	2191
ATD13660W	ABC TRANSPORTER -LIKE	7.25	Z70524 SPIRODELA PDR5-LIKE ABC TRANSPORTER	350	1462
ATD14705W	AQUAPORIN -LIKE	7.99	X93953 HELIANTHUS AQUAPORIN	1056	1184
ATD14130W	CHLOROPLAST PORE PROTEIN -LIKE	8.02	Z73533 PEA CHLOROPLAST PORE PROTEIN	192	711
ATD13120W	SEC 23 -LIKE	8.07	X97064 HUMAN SEC23 PROTEIN	2060	3814
ATD13930C	SYNAPTOBREVIN -LIKE	8.07	SYBR ARATH. SYNAPTOBREVIN-RELATED PROTEIN	191	816
ATD14900W	SYNTAXIN -LIKE	8.07	L41651 ARATH. SYNTAXIN	654	2209
ATD13490C	HYDROXYPROLINE-RICH GLYCOPROTEIN -LIKE	9.01	S06733 TOBACCO HYDROXYPROLINE RICH GLYCOPROTEIN	191	2479
ATD13625W	CELL WALL PROTEIN -LIKE	9.01	S71558 RAPE CELL WALL MEMBRANE LINKING PROT.	725	2202
ATD13770C	GLYCINE-RICH PROTEIN -LIKE	9.01	X95262 TOMATO TFMS GENE	194	689
ATD14110W	EXTENSIN -LIKE	9.01	X91836 VIGNA EXTENSIN CLASS 1	178	917
ATD14155W	APG CELL WALL PROTEIN -LIKE	9.01	S21961 ARATH. PROLINE-RICH PROTEIN APG	257	1743
ATD14645C	RYGANS ALLERGEN -LIKE	9.01	S18614 RYGANS ALLERGEN L01.P1	300	1433
ATD14645C	TRICHOHYALIN -LIKE	9.01	S28589 RABBIT TRICHOHYALIN	200	2469
ATD13115C	KINESIN -LIKE	9.04	X LAEVIS KLP2 KINESIN PROTEIN	558	7850
ATD13205W	KINESIN -LIKE	9.04	S45351 C.ELEGANS KINESIN OSM-3	371	4473
ATD13220C	ANKYRIN 2 -LIKE	9.04	S37431 HUMAN ANKRYN2 LONG FORM	181	4813
ATD14285C	MICROTUBULE-ASSOCIATED PROTEIN 1 LIGHT CHAIN 3 -LIKE	9.04	A53624 MICROTUBULE-ASSOCIATED PROTEIN 1 LIGHT CHAIN 3	167	808

GENE	IDENTITY	FUNCAT.	HOMOLOG	FASTA	SELF FASTA
ATDL4640C	MYOSIN HEAVY CHAIN ATPASE -LIKE	9.04	A24922 RAT EMBRYO MYOSIN HEAVY CHAIN ATPASE	227	2353
ATDL3415W	ACROSIN -LIKE	9.04	S72273 BOVINE ACTIN DEPOLYMERISING PROTEIN	121	1510
ATDL3520C	TUBULIN ALPHA-6 CHAIN	9.04	Q1597 ARATH. TUBULIN ALPHA-6 CHAIN	2230	2230
ATDL4005W	DYNEIN LIGHT CHAIN -LIKE	9.04	A56444 CHLAMYDOMONAS DYNEIN LIGHT CHAIN	167	495
ATDL3190W	RIBONUCLEOPROTEIN -LIKE	9.13	S40778 XENOPUS RIBONUCLEOPROTEIN	505	1981
ATDL4165C	PHYTOCHROME D	10.01	S46312 ARATH. PHYTOCHROME D	4360	5263
ATDL4205C	CALCIUM SENSOR -LIKE	10.01	Z70783, H. C. ELEGANS CA SENSOR PROTEIN	141	724
ATDL3360W	CALMODULIN -LIKE	10.04	L01433 SOYBEAN CALMODULIN	609	693
ATDL3210C	CASEIN KINASE 1	10.04	F88141 ARATH. CASEIN KINASE 1	2309	2309
ATDL3215C	PROTEIN KINASE -LIKE	10.04	S42864 ICE PLANT PROTEIN KINASE	971	3169
ATDL3280C	SPS-1 KINASE -LIKE	10.04	HJ1080 YEAST SPS-1 SPORE SPECIFIC KINASE	443	2327
ATDL3390C	SNF KINASE -LIKE	10.04	A53467 WHEAT SNF1 HOMOLOG	770	2075
ATDL3430W	PROTEIN KINASE -LIKE	10.04	S29851 SOYBEAN PROTEIN KINASE 6	472	1833
ATDL3865W	PROTEIN KINASE -LIKE	10.04	S49313 SLIME MOULD PROTEIN KINASE	475	1050
