Structure and Evolution of Genes Encoding Polyubiquitin and Ubiquitin-Like Proteins in *Arabidopsis thaliana* Ecotype Columbia

Judy Callis,* Tami Carpenter,[†,1] Chih-Wen Sun* and Richard D. Vierstra[†]

**Section of Molecular and Cellular Biology, University of California, Davis, California 95616 and [†]Department of Horticulture, University of Wisconsin-Madison, Madison, Wisconsin 53706*

Manuscript received August 17, 1993
Accepted for publication October 26, 1994

ABSTRACT

The *Arabidopsis thaliana* ecotype Columbia ubiquitin gene family consists of 14 members that can be divided into three types of ubiquitin genes; polyubiquitin genes, ubiquitin-like genes and ubiquitin extension genes. The isolation and characterization of eight ubiquitin sequences, consisting of four polyubiquitin genes and four ubiquitin-like genes, are described here, and their relationships to each other and to previously identified Arabidopsis ubiquitin genes were analyzed. The polyubiquitin genes, *UBQ3*, *UBQ10*, *UBQ11* and *UBQ14*, contain tandem repeats of the 228-bp ubiquitin coding region. Together with a previously described polyubiquitin gene, *UBQ4*, they differ in synonymous substitutions, number of ubiquitin coding regions, number and nature of nonubiquitin C-terminal amino acid(s) and chromosomal location, dividing into two subtypes; the *UBQ3*/*UBQ4* and *UBQ10*/*UBQ11*/*UBQ14* subtypes. Ubiquitin-like genes, *UBQ7*, *UBQ8*, *UBQ9* and *UBQ12*, also contain tandem repeats of the ubiquitin coding region, but at least one repeat per gene encodes a protein with amino acid substitutions. Nucleotide comparisons, K_s value determinations and neighbor-joining analyses were employed to determine intra- and intergenic relationships. In general, the rate of synonymous substitution is too high to discern related repeats. Specific exceptions provide insight into gene relationships. The observed nucleotide relationships are consistent with previously described models involving gene duplications followed by both unequal crossing-over and gene conversion events.

UBIQUITIN is a highly conserved 76-amino acid eukaryotic protein that is covalently attached posttranslationally to cellular proteins in a three-step pathway [recently reviewed by FINLEY and CHAU (1991) and by HERSHKO and CIECHANOVER (1992)]. The importance of ubiquitin in cellular function (FINLEY and CHAU 1991), which includes but is not limited to intracellular proteolysis, has prompted studies of ubiquitin protein structure, amino acid sequence, gene structure and gene expression. X-ray crystallographic structures of the human, oat and yeast ubiquitins are nearly identical (VIJAY-KUMAR *et al.* 1987). The remarkable conservation of ubiquitin amino acid sequences is evident from comparisons of coding regions from the fungal, plant and animal kingdoms. All expressed ubiquitin genes from higher plants that have been sequenced to date encode identical proteins; these differ by only one amino acid from Chlamydomonas ubiquitin (CALLIS *et al.* 1989), by two amino acids from *Saccharomyces cerevisiae* ubiquitin and by three amino acids from animal ubiquitin [reviewed in SCHLESINGER and BOND (1987) and by CALLIS and VIERSTRA (1989)]. Thus, ubiquitin rivals histone 4 as the most highly conserved protein yet described (SHARP and LI 1987).

Corresponding author: Judy Callis, Section of Molecular and Cellular Biology, University of California-Davis, Davis, CA 95616.

[1] *Present address:* Department of Nutrition, University of Wisconsin-Madison, Madison, WI 53706.

All ubiquitin genes characterized—from unicellular eukaryotes to angiosperms and mammals—have a unique structure in that they encode fusion proteins. The initial translation products are processed rapidly into ubiquitin monomers by specific proteases (JONNALAGADDA *et al.* 1989; MAYER and WILKINSON 1989; BAKER *et al.* 1992). Ubiquitin fusion polypeptides contain either multiple ubiquitin coding regions linked in tandem (polyubiquitin) or ubiquitin monomers fused to one of two unrelated proteins that localize to ribosomes (ubiquitin extension proteins; reviewed in SCHLESINGER and BOND 1987; CALLIS and VIERSTRA 1989). The last repeat of polyubiquitin genes almost always encodes a C-terminal peptide extension of one to several amino acids; these extra amino acid(s) are hypothesized to serve as a blocking group ensuring that only the monomeric protein enters the protein conjugation pathway. Why ubiquitin is encoded as a polyprotein that must be proteolytically processed to yield functional monomers is unknown. A gene expressing a single ubiquitin coding region can complement a deletion of the polyubiquitin gene in yeast, indicating that the polyprotein *per se* is not necessary for ubiquitin function (FINLEY *et al.* 1989).

Multiple ubiquitin genes are characteristic of most eukaryotes investigated to date. In our study of the ubiquitin gene family in *A. thaliana* ecotype Columbia, we previously have characterized the class of ubiquitin

genes encoding ubiquitin extension proteins (CALLIS *et al.* 1990). As described above, these genes encode a ubiquitin monomer fused to one of two unrelated ribosomal proteins either 52 or 81 amino acids in length. Four ubiquitin extension genes have been isolated, two of each type, and these four appear to represent all the ubiquitin extension genes present (CALLIS *et al.* 1990).

Two genes from Arabidopsis that contain tandem repeats of the ubiquitin coding region also have been described previously. One is an expressed polyubiquitin gene, designated *UBQ4* (BURKE *et al.* 1988). The other, *UBQ13*, has tandemly repeated ubiquitin coding regions, but contains an insertion of mitochondrial DNA in the coding region and appears to be transcriptionally silent (SUN and CALLIS 1993). We report here the characterization of eight additional ubiquitin genes: four functional polyubiquitin genes and four genes encoding ubiquitin-like proteins that appear to be pseudogenes in this organism. These genes represent the complement of ubiquitin genes from *A. thaliana* ecotype Columbia.

MATERIALS AND METHODS

Description of the Arabidopsis ubiquitin gene family: All cDNA and genomic libraries used *A. thaliana* ecotype Columbia nucleic acids as starting materials. No differences in the pattern of ubiquitin hybridizing fragments in genomic DNA from the Columbia ecotype were observed over a 5-year period. Two accessions of the Columbia ecotype available from the Arabidopsis Stock Center (CS Nos. 907 and 908) gave identical patterns of ubiquitin hybridizing fragments. In this study, genomic sequences corresponding to genes *UBQ3, 7, 8, 9, 12* and *14* and cDNA sequences corresponding to *UBQ3, 10* and *11* were isolated (see below for details). In other studies, genomic sequences for *UBQ10* and *UBQ11* were isolated (NORRIS *et al.* 1993; J. CALLIS, unpublished data). In all cases, the nucleotide sequences of ubiquitin coding regions determined from genomic clones were identical to the corresponding regions of cDNA clones (data not shown). The other previously isolated and published ubiquitin genes are the polyubiquitin gene *UBQ4*, (BURKE *et al.* 1988), the four ubiquitin extension genes, *UBQ1, UBQ2, UBQ5* and *UBQ6* (CALLIS *et al.* 1990), and the ubiquitin-like gene *UBQ13* (SUN and CALLIS 1993).

Isolation and sequence determination of Arabidopsis ubiquitin genes: *UBQ3, 7, 8* and *9* genomic sequences were isolated from a screen of 50,000 plaques of an *A. thaliana* λ EMBL3 genomic library (N. CRAWFORD) using a chicken polyubiquitin cDNA as a hybridization probe; the probe and hybridization conditions are described by CALLIS and coworkers (1990). Based on a genome size of 10^8 bp, this represents 10 genome equivalents. *UBQ12* was isolated from a screen of an equivalent number of phage from a second λEMBL3 genomic library (H. Klee) using a *Hin*dIII/*Xho*I fragment from the *UBQ3* gene as a probe (referred to as ubiquitin coding region probe) under the same hybridization conditions. *UBQ14* was isolated from a plasmid library of *Hin*dIII fragments of genomic DNA ~2 kb in size, prepared and screened as described in NORRIS *et al.* (1993).

A cDNA library in λ ZAP (Stratagene) was prepared from *A. thaliana* mature leaf poly A^+ RNA as described by HATFIELD *et al.* (1990); 200,000 plaques were screened with the ubiquitin coding region probe as above. Four different cDNAs for each of the genes, *UBQ3, UBQ10* and *UBQ11* were isolated from this library. At least one cDNA for each gene contained the entire coding region.

Phagemids were rescued from the phage according to the manufacturer's directions (Stratagene). For all genes, ubiquitin hybridizing fragments were subcloned into phagemids (VIEIRA and MESSING 1987), single-stranded DNA was prepared and the templates were sequenced by dideoxy sequencing (SANGER *et al.* 1977). Both strands of the coding region and 3′ untranslated regions were sequenced using appropriate subclones or oligonucleotides as primers. DNA and protein alignments were performed using the UW-GCG GAP and PRETTY programs (DEVEREUX *et al.* 1984). DNA sequences presented here corresponding to *UBQ3, 7, 8, 9, 10, 11, 12* and *14* have been entered in Genbank with accession numbers L05363, L05364, L05917, L03565, L05361, L05362, L05482 and L05394, respectively.

Isolation of Arabidopsis genomic DNA, hybridizations, and chromosomal mapping: Total cellular DNA was isolated from pools of several hundred 3-wk-old plants of the Columbia and Landsberg erecta ecotypes according to SANDERS *et al.* (1987) and from individual plants according to DELLAPORTA (1983). To determine the approximate number of ubiquitin genes in the genome, aliquots of DNA from the Columbia ecotype were incubated separately with the restriction enzymes *Bgl*II, *Bam*HI, *Pst*I, *Eco*RI and *Hin*dIII, fractionated by agarose gel electrophoresis, transferred to nylon membrane and hybridized with a chicken ubiquitin cDNA restriction fragment as described (CALLIS *et al.* 1990). To determine the correspondence between cloned genes and genomic DNA restriction fragments, total cellular DNA and plasmid DNA containing *Hin*dIII fragments that include the ubiquitin coding region for *UBQ3, 7, 9* and *14*, were incubated with *Hin*dIII and cofractionated by agarose gel electrophoresis. The resulting DNA gel blots were hybridized with either the chicken ubiquitin cDNA, the ubiquitin coding region (defined above) or a *Bam*HI/*Bgl*II fragment of the *UBQ4* gene (BURKE *et al.* 1988) that contained ubiquitin coding region sequences (see figure legends for specific hybridizations) under conditions described above.

To identify restriction fragment length polymorphisms (RFLPs) between two ecotypes, Columbia and Landsberg erecta DNA were digested with restriction enzymes *Hin*dIII and *Xba*I and cofractionated on agarose gels, and the resulting DNA gel blot hybridized as previously described (CALLIS *et al.* 1990) with either radiolabelled ubiquitin coding region or DNA fragments flanking the coding region of specific genes. RFLPs for seven ubiquitin genes were identified. The DNA fragments identifying polymorphisms subsequently were used as hybridization probes to genomic blots containing DNA from 155 F2 individuals (digested with the same enzyme that revealed the polymorphism) from a cross between the two ecotypes. The DNA samples on these gel blots have previously been used to map 306 RFLP markers (HAUGE *et al.* 1993). The chromosomal locations of ubiquitin genes were determined relative to these markers using three point and N-point analysis programs of MAPMAKER (LANDER and GREEN 1987) as described in detail in CHANG *et al.* (1988). Two loci were considered linked only if their LOD score was >3.0 (LANDER *et al.* 1987).

The DNA fragments from the Columbia ecotype that were used to detect RFLPs were radiolabelled by either nick translation (RIGBY *et al.* 1977) (No. 6-12-1 only) or oligo labeling (FEINBERG and VOGELSTEIN 1983) and hybridized under conditions described in CALLIS *et al.* (1990). For mapping *UBQ3*, phage No. 6-12-1 containing the genomic coding and flanking regions of *UBQ3* was used and detected a polymorphism

with *Xba*I. For *UBQ4*, the 1.3-kb *Hin*dIII fragment (plasmid p1367) that maps immediately 5′ of a two-kb *Hin*dIII fragment containing the coding region revealed a polymorphism with *Xba*I. All other genes were mapped using a *Hin*dIII restriction fragment length polymorphism. A *Hin*dIII / *Ssp*I 800-bp fragment from p1316 that maps 3′ of the *UBQ9* ubiquitin coding region was used. For *UBQ10*, an *Afl*II / *Hin*dIII restriction fragment (isolated from p3137), mapping immediately 3′ of the coding region was used. For *UBQ11*, the probe was a 3-kb *Sac*I fragment from p3169, that mapped 3′ of the coding region. For *UBQ13*, the *Hin*dIII / *Xho*I fragment (p5035) containing the *UBQ13* ubiquitin coding region was used with high-stringency hybridization conditions as described in SUN and CALLIS (1993). For *UBQ14*, a 5-kb *Hin*dIII fragment (p3190) that maps 3′ of the coding region was used.

Oligonucleotide hybridization: Oligonucleotides (24-mers) corresponding to the antisense strand of the 3′ untranslated regions of *UBQ10* (5′ CTTCTTAAGCATAACAGAGACGAG 3′) and *UBQ11* (5′ ACTTGGTTCAGTAACCATAAGAGA 3′) were end labeled with gamma-labeled ^{32}P-ATP and T4 polynucleotide kinase and purified by gel electrophoresis (SUGITA and GRUISSEM 1987). The sense sequences for *UBQ10* and *UBQ11* that hybridize to the corresponding oligonucleotide are designated in Figure 4. Oligonucleotide hybridization and wash conditions were as described (SUGITA and GRUISSEM 1987), with the final wash at $T_m - 5°$. The calculated T_m of the *UBQ10* oligonucleotide with its *UBQ10* sequence is 68° (THEIN and WALLACE 1986). Using the programs GAP and BESTFIT (DEVEREUX *et al.* 1984), the maximum identity found between the *UBQ10* oligonucleotide and all other ubiquitin DNA sequences (testing both strands) that have been determined to date (the coding regions of all other 13 genes and the flanking regions that have been sequenced) corresponded to a T_m of 50° (*UBQ14* with the *UBQ10* oligonucleotide), with all other combinations having T_ms <50°. Similarly, the *UBQ11* oligonucleotide has a calculated T_m of 66° with its corresponding *UBQ11* sequence. Using GAP and BESTFIT, the maximum T_m obtainable with all other ubiquitin genes (testing both strands) was 44° (*UBQ4* sequence with the *UBQ11* oligonucleotide). These maximum heterologous T_ms of 50 and 44° are 18 and 22° below the T_m, well below the hybridization wash conditions of $T_m - 5°$. The specificities of the radiolabeled oligonucleotides also were verified empirically by slot blot hybridizations to DNA fragments from each of the polyubiquitin genes and ubiquitin extension genes using the same hybridization stringency used for the genomic DNA gel blot described above. Under the hybridization conditions used, neither oligonucleotide showed significant hybridization to any other ubiquitin gene (data not shown and Figure 9).

Analysis of ubiquitin sequences: The proportion of synonymous nucleotide sites at which two polyubiquitin repeats were different (p_s value), and the proportion of total nucleotide sites at which ubiquitin-like repeats were different (p-distance value) were determined using the program MEGA (KUMAR *et al.* 1993). Each ubiquitin repeat from polyubiquitin, ubiquitin-like and ubiquitin-extension genes was compared pairwise with itself and every other ubiquitin repeat for estimation of the number of synonymous substitutions per site using the method of LI *et al.* (1985) with correction factor (LI 1993). Ubiquitin repeat identities were compared using the neighbor-joining method of NEI and SAITOU (1987) in the Phylip program (version 3.54c) developed by FELSENSTEIN (1989). For this, ubiquitin-like or polyubiquitin sequences initially were aligned using the program PILEUP (GCG, Madison, WI), imported into PAUP (version 3.0, SWOFFORD 1990) and exported as a Phylip compatible input files. One hundred replicas of bootstrapping were performed to assess the confidence of the observed relationships found in the minimum consensus tree (FELSENSTEIN 1988) and the bootstrap values >40% are shown. For the polyubiquitin repeat tree, the yeast *UBI1* ubiquitin sequence (OZKAYNAK *et al.* 1987) was used as an outgroup.

RESULTS

Isolation of ubiquitin genes and sequence analysis: Using as hybridization probes DNA restriction fragments from either a chicken ubiquitin cDNA (BOND and SCHLESINGER 1985) or an Arabidopsis ubiquitin coding region (see MATERIALS AND METHODS), eight new ubiquitin coding sequences were isolated from *A. thaliana* ecotype Columbia genomic and cDNA libraries. The nucleotide sequence of the coding region for each ubiquitin gene was determined. All contained tandem repeats encoding ubiquitin or proteins with similarity to ubiquitin. A diagrammatic representation of the coding region for each of the eight genes is shown in Figure 1A. *UBQ3, UBQ10, UBQ11* and *UBQ14*, designated polyubiquitin genes, encode ubiquitin polyproteins whose monomer sequence is identical to that of purified oat ubiquitin protein determined by direct protein sequencing (VIERSTRA *et al.* 1986). *UBQ7, UBQ8, UBQ9* and *UBQ12* also encode ubiquitin polyproteins, but each contained at least one ubiquitin coding region with amino acid substitutions compared with the higher plant ubiquitin sequence. These genes were designated ubiquitin-like genes.

For comparison, Figure 1B shows diagrammatic representations of previously isolated and published ubiquitin genes. These include one polyubiquitin gene, *UBQ4*, (BURKE *et al.* 1988), four ubiquitin extension protein genes, *UBQ1, UBQ2, UBQ5* and *UBQ6* that represent all of the members of this class of ubiquitin genes (CALLIS *et al.* 1990), and one ubiquitin-like gene, *UBQ13* (SUN and CALLIS 1993). The ubiquitin extension genes are distinctive in the presence of only a single ubiquitin coding region followed in frame by one of two different ribosomal proteins. These genes have been discussed in detail previously (CALLIS *et al.* 1990).

The polyubiquitin genes: The polyubiquitin genes include *UBQ3, UBQ10, UBQ11, UBQ14* (presented here) and *UBQ4* an expressed polyubiquitin gene described previously (BURKE *et al.* 1988). While encoding the same protein, these genes differ from each other in synonymous substitutions, in the number of ubiquitin coding regions (only *UBQ3* and *UBQ14* share the same number of ubiquitin repeats) and in the nature of the additional C-terminal residues (Figures 1 and 2). The polyubiquitin genes contain one of two different C-terminal additional amino acids at the end of the open reading frame. *UBQ3* and *UBQ4* open reading frames terminate in the same amino acids, serine-phenylalanine, whereas the single amino acid, phenylalanine, is found at the C-terminus of *UBQ10, UBQ11* and *UBQ14*. This difference in the C-terminus divides the five polyubiquitin genes into two subtypes: the *UBQ3/ UBQ4* subtype and the *UBQ10/ UBQ11/ UBQ14* subtype.

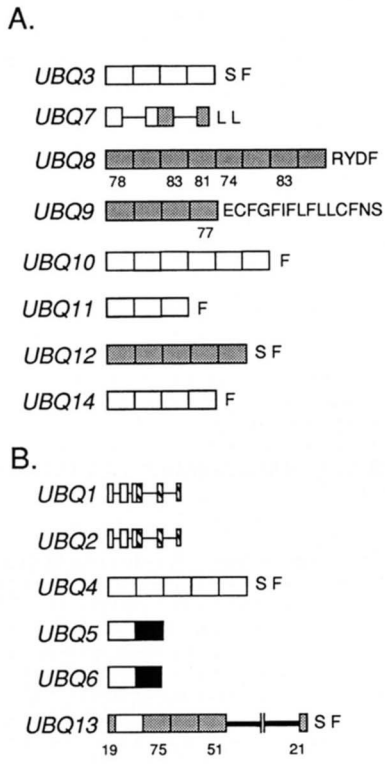

FIGURE 1.—Diagrammatic representations of the coding regions for ubiquitin genes from *Arabidopsis thaliana* ecotype Columbia. (A) Genes whose sequences are reported here. The coding regions for *UBQ3, UBQ7, UBQ8, UBQ9, UBQ12* and *UBQ14* were determined from genomic sequences. The coding regions for *UBQ3, UBQ10* and *UBQ11* were determined from cDNA sequences. (B) Genes whose sequences have been published previously and are shown for comparison purposes. Open boxes represent individual ubiquitin coding regions of 228 bp, encoding plant wild-type ubiquitin. Hatched boxes represent ubiquitin coding regions that contain at least one amino acid change from wild-type ubiquitin, either an amino acid substitution, deletion or insertion. Striped boxes represent the coding region for the 52 amino acid ribosomal protein and filled boxes the coding region for the 81 amino acid ribosomal protein (CALLIS *et al.* 1990). Thin lines represent the positions and relative lengths of intervening sequences. The thick line in *UBQ13* represents an insertion of 3.9 kb of mitochondrial DNA, interrupting a ubiquitin repeat of 72 amino acids (SUN and CALLIS 1993). The vertical lines interrupting the insertion indicate that the mitochondrial insertion is not drawn to scale. The first repeat of *UBQ13* contains only the last 19 amino acids of an ubiquitin coding region. The additional amino acids at the C terminus of polyubiquitin and ubiquitin-like open reading frames are shown. Numbers underneath ubiquitin-like repeats represent the number of amino acids encoded by that repeat if they differ from 76-amino acid residues.