ATDL4210W	AMP-ACTIVATED PROTEIN KINASE -LIKE	10.04	U42411 RAT AMP-ACTIVATED PROTEIN KINASE BETA SUBUNIT	381	1961
ATDL4255W	PROTEIN KINASE -LIKE	10.04	H54024 HUMAN GDC-2 RELATED PROTEIN KINASE	104	607
ATDL4515C	PROTEIN KINASE -LIKE	10.04	S56143 S. POMBE PROTEIN KINASE HSK1	164	3806
ATDL4855W	CASEIN KINASE 2 BETA CHAIN	10.04	S47968 ARATH. CASEIN KINASE 2 BETA CHAIN	1444	1444
ATDL4865W	PROTEIN KINASE -LIKE	10.04	S38326 MOUSE. PROTEIN KINASE	866	1820
ATDL3750W	PROTEIN PHOSPHATASE -LIKE	10.041	U38193 ORYCTOLAGUS PROTEIN PHOSPHATASE	1413	4517
ATDL3995C	COP9	10.99	A54842 ARATH. COP9	1016	1016
ATDL3990W	PR1-1 G-BETA PROTEIN	10.99	S49820 ARATH. PR11 PROTEIN	1806	1806
ATDL3000W	DISEASE RESISTANCE GENE -LIKE	11.01	U42445 SOLANUM PIMP. CF-2.2	1131	4278
ATDL3225C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	1392	9717
ATDL3345C	DISEASE RESISTANCE PROTEIN -LIKE	11.01	A54809 ARATH. RPP2 DISEASE RESISTANCE PROTEIN	552	2668
ATDL4460C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	4326	6448
ATDL4470C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	387	1246
ATDL4475C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	3495	6010
ATDL4480C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	2567	2567
ATDL4490C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	5065	6675
ATDL4495C	RPP5 -LIKE (FRAGMENT)	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	408	753
ATDL4500C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	2723	5568
ATDL4505C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	3116	6705
ATDL4510C	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	2354	5261
ATDL4525W	DOWNY MILDEW RESISTANCE PROTEIN RPP5 -LIKE	11.01	G2109275 ARATH. DOWNY MILDEW RESISTANCE PROTEIN RPP5	1286	4556
ATDL3070W	MAJOR LATEX PROTEIN TYPE 1 -LIKE	11.02	R91914 ARATH. MAJOR LATEX HOMOLOG TYPE 2	641	763
ATDL3155C	70S RELATED PROTEIN -LIKE	11.05	U41538, H. C. ELEGANS HSP-100 PROTEIN	1178	1878
ATDL3375W	HEAT SHOCK PROTEIN -LIKE	11.05	L35272 SOYBEAN HEAT SHOCK PROTEIN	2043	3482
ATDL3385W	EARLY LIGHT INDUCED PROTEIN (ELIP) -LIKE	11.05	S71560 SUNELOWER ELIP SDI-1 DROUGHT INDUCED PROTEIN	475	1548
ATDL3995W	DROUGHT-INDUCED PROTEIN D121 -LIKE	11.05	A51479 ARATH. DROUGHT-INDUCED PROTEIN D121	318	404
ATDL4135W	CYSTEINE PROTEINASE -LIKE	11.05	D00718 ARATH. DROUGHT-IND. CYSTEINE PROTEINASE	1377	1888
ATDL4295C	HEAT SHOCK PROTEIN 21 -LIKE	11.05	S16004 PEPUNIA HSP 21	110	1128
ATDL4300C	HEAT SHOCK PROTEIN 21 -LIKE	11.05	S65049 SOYBEAN HSP 23.9 PRECURSOR	125	3721
ATDL4615C	GTP-BINDING RAB2A -LIKE	11.05	Z73937 LOTUS JAPONICUS RAB2A	1004	2248
ATDL4620C	GTP-BINDING RAB2A -LIKE	11.05	Z73937 LOTUS JAPONICUS RAB2A	1010	2248