The 3′ untranslated regions of the five polyubiquitin genes are shown in Figure 3 and terminate at the major poly A addition sites for *UBQ3* and *UBQ10*. From sequence analysis of four cDNAs isolated for *UBQ11,* two different poly A addition sites were found at 237 and 127 nucleotides downstream from the stop codon; the longest cDNA sequence is shown. Because cDNAs corresponding to *UBQ4* and *UBQ14* have not yet been isolated, we do not know the location of their poly A addition sites. Figure 3 shows *UBQ4* sequences 3′ of the translation stop corresponding in identity to that of *UBQ3* and shows all of the *UBQ14* sequence 3′ of the translation stop that has been obtained.

Nucleotide identity in the 3′ untranslated region also can be used to distinguish among the genes. Using this criteria, the five polyubiquitin genes divide into the same two subtypes discussed above. Nucleotide identity is highest between members of the same subtype; 75% between *UBQ3* and *UBQ4,* 83% between *UBQ10* and *UBQ11,* 87% between *UBQ10* and *UBQ14* and 75% between *UBQ11* and *UBQ14.* Nucleotide identity is only 36% between *UBQ3* and *UBQ10* and only 44% between *UBQ3* and *UBQ14* (even with several gaps introduced to increase the similarity).

The nucleotide sequence of the coding regions for all five polyubiquitin genes is shown in Figure 2. They are aligned such that the ubiquitin repeats are directly below each other. Because all polyubiquitin genes encode the same protein, all nucleotide substitutions are synonymous. The p_s value representing the proportion of synonymous nucleotide substitutions between individual repeats within genes and between genes was determined (KUMAR *et al.* 1993) and is shown in Figure 4. The extent of synonymous substitution between ubiquitin repeats is high both within and between genes (see below). This analysis also revealed that a specific repeat has smaller number of nucleotide substitutions with a specific repeat in another gene in the same subtype than when compared with any other repeat (including other repeats of the same gene) or when compared with the average p_s value of all repeat comparisons between genes. In general, this identity occurs between repeats in the same relative order in each gene. This cannot occur for all repeats, because polyubiquitin genes differ in the number of repeats. Strikingly, the terminal repeats (independent of the number of repeats) share the highest identity. These specific relationships will be described in detail.

The mean *intragenic* repeat p_s values are 0.753 and 0.736 for *UBQ3* and *UBQ4,* respectively. The mean *intergenic* differences for all the repeat comparisons between these two genes is only slightly lower at 0.654. This slightly lower intergenic level of nucleotide substitution is a reflection of specific repeats having a lower number of nucleotide substitutions (Figure 4). The first three repeats of *UBQ3* have p_s values of 0.313, 0.385 and 0.361 to the first three repeats of *UBQ4,* respectively. The fourth repeat (and last) of *UBQ3* has a p_s value of 0.556 to the fourth repeat of *UBQ4,* but a p_s value of 0.210 to the fifth (and last) repeat of *UBQ4.* The statistical significance of the difference in the extent of nucleotide substitutions between specific repeats can be assessed using the chi-square test. The null hypothesis is that the number of nucleotide substitutions are equivalent between *UBQ3* repeat one and each re-

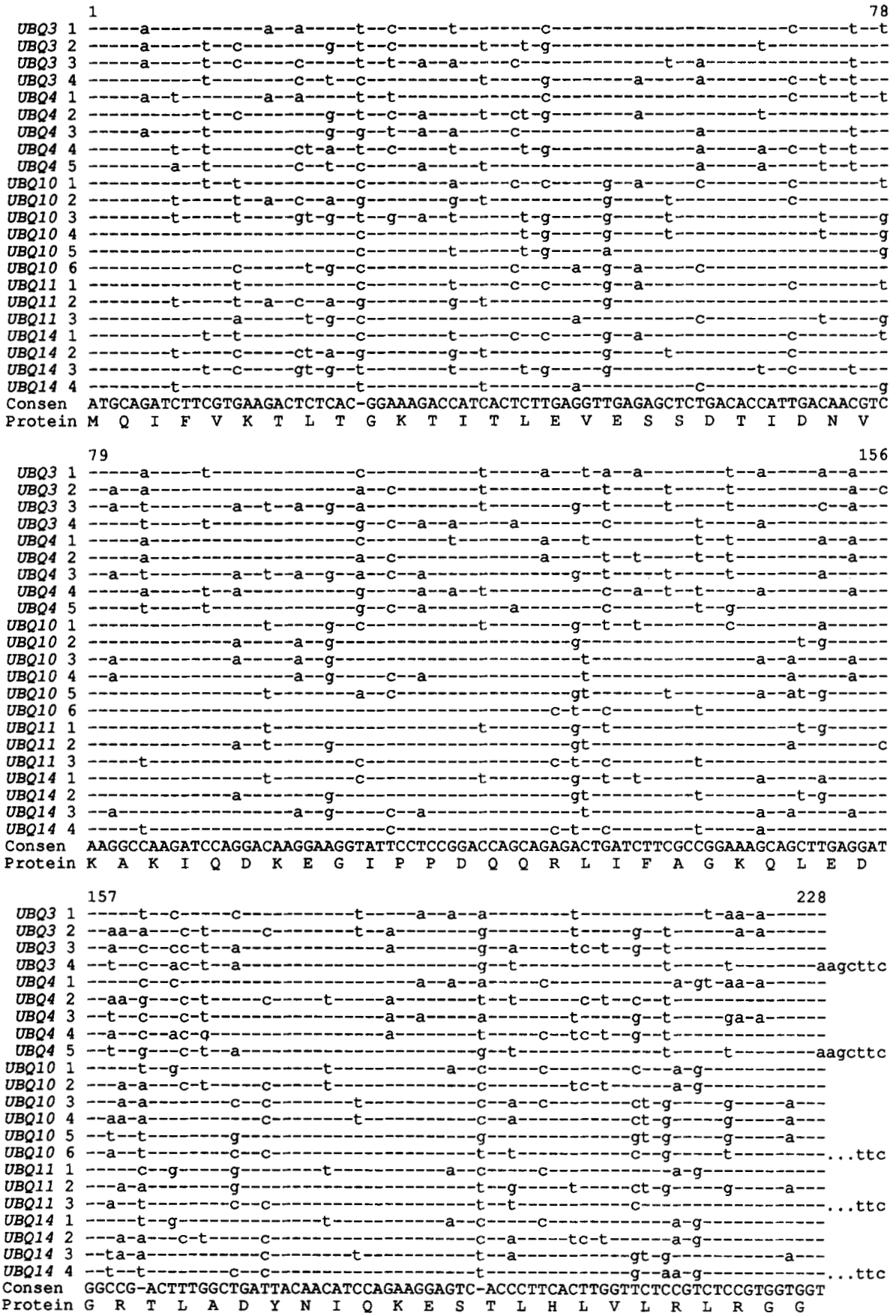

FIGURE 2.—Nucleotide sequences of the polyubiquitin genes. The nucleotide sequence of each repeat of the five polyubiquitin genes, *UBQ3, UBQ4, UBQ10, UBQ11* and *UBQ14* (denoted to the left) is aligned. The repeats within a gene are number sequentially from the 5′ end. Using the GCG PRETTY program (DEVEREUX *et al.* 1984), a consensus nucleotide sequence was generated (bottom sequence row). A dash indicates that the nucleotide is identical to the consensus; differences from the consensus are indicated. Bottom row contains the corresponding amino acid sequence.

```
                                                                                                  87
UBQ3   TaAgctttTtGTgaTcTg......ATgat.Aagt.....gGTT....GTt.CGtgTcTCATgc...ACTTgGGaGGtgatctatttc
UBQ4   TgAgc ttTtGTgaTgTg......ATcaa Aagt.....gGTT....GTt.CGagTcTCATgc...ACTTgGGaGGtgagatc ttc
UBQ10  TaAa..tcTcGTc.TcTg....ttATgcttAag.....aaGTTcaatGTttcg..TgTCATgtaaaACTTtGGtGG..........
UBQ11  TaAa..ccTtGTc.TcTctctcttATggttActgaaccaaGTTca.tGTatCG..TgTCATctagtACTTtGGtGG..........
UBQ14  TAAaaac.TtaTc.TcTg....ttATgaatca.gaagaa.GTTca.tgT..ctcgTTTCATttaaaACTTtGGtGG..........

                                                                                                 174
UBQ3   acctggtgtagTTTgTGtTTccG..tCa.GttggaaaaacttatCcc.tATcgA.....Tctcgt......TTTCatTTtCTgcTTt
UBQ4   acctggtttggTTTcTGcTTcgG..tCt.GcgggaaaaacttatCc.ttATcgA.....Tgttct..gaagTTTCatTTtCTg.TTt
UBQ10  ..........TTTgTGtTTtgGggcCttGt........ataatCcctgAT.gAataagTgttctactatgTTTCcgTTcCTg.TTa
UBQ11  ..........TTTaTGtTTtgGggcCatGt........acagcCtctgAT.aAataatTgatcgactatgTTTCcgTTtCT..TTc
UBQ14  ..........TTTgTGtTTtgGggcCttGt........aaagcCcctgAT.gAataatTgttcaactatgTTTCcgTTcCTg.Tgt

                                                                                                 260
UBQ3   tcttttatgTaCcTt.....cgtTtGggcttgtaacgggcctTTG.taTTtcaaCtcTcaATaataaTccaagtgCaTgtT.aaacpA
UBQ4   tcttttattTtC.Ttatgaacct TtGg..............TTGctgTTtcaaCatTtaATaataaTccaagtgCaTgtTaaaac
UBQ10  tctctttctTtC.TaatgacaagTcGaacttcttcttpA
UBQ11  atctctcttTtCtTtcaaacaacaaatcgaacttattctctaTTGcaaTTatctCttTcgATtcactTttgtcatCgTtgTctctt
UBQ14  tatacc..tTtCtTtctaat

                          279
UBQ11  tatatgatgtgcttagtttpA
```

FIGURE 3.—Nucleotide sequence comparisons of the 3′ untranslated regions of the Arabidopsis polyubiquitin mRNAs. For *UBQ3, 10* and *11*, these represent the sequence from cDNAs with the longest 3′ ends. No cDNAs corresponding to *UBQ4* or *UBQ14* have been isolated, so the genomic sequence of *UBQ4* corresponding in length to that of *UBQ3* is shown. All the *UBQ14* sequence available 3′ of the translation stop is shown. The 3′ sequences used to generate gene-specific antisense oligonucleotide probes are underlined. See MATERIALS AND METHODS for the oligonucleotide sequence. pA, position of the poly A tract found in the longest cDNAs. Alignments were obtained using the GCG GAP program (DEVEREUX *et al.* 1984). Uppercase nucleotides are identical among all four sequences, lower case nucleotides are not.

peat of *UBQ4*. Such a test revealed that substitutions are not randomly distributed ($0.001 < P < 0.01$).