ATDL4800C	GTP-BINDING RAB1C -LIKE	11.05	Z73882 LOTUS JAPONICUS RAB1C	127	127
ATDL4835W	HEAT SHOCK TRANSCRIPTION FACTOR 21 -LIKE	11.05	S59537 SOYBEAN HEAT SHOCK TRANSCRIPTION FACTOR 21	1727	1729
ATDL3055C	SELENIUM BINDING PROTEIN -LIKE	11.06	S27828 MOUSE. HEPATIC SELENIUM-BINDING PROTEIN	1609	2542
ATDL3060C	SELENIUM BINDING PROTEIN -LIKE	11.06	L_41488, D. WALLEYE DERMAL SARCOMA VIRUS GAG PROTEIN	1211	5455
ATDL3015C	PREDICTED	12	A9282 SUGAR BEET YELLOW VIRUS FUSION PROTEIN 1A/1B	138	2254
ATDL3020C	PREDICTED	12	IS2418 HUMAN CYTOCHROME P450	98	2593
ATDL3035W	PREDICTED	12	TRANSLATION INITIATION FACTOR IF-2 PRECURSOR	81	542
ATDL3075C	AMP-BINDING PROTEIN -LIKE	12	Z72151 BRASSICA NAPUS AMP BINDING PROTEIN	282	2982
ATDL3085W	PREDICTED	12	G1374712 PROTON-COUPLED PEPTIDE TRANSPORTER PEPT	105	5404
ATDL3100W	PREDICTED	12	U66560, C. MYCOBACTERIUM AVIUM EMBB	98	1134
ATDL3125C	MEMBRANE PROTEIN -LIKE	12	U08285 TOBACCO SALT-INDUCE MEMBRANE PROTEIN	123	2236
ATDL3135C	MEMBRANE PROTEIN -LIKE	12	U08285 TOBACCO SALT-INDUCE MEMBRANE PROTEIN	212	2988
ATDL3140C	PREDICTED	12	S52076 SLIME MOULD PROTEIN KINASE	102	3142
ATDL3150W	PREDICTED	12	A43444 HUMAN TAU PROTEIN	105	1518
ATDL3170C	PREDICTED	12	S17286 DROSOPHILA PER PROTEIN	83	2308
ATDL3175W	PREDICTED	12	T1818 BARLEY TRYPSIN INHIBITOR	96	2316
ATDL3185C	PREDICTED	12	S29274 PHBC, CHRVI POLYHYDROXYBUTYRATE POLYMERASE	96	2524
ATDL3195C	PREDICTED	12	MMMSB1 MOUSE LAMININ CHAIN B1 PRECURSOR	119	5907
ATDL3230W	PREDICTED	12	A36985 S. POMBE P TYPE MATING FACTOR	101	978
ATDL3235W	PREDICTED	12	U00111 CHICKEN NFKB	125	3453
ATDL3240W	PREDICTED	12	A39767 HUMAN RFX	144	378
ATDL3265W	PREDICTED	12	A41541 RAT ADENYLATE CYCLASE	83	744
ATDL3285C	PREDICTED	12	D50867 HALOCYNTHTA TROPONIN T	101	1759
ATDL3290W	PREDICTED	12	S57929 BOVINE PHOSPHATIDYL CHOLINE TRANSFER PROTEIN	102	2138
ATDL3305C	PREDICTED	12	U09822 ARGENTEM MYOSIN HEAVY CHAIN	118	607
ATDL3305C	PREDICTED	12	S05518 CHICKEN LAMIN B-1	98	612
ATDL3315C	IAA7 PROTEIN -LIKE	12	S58494 ARATH. IAA7 PROTEIN	875	939
ATDL3320W	AUXIN-RESPONSIVE PROTEIN IAA1	12	AX11 ARATH. AUXIN-INDUCED PROTEIN	826	827
ATDL3335W	PREDICTED	12	S44516 TEL1 PROTEIN YEAST	119	618
ATDL3340W	PREDICTED	12	S74555 SYNECHOCYSTIS ABC TRANSPORTER PROTEIN	101	2874
ATDL3350C	PREDICTED	12	S34961 RAT SYNAPTIC VESICLE PROTEIN 2 FORM B	92	1697
ATDL3365C	PREDICTED	12	S00485 PLASMODIUM GENE 11-1 PROTEIN PRECURSOR	181	2846
ATDL3405C	PREDICTED	12	S67539 YEAST MNE1 GENE	117	1517
ATDL3410C	PREDICTED	12	S19586 RAT N-ME-D-ASPARTATE RECEPTOR PROTEIN	130	523
ATDL3425C	PREDICTED	12	A56678 DROSOPHILA YEMANUCLEIN-ALPHA	96	2257
ATDL3435W	PREDICTED	12	S39162 HUMAN CREB BINDING PROTEIN	92	3031
ATDL3450C	PREDICTED	12	L29389 YEAST FUN12P	121	958
ATDL3455W	PREDICTED	12	M92439 HUMAN LEUCINE-RICH PROTEIN	173	3677
ATDL3460C	PREDICTED	12	U47087 CARROT PATHOGENESIS-RELATED PROTEIN	87	741
ATDL3465W	PREDICTED	12	D45163 HALOCYNTHTA MYOSIN HEAVY CHAIN	127	2524
ATDL3470W	PREDICTED	12	S29496 HALOCYNTHTA MYOSIN HEAVY CHAIN	90	2928
ATDL3475W	PREDICTED	12	L39769, E. STREPTOCOCCUS PLASMID pSP90 GENES	88	881
ATDL3505W	PREDICTED	12	SECE_THEME PREPROTEIN TRANSLOCASE	124	786
ATDL3515W	PREDICTED	12	G64369 M.JANNSCHII SURE SURVIVAL PROTEIN	113	1362
ATDL3525W	PREDICTED	12	KC5260 HUMAN PROGESTERONE MEMBRANE BINDING PROTEIN	143	3999
ATDL3530C	PREDICTED	12	U35238 ORYCTOLAGUS NA GUNNEL	106	3972
ATDL3535C	PREDICTED	12	TVHUBF HUMAN PROTEIN KINASE B-RAF	92	737
ATDL3535C	PREDICTED	12	RXRB MOUSE RETINOIC ACID RECEPTOR RXR-BETA	103	5008
ATDL3570C	PREDICTED	12	C64212 MYCOPLASMA GENTILIIUM PGIB	95	2093
ATDL3580C	PREDICTED	12	S14422 C.ELEGANS ENDOPROTEINASE	135	4070
ATDL3605W	PREDICTED	12	F47021 ERWINIA SP. OUTH PECTIC ENZYME SECRETION	94	831
ATDL3615C	PREDICTED	12	WMVN 49 40.9K AUTOGRAPHIA NUCLEAR POLYHEDROSIS VIRUS	96	906
ATDL3620C	PREDICTED	12	X05285 DROSOPHILA FIBRILLARIN	119	481
ATDL3640C	PREDICTED	12	S59411 YEAST PEPT	81	604
ATDL3740C	PREDICTED	12	X95343 TOBACCO HSR201 PROTEIN	247	2126
ATDL3745C	PREDICTED	12	X95343 TOBACCO HSR201 PROTEIN	237	2090
ATDL3795W	PREDICTED	12	A41369 CABBAGE S RECEPTOR KINASE	86	1363
ATDL3810W	PREDICTED	12	A40393 MOUSE CYSTIC FIBROSIS CONDUCTANCE RECEPTOR	130	2718
ATDL3845W	PREDICTED	12	I51270 POEPHILA MYELIN PROTEOLIPID PROTEIN	86	718
ATDL3850W	PREDICTED	12	U52866, C. RHIZOBIUM CCM8 CYTOCHROME ASSEMBLY	93	913
ATDL3855W	PREDICTED	12	S78467 SYNECHOCYSTIS SPORE MATURATION PROTEIN B	97	885
ATDL3860C	PREDICTED	12	L76937 HUMAN WERNER SYNDROME	149	909
ATDL3895W	PREDICTED	12	A44357 SLIME MOULD DYNEIN HEAVY CHAIN	90	739
ATDL3900W	MEMBRANE PROTEIN -LIKE	12	U08285 TOBACCO MEMBRANE ASSOC.SALT-INDUCED PROTEIN	146	2972
ATDL3910C	PREDICTED	12	B37237 XENOPUS PROTEIN KINASE C	109	2237
ATDL3935W	PREDICTED	12	G155761 NAELICERIA MYOSIN II HEAVY CHAIN	112	517
ATDL3940C	IAF86 GTP-BINDING PROTEIN -LIKE	12	PSIAF86A, 1 PEA IAF86 GTP-BINDING PROTEIN	405	2260
ATDL3950W	PREDICTED	12	A38713 SEA URCHIN KINESIN HEAVY CHAIN	108	2807
ATDL3955C	PREDICTED	12	U18792 BABESIA GLUTAMATE DEP. CARBAMOYL P SYNTH.	90	1178
ATDL3960W	PREDICTED	12	U20449 PARAMECIUM DYNEIN HEAVY CHAIN	89	3278
ATDL4025W	ZINC-FINGER PROTEIN -LIKE	12	Z72511, C. G. ELEGANS ZFN FINGER PROTEIN	195	2310