This pattern is also observed within the *UBQ10/UBQ11/UBQ14* subtype of polyubiquitin genes. The mean *intragenic* p_s values are 0.615, 0.629 and 0.651 for *UBQ10, UBQ11* and *UBQ14*, respectively. Similarly, the mean *intergenic* p_s values for repeat comparisons between *UBQ10* and *UBQ11* is 0.540, between *UBQ10* and *UBQ14* is 0.567 and between *UBQ11* and *UBQ14* is 0.561. Again, the observed slightly lower intergenic nucleotide substitution level results from the presence of a lower level of nucleotide substitution between specific repeats *between* genes in this subtype (Figure 4). The first repeats of *UBQ10* and *UBQ11* have p_s values of 0.234 to each other, and the third repeat of *UBQ11* has the lowest p_s value, 0.193, with the last repeat of *UBQ10* (repeat six). *UBQ14* repeat one has p_s values of 0.116 and 0.156 when compared with repeat one of *UBQ10* and *UBQ11*, respectively. *UBQ14* repeat four (the last) has the lowest p_s values, 0.346 and 0.270, with the termi-

repeat	3-1	3-2	3-3	3-4	4-1	4-2	4-3	4-4	4-5	10-1	10-2	10-3	10-4	10-5	10-6	11-1	11-2	11-3	14-1	14-2	14-3	14-4
3-1		0.064	0.052	0.052	0.065	0.050	0.053	0.063	0.044	0.064	0.047	0.048	0.046	0.058	0.052	0.065	0.046	0.060	0.068	0.045	0.055	0.060
3-2	0.698		0.069	0.057	0.058	0.067	0.069	0.065	0.060	0.050	0.058	0.056	0.067	0.065	0.063	0.052	0.058	0.064	0.052	0.060	0.069	0.066
3-3	0.830	0.557		0.053	0.049	0.065	0.066	0.065	0.059	0.058	0.062	0.054	0.054	0.058	0.055	0.056	0.057	0.055	0.052	0.064	0.058	0.050
3-4	0.828	0.786	0.820		0.056	0.063	0.059	0.069	0.056	0.050	0.053	0.035	0.060	0.064	0.067	0.058	0.052	0.068	0.058	0.060	0.063	0.068
4-1	**0.313**	0.778	0.852	0.792		0.056	0.057	0.060	0.052	0.066	0.058	0.048	0.051	0.053	0.056	0.068	0.051	0.063	0.068	0.061	0.050	0.065
4-2	0.850	**0.385**	0.670	0.707	0.795		0.063	0.063	0.066	0.061	0.059	0.052	0.063	0.058	0.069	0.058	0.052	0.068	0.061	0.061	0.065	0.067
4-3	0.826	0.554	**0.361**	0.759	0.790	0.705		0.057	0.064	0.059	0.057	0.044	0.058	0.065	0.055	0.050	0.060	0.055	0.053	0.052	0.060	0.059
4-4	0.717	0.675	0.672	0.556	0.759	0.712	0.783		0.066	0.044	0.052	0.051	0.046	0.054	0.052	0.049	0.047	0.060	0.054	0.060	0.063	0.061
4-5	0.886	0.748	0.763	**0.210**	0.831	0.650	0.684	0.652		0.041	0.053	0.040	0.060	0.063	0.067	0.047	0.057	0.069	0.045	0.060	0.058	0.068
10-1	0.699	0.850	0.769	0.845	0.662	0.732	0.766	0.889	0.902		0.067	0.054	0.061	0.066	0.069	0.059	0.064	0.068	0.045	0.065	0.056	0.064
10-2	0.870	0.769	0.727	0.822	0.775	0.767	0.781	0.827	0.822	0.636		0.067	0.067	0.063	0.065	0.070	0.068	0.064	0.067	0.050	0.064	0.067
10-3	0.863	0.800	0.816	0.931	0.866	0.836	0.889	0.839	0.911	0.821	0.642		0.064	0.069	0.066	0.061	0.070	0.067	0.058	0.068	0.069	0.064
10-4	0.879	0.642	0.813	0.754	0.843	0.717	0.770	0.875	0.754	0.740	0.640	**0.295**		0.070	0.069	0.064	0.070	0.070	0.060	0.068	0.066	0.068
10-5	0.784	0.683	0.777	0.698	0.826	0.778	0.676	0.820	0.717	0.664	0.720	0.592	0.452		0.069	0.070	0.070	0.070	0.069	0.065	0.070	0.070
10-6	0.832	0.713	0.805	0.612	0.796	0.538	0.801	0.828	0.612	0.579	0.672	0.662	0.544	0.565		0.069	0.069	0.055	0.068	0.069	0.068	0.066
11-1	0.684	0.836	0.794	0.773	0.627	0.775	0.848	0.856	0.870	**0.234**	0.523	0.748	0.706	0.511	0.544		0.068	0.068	0.051	0.069	0.059	0.067
11-2	0.875	0.774	0.790	0.827	0.839	0.830	0.748	0.871	0.788	0.698	**0.386**	0.529	0.546	0.489	0.580	0.624		0.067	0.063	0.069	0.068	0.067
11-3	0.757	0.696	0.808	0.595	0.721	0.597	0.804	0.754	0.557	0.619	0.693	0.645	0.526	0.547	**0.193**	0.623	0.640		0.067	0.068	0.065	0.062
14-1	0.602	0.831	0.827	0.768	0.623	0.732	0.823	0.812	0.883	**0.116**	0.616	0.782	0.760	0.567	0.598	**0.156**	0.717	0.639		0.066	0.061	0.066
14-2	0.881	0.760	0.698	0.755	0.746	0.738	0.830	0.760	0.755	0.683	**0.155**	0.611	0.609	0.690	0.564	0.570	0.410	0.605	0.663		0.068	0.069
14-3	0.807	0.587	0.779	0.720	0.849	0.683	0.756	0.724	0.778	0.804	0.702	**0.396**	**0.336**	0.535	0.625	0.770	0.628	0.686	0.745	0.633		0.067
14-4	0.755	0.655	0.843	0.593	0.680	0.634	0.763	0.732	0.574	0.695	0.614	0.701	0.621	0.487	**0.346**	0.641	0.638	**0.270**	0.656	0.564	0.645	

FIGURE 4.— *P* distance values for synonomous substitutions (p_s) from the pairwise comparison of all the Arabidopsis polyubiquitin repeats. Each repeat is indicated by the gene number, a dash, and the number of the repeat within each gene (numbering as in Figure 2). The fraction of synonomous nucleotide sites that differ between repeats was determined using MEGA (KUMAR *et al.* 1993) and these values are shown in the lower left hand matrix. The corresponding standard errors are shown in the upper righthand matrix. *P* distance values ≤0.4 are in boldface type.

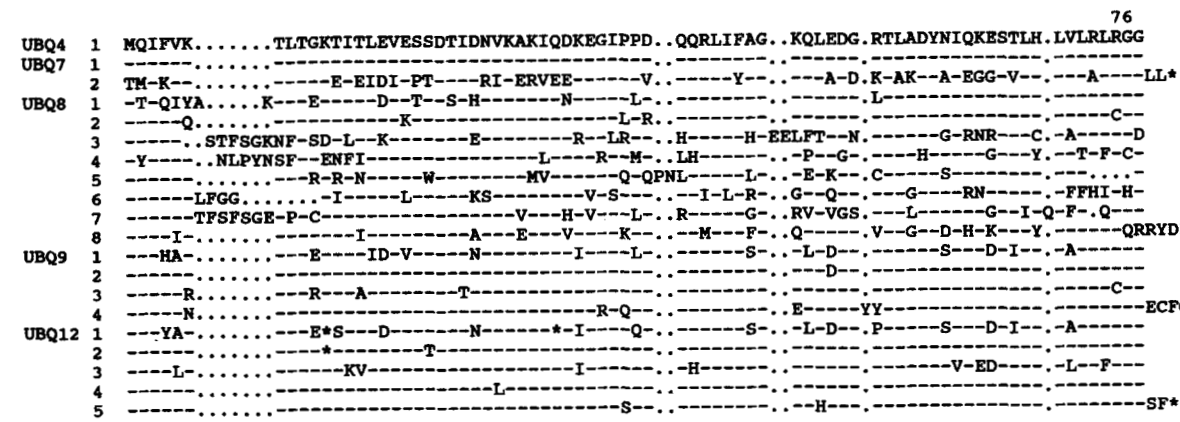

FIGURE 5.—Amino acid sequence of the Arabidopsis ubiquitin-like genes. The derived amino acid sequences of the ubiquitin-like genes are aligned with that of ubiquitin derived from the polyubiquitin gene *UBQ4* reported previously that is identical for all the polyubiquitin genes (top line). Positions of amino acid identity are indicated with a dash, and the replacements are denoted. Dots indicate the position of gaps that were introduced to increase the alignment. *, stop codons. *UBQ13*, not shown.

nal repeats of *UBQ10* and *UBQ11*, respectively. Chi-square tests support the statistical significance of this pattern (data not shown).

In contrast, this marked pattern of specific low level of nucleotide substitution is not seen when comparing repeats between genes from *different* subtypes (Figure 4). For example, the p_s value between repeat two of *UBQ3* and repeat two of *UBQ10* is 0.769, close to the overall p_s value of 0.781 between these two genes and not very different than the p_s value of 0.642 between repeat two of *UBQ3* and repeat four of *UBQ10*. The differences between repeats from different subtypes also is supported by statistical tests with, for example, the null hypothesis that nucleotide changes between *UBQ3* repeat one and the repeats of *UBQ14* are distributed randomly. This hypothesis is supported ($P > 0.3$). Additional comparisons between corresponding repeats of these two genes and other genes in different subtypes reveals the same results (Figure 4).

The ubiquitin-like genes: Ubiquitin-like genes are similar in structure to polyubiquitin genes, each encoding different numbers of tandem ubiquitin repeats with additional nonubiquitin amino acids at the C-termini (Figure 1). *UBQ12*, with five repeats, contains the same number of repeats as *UBQ4*, whereas *UBQ9*, with four repeats, contains the same number as *UBQ3* and *UBQ14*. *UBQ7* and *UBQ8* genes, with two and eight repeats, respectively, do not correspond in repeat number with any other Arabidopsis ubiquitin gene. Ubiquitin-like genes are distinguished from the polyubiquitin genes by the presence of amino acid substitutions in at least one of the ubiquitin repeats.

The derived amino acid sequences of the ubiquitin-like genes reported here are shown in Figure 5 and compared with the amino acid sequence of the ubiquitin protein present in all higher plants analyzed (CALLIS and VIERSTRA 1989). The percent amino acid identity of each ubiquitin-like repeat to ubiquitin is shown in Table 1. The number of amino acid replacements in each repeat within one gene does vary. For example, within *UBQ12* there are 18 changes in repeat one, but only one change in repeat four and two changes in repeat five. Although the first repeat of *UBQ7* encodes a 76-amino acid ubiquitin protein, the second repeat

TABLE 1

Amino acid identity of ubiquitin-like repeats to ubiquitin

Repeat	Percent identity to ubiquitin
UBQ7-1	100
UBQ7-2	57.6
UBQ8-1	86.4
UBQ8-2	94.9
UBQ8-3	66.1
UBQ8-4	71.2
UBQ8-5	76.3
UBQ8-6	66.1
UBQ8-7	64.4
UBQ8-8	77.6
UBQ9-1	76.3
UBQ9-2	98.7
UBQ9-3	93.4
UBQ9-4	93.4
UBQ12-1	76.3
UBQ12-2	97.3
UBQ12-3	86.8
UBQ12-4	98.7
UBQ12-5	97.4
UBQ13-2	100
UBQ13-3	100
UBQ13-4	97.4
UBQ13-5	97.4

Amino acid identity of ubiquitin-like repeats to ubiquitin. The amino acid sequence of each repeat of the five ubiquitin-like genes was compared with the ubituitin amino acid sequence encoded by the polyubiquitin genes and is represented as a percentage.

encodes a 76-amino acid protein only 58% identical to ubiquitin. *UBQ8* repeats range in amino acid identity to ubiquitin from a high of 95% to a low of 66%. Two ubiquitin-like genes share the same amino acid replacements. Of the 16 amino acid replacements in repeat one of *UBQ9,* 12 are found in repeat one of *UBQ12.* Two additional replacements are shared in position, but not in the nature of the amino acid. This relationship is also evident at the nucleotide level (Figure 7). No other repeats of these two genes share more than one amino acid replacement.

Specific amino acid replacements of note in *UBQ8* involve changes in the C-terminal glycine residue at the ends of the third and last repeats. Processing of ubiquitin polyproteins by ubiquitin protein hydrolases requires the presence of this Gly residue (JONNALAGADDA *et al.* 1989). Amino acid replacements in specific repeats suggest that if expressed, these polyproteins would not be cleaved into ubiquitin-like monomers at the variant junctions. An additional distinctive feature of the ubiquitin-like genes is the diversity of the C-terminal peptide extensions. The dipeptide Ser-Phe in *UBQ12* is identical to that of polyubiquitin proteins encoded by *UBQ3* and *UBQ4,* whereas a unique 15-amino acid peptide terminates the last repeat of *UBQ8.*

In addition to amino acid replacements, several ubiquitin-like repeats differ in the number of amino acids from the 76-amino acids found in ubiquitin, with several cases amino acid insertions and one of an amino acid deletion (Figure 5). Whereas all repeats of *UBQ7* and *UBQ12* encode 76-amino acid proteins, one of the four repeats in *UBQ9* is 77 amino acids long. Four of the eight repeats in *UBQ8* differ in length, with repeats one, three, four and seven containing 78, 83, 81 and 83 amino acids, respectively. In addition to extra amino acids near the amino terminus, *UBQ8* repeat three has two extra amino acids at position 47 of wild-type ubiquitin and repeat seven has one additional amino acid at position 68. Only one *UBQ8* repeat, repeat five, encodes a smaller protein of 74 amino acids. This results from a four amino acid deletion beginning at residue 71 and a two amino acid insertion at beginning position 39. *UBQ12* has three in-frame translational stop codons, two in the first repeat and one in the second (asterisks in Figure 5). Notable is the absence of in-frame stop codons in *UBQ8* even though it encodes eight ubiquitin-like coding regions encompassing 627 codons.

The nucleotide sequence corresponding to *UBQ7* is shown in Figure 6. Two observations suggest the presence of two introns within the coding region, one in each repeat at different positions. First, the amino acid sequence identity to ubiquitin in both repeats disappears immediately after a GT nucleotide sequence and resumes after an AG sequence (Figure 6). These features are characteristic of the 5′ and 3′ borders of introns (HANLEY and SCHULER 1988). Second, removal of the putative introns restores an in-frame ubiquitin coding region to both repeats. The three other ubiquitin-like genes, *UBQ8, UBQ9* and *UBQ12* differ from *UBQ7* in that they appear to have uninterrupted coding regions (data not shown).

The extent of nucleotide substitutions between all ubiquitin-like repeats was determined (*p*-distance value) and shown in Figure 7. The level of nucleotide substitution between *UBQ7* repeat two and any other repeat is much higher than any other comparison. The observed amino acid identity (see above) between *UBQ9* repeat one and *UBQ12* repeat one was confirmed and extended when observed at the nucleotide level. Repeats 9–1 and 12–1 are only 4% diverged at the nucleotide level. In addition, other comparisons between these two genes revealed highly similar repeats; repeats 9–2 and 12–2 are only 10% diverged and repeats 9–4 and 12–5 (the terminal repeats) are only 8% diverged (Figure 7).

Arabidopsis ubiquitin gene family: To determine the approximate number of ubiquitin genes in *A. thaliana* ecotype Columbia, total cellular DNA was digested singly with five different restriction enzymes, and the number of ubiquitin fragments present in the genome were visualized by hybridization of the resulting DNA gel blot with the chicken ubiquitin cDNA probe (Figure 8). The number of distinct ubiquitin fragments visualized with five different enzymes ranged from a minimum of 9 to a maximum of 12.

The ubiquitin-hybridizing pattern generated with *Hin*dIII consisted of well-separated fragments that were of lower molecular weight than those produced by the other enzymes. Furthermore, from sequence determination there were no *Hin*dIII sites in any of the eight ubiquitin coding regions (data not shown). For these reasons, the correspondence of ubiquitin genes to *Hin*dIII genomic restriction fragments was determined in one of two ways (Figure 9). For *UBQ3, UBQ7, UBQ9* and *UBQ14,* cloned *Hin*dIII fragments containing the coding regions were electrophoresed adjacent to genomic DNA, and the specific fragment identified by comigration of ubiquitin hybridizing fragments (Figure 9, A and B; data not shown for *UBQ9* and *UBQ14*). Alternatively, DNA gel blots containing genomic DNA digested with *Hin*dIII were hybridized first with a restriction fragment mapping adjacent to the coding region (*UBQ8* and *UBQ12*), or with a gene-specific oligonucleotide (*UBQ10* and *UBQ11*), and then the blots were rehybridized with the ubiquitin coding region to identify the specific fragment (Figure 9, C and D; data not shown for *UBQ8* and *UBQ12*). Genomic *Hin*dIII fragments corresponding to *UBQ1, 2, 4, 5, 6* and *13* genes were assigned previously (BURKE *et al.* 1988; CALLIS *et al.* 1990; SUN and CALLIS 1993). The relative hybridization differences of ubiquitin genomic fragments observed between Figure 9, A and B and C and D are a reflection of the different hybridization probes used (see legend), the fragment's nucleotide identity to the hybrid-

```
gctttggagtctgaagttacataagcaaattcgccattttgttacttctctccaattaat   60
gtgaagcggctagatttcaatggagttagtgtgtatagtggtgggagagtctgggagctg  120
atggtgagtgtgccagattctttttaatatgagacaaaatattattctttctatgcttgt  180
ttacatctttttttcaaccactagtagaaacctgaatattctctcgactctgtttagata  240
tgattctgatgtagttggcatatcttgtgtaatccctccactctcaaatttctaacaatt  360
cttaatttttgggttgaaacggaagtaatattttgttcttccatctaaccaaaggtcgta  420
tatttgcaaaccaattgaagtagaaattaacttaaaccgggaactggtttaagagatatt  480
aatatatgaccgaagacattgggtttaaaaagtcagaataataaaaagttggcgtagtct  540
aaacgccgacccagaggccctctcatacctttaccagtgcctatatatatatatatatat  600
aaacagactaattactattccccatcgcacagaccccctaagaatccgagagagaagaa  660
gagataatgcagatcttcgtcaaaaccctcaccggcaaaactataaccctagaagttgag  720
      METGlnIlePheValLysThrLeuThrGlyLysThrIleThrLeuGluValGlu

agcagcgacaccatcgacaatgttaaagccaaaatccaggtttaatttagggtttcttct  780
SerSerAspThrIleAspAsnValLysAlaLysIleGln

cttctttatcccctcaaacgattccttagttcctctgaatctctacttgttcttaccttg  840
aattgtataattaagaattgtttcgagtaggaatctctagtaatcttgaagaacatggat  900
ttgatttacttgagaatagctaaaaagtgtttgtgaggattgatagtttgtttaacgatc  960
tttatctgagctcgcttttgttcgaatctttggtaaatcattttggggtttcaggacaaa 1020
                                                        AspLys

gagggcataccacctgatcaacagaggctgatttttgctggtaagcaattggaagatggc 1080
GluGlyIleProProAspGlnGlnArgLeuIlePheAlaGlyLysGlnLeuGluAspGly

cggaccttagctgattacaacatccagaaagagtctactcttcatcttgtcctcaggctc 1160
ArgThrLeuAlaAspTyrAsnIleGlnLysGluSerThrLeuHisLeuValLeuArgLeu

agaggtggaaccatgatcaaggtgaagacactcactggaaaagaaatcgagattgatatc 1200
ArgGlyGlyThrMetIleLysValLysThrLeuThrGlyLysGluIleGluIleAspIle

gaaccaaccgacactattgatcggatcaaagaacgtgttgaagagaaagaaggcatccct 1260
GluProThrAspThrIleAspArgIleLysGluArgValGluGluLysGluGlyIlePro

cctgttcaacaaaggtaaaaacaaaacttacgtttttctcttatgacttgacgagtttgt 1320
ProValGlnGlnArg
gtttgtcttggtgcatttaagcctcaattgagtcccatcctgaacattgtttgcagtatt 1380
tttattgcatgtgttagtcccatcctgaacattgtttgcagtatataactctgtggtacg 1440
ctctgaattctcaggctcatatatgccggaaaacagcttgctgatgacaaaacggccaaa 1500
               LeuIleTyrAlaGlyLysGlnLeuAlaAspAspLysThrAlaLys

gattatgcgatagagggaggctctgttcttcatttggttcttgctcttaggggtggtctt 1560
AspTyrAlaIleGluGlyGlySerValLeuHisLeuValLeuAlaLeuArgGlyGlyLeu

ctctgatcttaataaataagctt 1583
Leu *
```