ATDL4065W	PREDICTED	12	A40253 YEAST ACIDIC NUCLEAR PROTEIN SPT5	145	4425
ATDL4070W	PREDICTED	12	RPOB, CYANOPHORA DNA-DIRECTED RNA POL. BETA CHAIN	103	1367
ATDL4085W	PREDICTED	12	S17857, E. COLI OUTER MEMBRANE PROTEIN	95	1312
ATDL4090W	PREDICTED	12	U31777 RAT ATROPHIN-1 PROTEIN	89	905
ATDL4095W	PREDICTED	12	B33862 E.COLI REGULATORY PROTEIN HYD	131	2974
ATDL4100C	PREDICTED	12	ODZJ1 BRADYRHIZOBIUM CYTOCHROME C OXIDASE	109	3298
ATDL4125C	PREDICTED	12	G64419 M. JANNSCHII SERINE AMINOTRANSFERASE	92	2310
ATDL4130C	PREDICTED	12	S23662 E.COLI CYTOSINE DEAMINASE	123	833
ATDL4160C	GLYCINE RICH PROTEIN -LIKE	12	QJ1060 ARATH. GLYCINE-RICH PROTEIN	155	207
ATDL4190W	PREDICTED	12	A42111 ENTEROCOCCUS NAPA Na/H EXCHANGING PROTEIN	84	1375
ATDL4220C	PREDICTED	12	S42629 RABBIT KERATIN	113	386
ATDL4225W	MEMBRANE PROTEIN -LIKE	12	U08285 TOBACCO MEMBRANE ASSOC.SALT-INDUCIBLE PROT.	225	2832
ATDL4233C	PREDICTED	12	X92429 STREPTOMYCES N-METHYL TRANSFERASE	98	916
ATDL4245C	PREDICTED	12	S31336 KLUVEROAMYCES LET1 PROTEIN	185	2827



GENE	IDENTITY	FUN.CAT.	HOMOLOG	FASTA	SELF FASTA
ATDL4260C	MEMBRANE PROTEIN -LIKE	12	U08285 TOCACC0 MEMBRANE-ASSOC. SALT-INDUCIBLE PROT.	120	2378
ATDL4290W	PREDICTED	12	U09274 CARCINUS NA/H EXCHANGER	122	3848
ATDL4305C	PREDICTED	12	L47741 PICEA MITOCHONDRIAL HSP 23.5	145	5215
ATDL4310W	PREDICTED	12	A41098 RAT CA CERNIELL ISOPORNIN A	100	2389
ATDL4330C	PREDICTED	12	S02035 DROSOPHILA PER PROTEIN	107	945
ATDL4335C	PREDICTED	12	U43629 BEET MEMBRANE PROTEIN	174	2907
ATDL4355W	PREDICTED	12	D85904 MOUSE APG-2	306	1323
ATDL4360W	PREDICTED	12	A40718 HUMAN HOST CELL FACTOR C1 PRECURSOR	115	1848
ATDL4420C	PREDICTED	12	A29356 KIDNEY BEAN HYDROXYPROLINE RICH PROTEIN	158	2672
ATDL4430C	PREDICTED	12	S28261 HUMAN CENTROMERE PROTEIN E	97	2866
ATDL4440W	PREDICTED	12	U26024 BOVINE NUCLEAR ANTIGEN	152	1826
ATDL4445W	MEMBRANE PROTEIN -LIKE	12	U02825 TOBRACCO MEMBRANE ASSOC. SALT-IND. PROT	148	4334
ATDL4450W	PREDICTED	12	A28755 N. CRASSA UBQ CYTO. C. REDUCTASE	115	1499
ATDL4520C	PREDICTED	12	A26099 SOYBEAN GLYCINE-RICH CELL WALL PROTEIN	132	480
ATDL4530C	PREDICTED	12	A45990 RABBIT TRIADIN PROTEIN	132	3046
ATDL4555W	PREDICTED	12	B42477 E. COLI PHOSPHOTRANSFERASE SYSTEM ENZYME II	90	1577
ATDL4560C	PREDICTED	12	A34840 XENOPUS HRNP 1a	133	4440
ATDL4565C	PREDICTED	12	S40507 RUMEN FUNGUS ENDOGLUCANASE	94	1620
ATDL4590C	PREDICTED	12	A48805 MOUSE INSULIN-LIKE GROWTH FACTOR 1	115	1722
ATDL4600C	PREDICTED	12	S79000 SINECHOCYSTIS MG CHELATASE SUBUNIT	104	3630
ATDL4605C	PREDICTED	12	U29335 C. ELEGANS SEX MUSCLE ABNORMAL PROTEIN 5	132	6666