FIGURE 6.—Nucleotide sequence of the *UBQ7* gene. The nucleotide sequence of a 1585-bp genomic *Hin*dIII fragment containing the coding region of the ubiquitin-like gene *UBQ7* is shown. Two nucleotides, AA, constituting the 5′ *Hin*dIII site, are not shown. The derived amino acid sequence for the region with similarity to ubiquitin is shown underneath. The first and last two nucleotides of the two putative introns in the coding region are underlined. *, stop codon.

ization probe used and the number of repeats contained within a particular fragment or band. As expected from sequence data, each gene was represented by a single genomic *Hin*dIII fragment that corresponded to a restriction fragment visualized by hybridization with the ubiquitin coding region. All of the *Hin*dIII fragments visualized in genomic DNA were accounted for, corresponding to one or more cloned gene.

DNA isolated from a population of individuals was

	7-1	7-2	8-1	8-2	8-3	8-4	8-5	8-6	8-7	8-8	9-1	9-2	9-3	9-4	12-1	12-2	12-3	12-4	12-5	13-2	13-3	13-4
7-2	0.44																					
8-1	0.28	0.45																				
8-2	0.26	0.44	0.24																			
8-3	0.32	0.51	0.35	0.30																		
8-4	0.31	0.50	0.30	0.33	0.31																	
8-5	0.26	0.47	0.31	0.24	0.27	0.36																
8-6	0.32	0.55	0.38	0.32	0.22	0.36	0.33															
8-7	0.35	0.51	0.32	0.34	0.35	0.24	0.37	0.34														
8-8	0.29	0.45	0.32	0.29	0.31	0.36	0.26	0.36	0.38													
9-1	0.30	0.44	0.17	0.27	0.37	0.36	0.32	0.40	0.38	0.32												
9-2	0.21	0.41	0.21	0.23	0.27	0.22	0.28	0.29	0.26	0.25	0.25											
9-3	0.23	0.45	0.27	0.20	0.29	0.33	0.21	0.33	0.33	0.25	0.30	0.22										
9-4	0.23	0.44	0.27	0.24	0.29	0.32	0.25	0.33	0.37	0.22	0.29	0.21	0.21									
12-1	0.31	0.45	0.17	0.27	0.37	0.37	0.32	0.42	0.40	0.33	**0.06**	0.26	0.30	0.29								
12-2	0.21	0.43	0.27	0.24	0.26	0.23	0.27	0.27	0.29	0.26	0.30	**0.10**	0.21	0.24	0.30							
12-3	0.27	0.45	0.29	0.23	0.34	0.35	0.23	0.38	0.38	0.28	0.30	0.25	0.13	0.24	0.31	0.26						
12-4	0.22	0.44	0.27	0.19	0.26	0.32	0.22	0.29	0.33	0.26	0.28	0.21	0.12	0.17	0.28	0.16	0.17					
12-5	0.24	0.43	0.24	0.23	0.31	0.32	0.26	0.34	0.36	0.23	0.25	0.22	0.20	**0.08**	0.25	0.25	0.23	0.16				
13-2	0.25	0.42	0.25	0.23	0.31	0.30	0.28	0.34	0.34	0.26	0.29	0.17	0.22	0.21	0.29	0.19	0.25	0.21	0.21			
13-3	0.25	0.42	0.25	0.23	0.29	0.32	0.22	0.31	0.33	0.26	0.28	0.23	0.17	0.21	0.29	0.19	0.20	0.15	0.20	0.22		
13-4	0.23	0.44	0.21	0.22	0.28	0.30	0.24	0.29	0.30	0.27	0.26	0.21	0.21	0.19	0.27	0.22	0.23	0.17	0.18	0.21	0.14	
13-5	0.26	0.45	0.22	0.24	0.31	0.32	0.27	0.34	0.33	0.30	0.26	0.22	0.22	0.24	0.26	0.25	0.24	0.19	0.21	0.22	0.14	**0.10**

FIGURE 7.— *P* distance values for all nucleotide subtitutions from the pairwise comparison of the Arabidopsis ubiquitin-like repeats. The fraction of nucleotide sites that differ between pairwise comparisons of repeats was determined using MEGA (KUMAR *et al.* 1993) and is indicated. *P* distance values ≤0.1 are in boldface type.

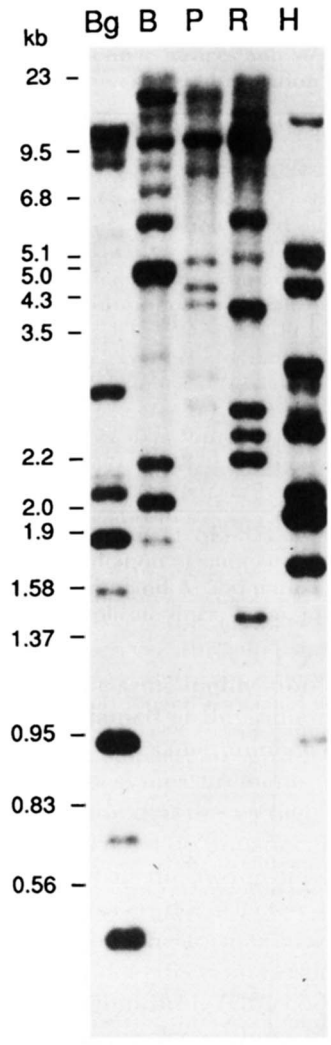

FIGURE 8.—Genomic DNA gel blot analysis of Arabidopsis ubiquitin genes. *A. thaliana* ecotype Columbia genomic DNA was isolated and digested singly with five different restriction enzymes noted above the lane. The DNA was fractionated on a 0.7% agarose gel, transferred to a nylon membrane and hybridized with radiolabelled ubiquitin cDNA as described in MATERIALS AND METHODS. Bg, *Bgl*II; B, *Bam*HI; P, *Pst*I; R, *Eco*RI; H, *Hin*dIII. Molecular weight markers indicated to the left are from lambda restricted with *Hin*dIII and *Eco*RI.

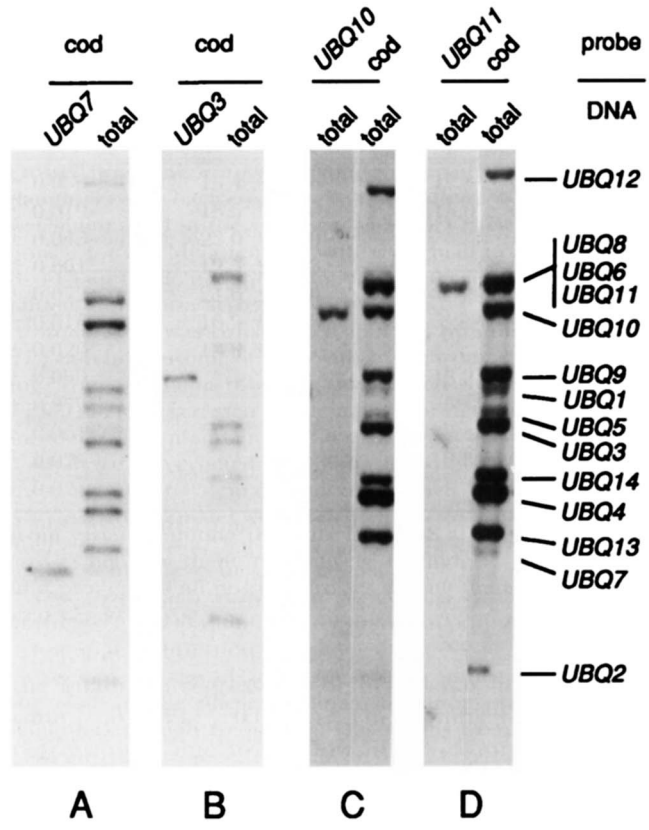

FIGURE 9.—Correspondence of ubiquitin *Hin*dIII genomic fragments with isolated ubiquitin genes. (A and B) Four micrograms of Arabidopsis ecotype Columbia total cellular DNA (lanes designated as total) and 200 ng of plasmid DNA containing the coding regions for *UBQ3* (p1585) and *UBQ7* (p1273) were digested with *Hin*dIII and subjected to DNA gel blot analysis using the radiolabeled ubiquitin coding regions either from the Arabidopsis *UBQ3* gene (A) or from a chicken ubiquitin cDNA (B) as hybridization probes (see MATERIALS AND METHODS). (C and D). A DNA gel blot containing 4 μg of Arabidopsis total cellular DNA digested with *Hin*dIII was hybridized sequentially with a gene-specific oligonucleotide for *UBQ10* (C) or *UBQ11* (D), then with radiolabeled ubiquitin coding region from the Arabidopsis *UBQ4* gene (see MATERIALS AND METHODS). The resulting correspondence of ubiquitin-hybridizing genomic fragments with isolated ubiquitin genes is designated to the right. The designations for *UBQ1, 2, 4, 5, 6* and *13* were determined previously (BURKE *et al.* 1988; CALLIS *et al.* 1990; SUN and CALLIS 1993); data not shown for *UBQ7, 8, 12* and *14*. Above the line indicates the probe used for each lane, below the line indicates whether total cellular Arabidopsis (total) DNA or a plasmid containing ubiquitin sequences (*UBQ3, UBQ7*) was fractionated in the lane below. Cod, ubiquitin coding region with the specific DNA used designated as above.

used for these analyses and multiple libraries were used as sources of genomic clones. Thus, it was possible that some of the *Hin*dIII fragments represented genes not present in every individual, but rather polymorphic alleles segregating in the population. To determine whether all 14 ubiquitin genes are loci present in every individual, DNA was isolated from 13 individual plants and from 10 progeny of 1 selfed plant. The ubiquitin-hybridizing *Hin*dIII fragment pattern for DNA from individuals was compared with the pattern obtained from DNA isolated from a population of plants. All DNA samples from individual plants, including siblings, yielded an identical hybridization pattern and this pattern was identical to that observed for DNA isolated from the pool of individuals (data not shown).

Chromosomal mapping of Ub genes: Members of the F2 generation from a cross between the *A. thaliana* ecotypes Landsberg erecta and Columbia were used to map ubiquitin genes. Using genomic DNA restriction fragments that flank the coding regions as hybridization probes, genomic DNA gel blot analysis with DNAs from the two ecotypes uncovered restriction fragment polymorphisms (RFLPs) for specific genes that could be used to

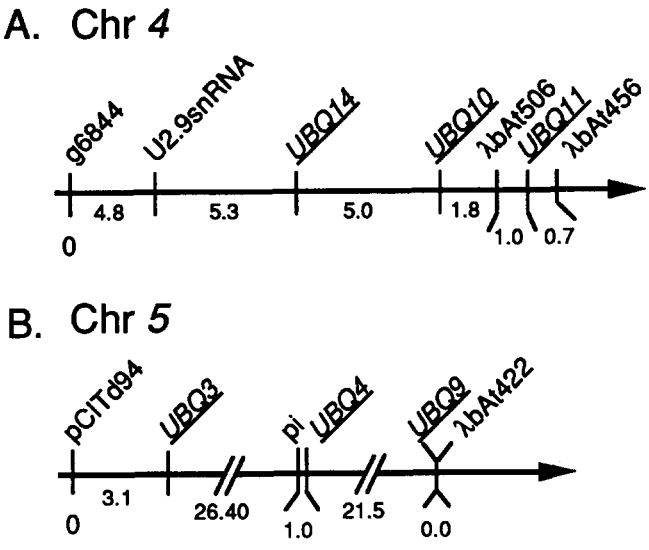

FIGURE 10.—Diagrammatic representation of the relative map positions of the ubiquitin genes *UBQ10, UBQ11* and *UBQ14* on chromosome *4* and *UBQ3, UBQ4* and *UBQ9* on chromosome *5*. Markers are indicated above the line; below the line are the relative map distances between markers in centimorgans for each interval. The most distal marker for one chromosome end is indicated below the line with a 0; conventions for the designation of Arabidopsis chromosomes and the source and identity of the adjacent markers are in (HAUGE *et al.* 1993). Diagonal lines interrupting the chromosome represent intervals not drawn to scale. The ubiquitin genes are underlined. Arrow indicates continuation of the rest of the chromosome. For both chromosomes the total recombinational distance is approximately 100 cM (HAUGE *et al.* 1993). The Arabidopsis centromeres have not been mapped (HAUGE *et al.* 1993).

locate chromosomal positions (data not shown). We were unable to identify usable RFLPs for *UBQ7, UBQ8* and *UBQ12*. The chromosomal positions for all the polyubiquitin genes, *UBQ3, UBQ4, UBQ10, UBQ11* and *UBQ14* and two ubiquitin-like genes, *UBQ9* and *UBQ13*, were determined. *UBQ10, UBQ11* and *UBQ14* are linked on chromosome *4* (Figure 10A). *UBQ3, UBQ4* and *UBQ9* are on chromosome *5*, but are located considerable distances apart from each other (Figure 10B). The previously characterized ubiquitin-like gene *UBQ13* mapped to chromosome *1* (data not shown).

Analysis of ubiquitin sequences from Arabidopsis: To further analyze the relationships between ubiquitin sequences in Arabidopsis, the extent of synonymous substitutions between polyubiquitin and ubiquitin extension protein repeats was determined by obtaining K_s, the number of synonymous substitutions per site using the method and correction factor of LI (LI 1985; LI 1993). For the polyubiquitin genes, all *intra*genic repeat comparisons gave K_s values that were either greater than one or were not calculable because the substitution rate was high; both results indicated saturation for synonymous substitutions (Figure 11, A and B). Also, all *inter*genic comparisons using repeats between the two subtypes of polyubiquitin genes also gave K_s values greater than one or was unable to give a specific value (data not shown), again indicating saturation.

Intergenic comparisons using repeats from polyubiquitin genes of the same subtype also indicated saturation or were nearly saturated for synonymous substitutions with the exceptions of the specific repeats with lower p_s values (Figure 11A for the *UBQ3/ UBQ4* subtype; B for the *UBQ10/ UBQ11/ UBQ14* subtype). For example, *UBQ3*-1 has a K_s value less than one for only one polyubiquitin repeat, *UBQ4*-1. *UBQ10*-1 has K_s values less than one for only two other repeats, *UBQ11*-1 and *UBQ14*-1. Similarly, the terminal repeats show saturation with almost all other repeats except for the terminal repeats from the same subtype. The one exception is *UBQ10*-6, which is also not saturated when compared with one of its internal repeats, *UBQ10*-4.

Several internal repeats of *UBQ10*, as well as repeats *UBQ11*-2, *UBQ14*-2 and *UBQ14*-3 have K_s values less than one when compared with each other (Figure 11B). Their relationships were examined further using neighbor-joining analysis (Figure 12A). A minimum consensus tree containing repeats 11–2, 10–2, 10–3, 10–4, 14–2, and 14–3 showed that repeats 10–2 and 14–2 form a separate branch. Most closely related to this branch was repeat 11–2. Weaker relationships that may not be significant were found between two internal repeats of *UBQ10*, 10–3 and 10–4 and *UBQ14*-3.

When compared with polyubiquitin and ubiquitin-like repeats, each ubiquitin repeat from the four ubiquitin extension protein genes was saturated with respect to synonymous substitutions so that their relationships to these genes can not be discerned. However, *UBQ1* is not saturated with respect to *UBQ2*, the other gene encoding ubiquitin and the same 52-amino acid extension. This is also true for the *UBQ5/ UBQ6* comparison; these two genes encode nearly the same 81-amino acid protein in addition to ubiquitin and are not saturated for synonymous substitutions (data not shown).

The ubiquitin-like sequences were compared with themselves and with the polyubiquitin repeats; K_s (see above) and K_a, the number of nonsynonymous substitutions per site, were calculated. None of the repeat comparisons were saturated for nonsynonymous substitutions. However, most repeat comparisons revealed saturation for synonymous substitutions. Both repeats of *UBQ7* were saturated for synonymous substitutions when compared individually with every other polyubiquitin or ubiquitin-like repeat (data not shown). Nonsaturation for synonymous sites was observed between specific repeats of *UBQ8, UBQ9* and *UBQ12* (Figure 13). Interestingly nonsaturation also was observed between the repeats *UBQ4*-1 and *UBQ8*-1, *UBQ4*-1 and *UBQ9*-1 and *UBQ4*-5 and *UBQ13*-5. Several repeats of the polyubiquitin gene *UBQ3* were also not saturated when compared with several *UBQ13* repeats (Figure 13).

The relationships among the ubiquitin-like genes

A.

	3-1	4-1	3-2	4-2	3-3	4-3	3-4	4-4
4-1	0.305 0.085							
3-2	+	+						
4-2	o	o	0.483 0.135					
3-3	o	o	0.980 0.255	+				
4-3	o	+	0.938 0.263	+	0.641 0.206			
3-4	o	o	+	+	o	+		
4-4	o	o	+	+	+	+	0.888 0.222	
4-5	o	o	+	+	+	+	0.232 0.080	+

B.

	10-1	11-1	14-1	10-2	11-2	14-2	10-3	11-3	14-3	10-4	14-4	10-5
11-1	0.237 0.077											
14-1	0.119 0.054	0.129 0.050										
10-2	+	0.827 0.210	+									
11-2	+	+	+	0.685 0.196								
14-2	o	+	+	0.129 0.050	0.680 0.178							
10-3	o	o	o	+	0.674 0.171	+						
11-3	+	+	+	+	+	+	+					
14-3	o	+	+	+	0.938 0.253	+	0.433 0.117	+				
10-4	o	+	o	o	0.818 .226	+	0.377 0.114	+	0.489 0.152			
14-4	+	+	+	+	+	+	+	0.348 0.101	+	+		
10-5	+	+	+	o	+	0.680 0.178	+	+	0.938 0.253	0.718 0.217	0.939 0.268	
10-6	+	+	+	+	0.973 0.251	+	+	0.226 0.080	+	0.49 0.131	0.490 0.131	+

FIGURE 11.—Determination of the rate of synonomous substitutions (K_s) between polyubiquitin genes. K_s values and standard errors were determined using the method of LI (LI *et al.* 1985; LI 1993). +, $K_s > 1$; o, number can not be calculated. (A) K_s values between repeats of the *UBQ3/4* subtype. (B) K_s values between repeats of the *UBQ10/11/14* subtype.

were analyzed further in a neighbor-joining analysis (Figure 12B). Specific repeats of *UBQ9* and *UBQ12* formed separate branches; 9–1 with 12–1, 9–3 with 12–3 and 12–4 and 9–4 with 12–5. The first repeats of *UBQ8, UBQ9* and *UBQ12* formed a separate branch linked to the second repeat of *UBQ7,* although some of these branches were weakly supported as evidenced by the bootstrap value (Figure 12B). Repeat 9–2 formed a separate branch with *UBQ12* two repeats of *UBQ8,* albeit weakly. In contrast to the polyubiquitin genes, Neighbor-joining analysis revealed that the ubiquitin-like genes had several instances of strong relationships between intragenic repeats (Figure 12B). Internal repeats of *UBQ8* formed separate branches with themselves; 8–4 with 8–7, 8–3 with 8–6 and, weakly, 8–5 with 8–8. Less significant relationships (bootstrap values were <40%) were found for repeats 8–1, 8–2 and for all the repeats of *UBQ7* and *UBQ13.*