ATDL4610W	PREDICTED	12	A35548 RHIZOBIUM MELILOTI ndb8 PROTEIN	107	1991
ATDL4635W	PREDICTED	12	A37353 XENOPUS MEMBRANE PROTEIN	91	1954
ATDL4655C	PREDICTED	12	JH0280 HUMAN H60K GOLGI ANTIGEN	119	1479
ATDL4660W	PREDICTED	12	U12925 A. CERATIS MITOCHONDRIAL NADH DEH'ASE	89	3073
ATDL4675C	PREDICTED	12	U18197 HUMAN ATP-CITRATE LYASE	100	1262
ATDL4695C	PREDICTED	12	S64942 YEAST PROBABLE MEMBRANE PROTEIN	111	8955
ATDL4710W	PREDICTED	12	YG85TB BACILLUS BREVIS TYROCIDINE SYNTHASE	89	1806
ATDL4740W	HUMAN DNA-BINDING PROTEIN 5 -LIKE	12	S26509 HUMAN DNA-BINDING PROTEIN 5	188	2114
ATDL4745C	PREDICTED	12	S51330 HUMAN FLAVIN-CONTAINING MONOOXYGENASE	95	1389
ATDL4755W	PREDICTED	12	A38712 HUMAN FIBRILLARIN	111	1044
ATDL4805W	PREDICTED	12	I54222 MOUSE HOUSEKEEPING PROTEIN	84	950
ATDL4815W	PREDICTED	12	S51232 PEA OVARY PROTEIN	89	3073
ATDL4820C	PREDICTED	12	A64251 MYCOPLASMA GLUTAMATE-TRNA LIGASE	106	3767
ATDL4825C	TEGT -LIKE	12	S42069 RAT TEGT PROTEIN	359	1345
ATDL4830C	PREDICTED	12	A48998 MOUSE NUCLEAR PROTEIN	115	2690
ATDL4840C	TRP-16 -LIKE	12	S62356 HUMAN TRP-16 PROTEIN	291	7303
ATDL4845W	PREDICTED	12	S60465 C.ELEGANS DOM3 PROTEIN	149	5793
ATDL4850C	PREDICTED	12	A33638 MOUSE ANION EXCHANGER AE3	117	921
ATDL4860W	MAJOR SPERM PROTEIN -LIKE	12	U23515 F. C. ELEGANS MAJOR SPERM PROTEIN	188	1357
ATDL4870C	PREDICTED	12	C53234 MAIZE GLOBULIN 40	99	3073
ATDL4875C	PREDICTED	12	S68451 HUMAN APOPTOSIS-INHIBITOR XIAP	122	1403
ATDL4895C	PREDICTED	12	A49465 BOVINE COATOMER ZETA CHAIN	95	1379
ATDL4915W	PREDICTED	12	A24302 RABBIT GLYCOGEN PHOSPHORYLASE	92	1252
ATDL4935C	PREDICTED	12	A17015 PSEUDOMONAS CEPHALOSPORIN CYCLASE	112	1360
ATDL3025C	HYPOTHETICAL	13	Z29639 C.ELEGANS HYPOTHETICAL PROTEIN F54E4.1	114	3454
ATDL3040W	HYPOTHETICAL	13	Z66514_B.C.ELEGANS HYPOTHETICAL PROTEIN	240	1330
ATDL3045C	HYPOTHETICAL	13			587
ATDL3065C	HYPOTHETICAL	13	U15181_AT MYCOBACTERIUM HYPOTHETICAL PROTEIN	109	1262
ATDL3130W	HYPOTHETICAL	13	D84066 SINECHOCYSTIS HYPOTHETICAL PROTEIN	113	3073
ATDL3160C	HYPOTHETICAL	13	U14548_C.C.ELEGANS HYPOTHETICAL PROTEIN	323	2402
ATDL3180W	HYPOTHETICAL	13	S25990 LIVERWORT HYPOTHETICAL PROTEIN	105	4025
ATDL3250C	HYPOTHETICAL	13	S55204 HYPOTHETICAL YEAST PROTEIN	88	762
ATDL3395C	HYPOTHETICAL	13	S36039 YEAST HYPOTHETICAL PROTEIN YMR09w	41	4348
ATDL3500C	HYPOTHETICAL	13	E24548_C.ELEGANS HYPOTHETICAL	153	5124
ATDL3540C	HYPOTHETICAL	13	H6451 METHANOCoccus HYPOTHETICAL PROTEIN MJEC508	103	1911
ATDL3550W	HYPOTHETICAL	13			791
ATDL3555W	HYPOTHETICAL	13	Z54342_0.C.ELEGANS F3C11.1	138	2332
ATDL3575W	HYPOTHETICAL	13	S44609_C.ELEGANS C02F3.7	140	2855
ATDL3585C	HYPOTHETICAL	13	S41011 HYPOTHETICAL PROTEIN ZK757.1_C. ELEGANS	168	3555
ATDL3590W	HYPOTHETICAL	13	S49183 STREPTOMYCES GRISEUS HYP. PROTEIN	453	8613
ATDL3630W	HYPOTHETICAL	13			148
ATDL3635W	HYPOTHETICAL	13	Z77655_G_C56A3.1_C.ELEGANS HYPOTHETICAL PROTEIN	148	11666
ATDL3645C	HYPOTHETICAL	13	A34043 POLYCHAETE HYPOTHETICAL PROLINE RICH PROTEIN	193	3063
ATDL3665C	HYPOTHETICAL	13	YR47_CAEEL HYPOTHETICAL PROTEIN	94	2693
ATDL3760W	HYPOTHETICAL	13	S51583 ARATH. HYPOTHETICAL PROTEIN HYP1	801	3843
ATDL3775W	HYPOTHETICAL	13	S44609_C.ELEGANS C02F3.7 HYPOTHETICAL PROTEIN	445	5048
ATDL3800W	OBP 33 PEP PROTEIN -LIKE	13	S71213 ARATH. OBP33 PEP PROTEIN	893	1389
ATDL3905C	HYPOTHETICAL	13	G1469195 HUMAN KIAA0136 PROTEIN	151	5028
ATDL3915C	HYPOTHETICAL	13	S10911 CARROT HYPOTHETICAL PROTEIN	105	834
ATDL3920C	HYPOTHETICAL	13	D84566 M.JANANS HYPOTHETICAL PROTEIN	106	3195
ATDL3925W	HYPOTHETICAL	13	S65230 YEAST HYPOTHETICAL PROTEIN YPL211w	267	368
ATDL3945C	HYPOTHETICAL	13	S55101 YEAST HYPOTHETICAL PROTEIN YMR219w	117	2268
ATDL3980W	HYPOTHETICAL	13	Y476_HUMAN HYPOTHETICAL MYELOID CELL LINE 6 PROTEIN	394	8268
ATDL3985W	HYPOTHETICAL	13	G479441 HUMAN ORF	727	460
ATDL4040C	HYPOTHETICAL	13			460
ATDL4060C	HYPOTHETICAL	13	U23176_C.ELEGANS COSMID F21H12	83	438
ATDL4075C	HYPOTHETICAL	13	U00051_G.C.ELEGANS COSMID F4209	188	3232
ATDL4080C	HYPOTHETICAL	13	Z70270_G.C.ELEGANS CS36.7	107	1618
ATDL4185W	HYPOTHETICAL	13	Z3562_C.C.ELEGANS HYPOTHETICAL PROTEIN	484	5906
ATDL4230W	HYPOTHETICAL	13	D79987 HUMAN KIAA0165 HYPOTHETICAL PROTEIN	117	1060
ATDL4270C	HYPOTHETICAL	13	S50446 YEAST HYPOTHETICAL PROTEIN YEL013w	165	2093
ATDL4280W	HYPOTHETICAL	13			521
ATDL4315C	HYPOTHETICAL	13	S46810 YEAST HYPOTHETICAL PROTEIN YHR076W	124	2104
ATDL4380W	HYPOTHETICAL	13	S64051 YEAST HYPOTHETICAL PROTEIN YGL047W	170	851
ATDL4535W	HYPOTHETICAL	13			386
ATDL4550C	HYPOTHETICAL NON-LTR	13	S75438 SINECHOCYSTIS HYPOTHETICAL PROTEIN	138	1124
ATDL4570W	HYPOTHETICAL	13	S74454 SINECHOCYSTIS HYPOTHETICAL PROTEIN SLR1485	331	2382
ATDL4580W	HYPOTHETICAL	13	Z69883_D.C.ELEGANS HYPOTHETICAL PROTEIN	94	726
ATDL4595C	HYPOTHETICAL	13	Z80220_C.C.ELEGANS T08G1	189	8294
ATDL4670W	HYPOTHETICAL	13	I57997 MOUSE HYPOTHETICAL CA BINDING PROTEIN	429	1108
ATDL4680W	HYPOTHETICAL	13	I54209 HUMAN HYPOTHETICAL PROTEIN	85	541
ATDL4690C	HYPOTHETICAL	13			271
ATDL4700C	HYPOTHETICAL	13	I51116 SEA LAMPREY HYPOTHETICAL PROTEIN NF-180	136	7495
ATDL4735W	HYPOTHETICAL	13	K_5441 YEAST HYPOTHETICAL PROTEIN	205	1340
ATDL4750C	HYPOTHETICAL	13	G64015 HAEMOPHILUS HYPOTHETICAL PROTEIN	106	2391
ATDL4795W	HYPOTHETICAL	13	U53154_K.C.ELEGANS HYPOTHETICAL PROTEIN	155	1621
ATDL4885W	HYPOTHETICAL	13	Z50875_B.C.ELEGANS HYPOTHETICAL PROTEIN	85	1803
ATDL4930W	HYPOTHETICAL	13	D26067 HUMAN ORF41	530	1353
ATDL3270W	RETROELEMENT POLYPROTEIN TA1-3 LIKE	14	S05465 ARATH. RETROELEMENT ITR	765	8744
ATDL3275W	RETROELEMENT TA11-1 LIKE	14	E248476 NON-LTR RETROTRANSPOSON	348	1702
ATDL3835C	RETROELEMENT NON-LTR	14	S65812 ARATH. RETROTRANSPOSON TA11-1 REV.TXASE	815	4838
ATDL3840C	RETROTRANSPOSON FINGER PROTEIN	14	S65811 ARATH. RETROTRANSPOSON TA11-1 FINGER PROT.	177	3204
ATDL3970C	RETROELEMENT NON-LTR	14	S23115 RETROVIRUS RELATED PROTEIN TA1-2	238	1345
ATDL4045W	RETROTRANSPOSON ITR	14	U12626 MAIZE HOPS/COTCH RETROTRANSPOSON	1112	9689
ATDL4050W	RETROTRANSPOSON ITR	14	U12626 MAIZE HOPS/COTCH RETROTRANSPOSON	585	1659
ATDL4465C	COPIA-LIKE LTR RETROTRANSPOSON	14	G531389 MAIZE COPIA-LIKE TRANSPOSON HOPS/COTCH	1459	7098
ATDL4485C	LTR RETROTRANSPOSON	14	I00791 RETROVIRUS YEAST RETROTRANSPOSON	238	3507
ATDL4760C	LTR RETROTRANSPOSON	14	U12626 MAIZE HOPS/COTCH RETROELEMENT PROTEIN	1252	7148
ATDL3600C	CYTOCHROME P450 -LIKE	20.1	S71663 PEA CYTOCHROME P450	2118	2948
ATDL3695C	CYTOCHROME P450 -LIKE	20.1	S55739 ARATH. CYTOCHROME P450	490	2569
ATDL3700W	CYTOCHROME P450 -LIKE	20.1	S55379 ARATH. CYTOCHROME P450	317	1548
ATDL3710C	CYTOCHROME P450 -LIKE	20.1	S62899 SOYBEAN CYTOCHROME P450	977	2617
ATDL3720W	CYTOCHROME P450 -LIKE	20.1	S62899 SOYBEAN CPY93 A1 CYTOCHROME P450	829	2574
ATDL3725W	CYTOCHROME P450 -LIKE	20.1	S62899 SOYBEAN CPY93 A1 CYTOCHROME P450	1016	2701
ATDL3735W	CYTOCHROME P450 -LIKE	20.1	S62899 SOYBEAN CPY93 A1 CYTOCHROME P450	784	2489
ATDL4175W	PEROXIDASE -LIKE	20.1	I37790 STYLOMATHES CATIONIC PEROXIDASE	548	1698
ATDL4195C	DIOXYGENASE -LIKE	20.1	L42466 PICEA ETHYLENE FORMING ENZYME	475	1217
ATDL4880W	PEROXIDASE -LIKE	20.1	L36158 MEDICAGO PEROXIDASE	789	1610
ATDL3715C	LUPULIN SYNTHASE -LIKE	20.2	ATU49919_1 ARATH LUPULIN SYNTHASE	1309	4144
ATDL3730C	LUPULIN SYNTHASE -LIKE	20.2	ATU49919_1 ARATH LUPULIN SYNTHASE	787	3274
ATDL3975C	SESQUITERPENE CYCLASE -LIKE	20.2	U20187 HYOSCAMUS VETISPIRADINE SYNTHASE	582	918
ATDL4390W	LIMONENE CYCLASE -LIKE	20.2	A48863 SPEARMINT LIMONENE CYCLASE	1189	3020
ATDL4395W	LIMONENE CYCLASE -LIKE	20.2	A48863 SPEARMINT LIMONENE CYCLASE	1041	5230
ATDL4150W	MYRININASE-ASSOCIATED PROTEIN -LIKE	20.6	U39289 BRASSICA MYRININASE-ASSOCIATED PROTEIN	128	960
ATDL3510W	COPPER AMINE OXIDASE -LIKE	20.99	C44239_PEA CU-AMINE OXIDASE	1684	7228
ATDL3595W	HYDROXYNITRILE LYASE -LIKE	20.99	S53311 SORGHUM HYDROXYNITRILE LYASE	762	2134
ATDL4370C	S-HYDROXYNITRILE LYASE -LIKE	20.99	U40402 HEVEA HYDROXYNITRILE LYASE	346	1325