DISCUSSION

This paper reports the nucleotide and derived amino acid sequence for eight ubiquitin genes from *A. thaliana* ecotype Columbia. Lack of segregating ubiquitin hy-

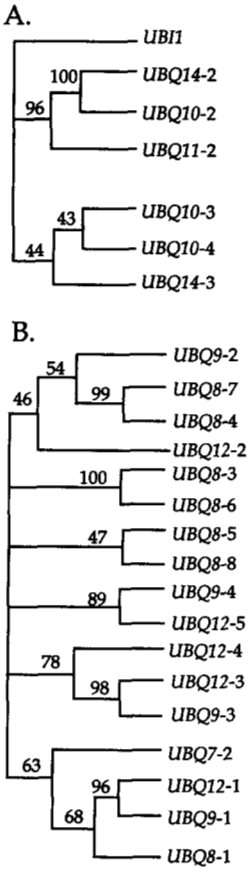

FIGURE 12.—Neighbor-joining analysis of Arabidopsis ubiquitin repeats. The nucleotide sequences of specific ubiquitin repeats were analyzed by 100 replicas of the neighbor-joining method and a consensus minimum tree obtained (see MATERIALS AND METHODS). Yeast *UBI1* ubiquitin sequence was used as the outgroup sequence. The tree is unrooted, and there is no significance to branch length. Bootstrap values above 40% for 100 replicas are shown above the appropriate branch point. (A). Neighbor-joining analysis of polyubiquitin repeats with Ks values less than one. (B). Neighbor-joining analysis of all ubiquitin-like repeats. Individual repeats with no relationship to other repeats are not shown.

bridizing restriction fragments in individual progeny from a selfed plant indicate that each ubiquitin sequence is present in every individual. Combining the previously isolated ubiquitin genes with the eight presented here, we have identified 14 ubiquitin genes in *A. thaliana* ecotype Columbia. It is likely that they represent the complement of ubiquitin genes in this organism because all genomic *Hin*dIII restriction fragments that hybridize to the ubiquitin coding region correspond to at least one of these 14 cloned ubiquitin genes, genomic DNA gel blot analysis with four other restriction enzymes suggests that the number of ubiquitin genes is not larger, hybridization and washing of genomic DNA blots with ubiquitin coding region probes at T_m −50° did not identify additional hybridizing bands (data not shown) and extensive screening of two total genomic DNA libraries and two libraries containing genomic DNA enriched for 2- and 5-kb *Hin*dIII restriction fragments failed to isolate additional ubiquitin genes. Although it is difficult to be certain, based on these criteria, we conclude that we have isolated all of the ubiquitin genes of the Arabidopsis genome from the Columbia ecotype.

In higher plants and animals, ubiquitin genes can be divided into three types of genes: the polyubiquitin genes, the ubiquitin-like genes and the ubiquitin-extension protein genes (reviewed in SCHLESINGER and BOND 1987; CALLIS and VIERSTRA 1989). The first two types encode ubiquitin polyproteins, with the latter encoding at least one repeat with amino acid replacements. The ubiquitin-extension protein genes contain one ubiquitin coding region followed by one of two different ribosomal proteins and have been described previously (CALLIS *et al.* 1990). The Arabidopsis ubiquitin gene family of 14 members consists of a total of five polyubiquitin genes, five ubiquitin-like genes and four ubiquitin-extension genes.

The Arabidopsis polyubiquitin genes differ in the number of ubiquitin coding regions per gene (from 3 to 6) and are all expressed. *UBQ4* has been determined previously to encode a mRNA (BURKE *et al.* 1988). Isolation of cDNAs for *UBQ3, UBQ10* and *UBQ11* indicates that they also are expressed. Although a cDNA has not yet been isolated for *UBQ14*, introduction of a chimeric gene containing the 5′ flanking region of the *UBQ14* gene upstream of the coding region for a marker enzyme into Arabidopsis leaves via particle bombardment resulted in enzyme activity, indicating that the promoter was functional (J. CALLIS, unpublished data). This suggests that *UBQ14* also is expressed *in vivo*.

The presence of multiple, expressed polyubiquitin genes that differ in repeat number is typical in eukaryotes, with the exception of *Saccharomyces cerevisiae* (OZKAYNAK *et al.* 1987) and possibly *Neurospora crassa* (TACCIOLI *et al.* 1989) where each has only one polyubiquitin gene. Dictyostelium expresses at least five different polyubiquitin genes including seven, five and three ubiquitin repeat genes (OHMACHI *et al.* 1989). To date, two expressed polyubiquitin genes have been identified in humans, one with nine and one with three repeats (WIBORG *et al.* 1985; BAKER and BOARD 1987a, 1989). The total number of ubiquitin genes or expressed polyubiquitin genes is not known for any other higher plant than Arabidopsis (this work). In maize, expression of two polyubiquitin genes each with seven repeats has been characterized, but the presence of an additional ubiquitin transcript-size class suggests that there is at least one additional polyubiquitin gene with a smaller number of repeats (CHRISTENSEN and QUAIL 1989; CHRISTENSEN *et al.* 1992). In parsley, there is only one major polyubiquitin transcript-size class that contains at least two different hexameric polyubiquitin transcripts; whether there are additional transcripts of the same size but from different genes is not known (KAWALLECK

A.

	3-1	3-2	3-3	3-4	4-1	4-2	4-3	4-4	4-5
8-1	+	+	o	o	0.691 0.179	o	+	o	o
9-1	0.938 0.268	o	o	o	0.743 0.223	o	o	o	o
12-1	+	o	o	o	0.829 0.317	o	o	o	o
13-2	o	0.611 0.165	+	o	+	+	0.819 0.199	+	o
13-3	o	o	o	0.610 0.189	o	o	o	0.870 0.249	+
13-4	o	+	o	+	o	+	+	0.980 0.294	+
13-5	o	+	o	0.573 0.170	o	+	0.888 0.222	+	0.679 0.195

B.

	8-1	8-2	8-3	8-4	8-5	8-6	8-7	9-1	9-2	9-3	9-4
8-3	o	0.946 0.283									
8-4	+	+	0.782 0.213								
8-5	o	+	0.900 0.303	0.249 0.046	+						
8-6	o	0.99 0.26	0.186 0.071	0.986 0.51	0.932 .25						
8-7	+	+	0.730 0.193	0.404 0.132	+	0.889 0.227					
12-1	0.456 0.131	+	o	o	o	o	+	0.106 0.050	+	o	o
12-2	o	o	0.881 0.224	0.618 0.187	+	0.855 0.240+	0.896 0.279	o	0.588 0.187	+	+
12-3	o	o	+	o	0.717 0.209	o	o	o	+	0.408 0.134	+
12-4	o	+	+	+	0.919 0.219	+	+	o	o	0.452 0.122	+
12-5	o	+	o	o	+	o	o	+	o	+	0.278 0.0941

FIGURE 13.—Determination of the rate of synonomous substitutions (K_s) between ubiquitin-like genes. Methods and notation as described for Figure 11. (A) K_s value for comparison among ubiquitin-like repeats. (B) Comparison between ubiquitin-like repeats and polyubiquitin repeats.

et al. 1993). In sunflower, two six-repeat and one four-repeat polyubiquitin genes are expressed (BINET *et al.* 1989, 1991). Potato (*Solanum tuberosum*) tubers express at least three polyubiquitin genes, one containing six repeats and two containing seven repeats (GARBINO *et al.* 1992). Three different polyubiquitin cDNAs from *Nicotiana sylvestris* have been isolated, one with seven or eight repeats, one with six repeats and two with five repeats (GENSCHIK *et al.* 1992). Three polyubiquitin genes from flax, *Linum usitatissimum,* have been described with four additional ubiquitin sequences identified, but not yet isolated (AGARWAL and CULLIS 1991).

Identification of chromosomal location combined with nucleotide comparisons revealed several interesting relationships among the Arabidopsis polyubiquitin genes. Three criteria presented here support the hypothesis that the polyubiquitin genes can be grouped into two subtypes: the nature of their C-terminal amino acids, the nucleotide identity relationships in the coding region and 3′ untranslated regions and their relative chromosomal locations. In contrast, the number of ubiquitin coding regions does not appear to be useful as a classifying criterion. Whereas *UBQ3* and *UBQ4* differ in the number of ubiquitin repeats, they have the same C-terminal amino acids, the highest nucleotide identity to each other than to any other gene and, although not closely linked, map to the same chromosome. Similarly, *UBQ10, 11* and *14* differ from each other in repeat number but have the same C-terminal amino acid, exhibit higher nucleotide identity to each other than to *UBQ3* and *UBQ4* and are genetically linked.

The presence of distinct subtypes within a gene family and the presence of both linked and dispersed family

members of varying relatedness are characteristics of several other plant gene families. A few examples are the tomato *cab* (PICHERSKY *et al.* 1985) and *rbcS* (SUGITA *et al.* 1987) families, the petunia actin family (MCLEAN *et al.* 1988), the Arabidopsis *rbcS* family (KREBBERS *et al.* 1988) and the maize 19- and 22-kD zeins (HEIDECKER *et al.* 1991).

Arabidopsis polyubiquitin genes are distinct from other gene families in that there are no nonsynonymous substitutions between genes and most of the repeat comparisons are saturated for synonymous substitutions. The polyubiquitin repeats also are saturated for synonymous substitutions when compared with ubiquitin repeats from the ubiquitin extension protein genes. This indicates that aside from a few comparisons, the relationships between repeats can not be discerned among these genes. The exception is comparisons between *UBQ10, UBQ11* and *UBQ14*. The p_s and K_s values and neighbor-joining analysis suggest that within this subtype, *UBQ10* and *UBQ14* are more closely related to each other than either is to *UBQ11*. Interestingly, the higher nucleotide identity does not correlate with genetic distance, with *UBQ10* and *UBQ11* having a smaller map distance than *UBQ10* and *UBQ14*.

The ubiquitin gene family contains five genes that encode ubiquitin-like proteins. Although not definitive, several pieces of evidence suggest that none of these genes are expressed. No corresponding cDNAs were isolated from a screen of 200,000 independent clones from cDNA library made from mature leaf poly A^+ mRNA. From this screening, 12 different polyubiquitin cDNAs corresponding to the *UBQ3, UBQ10* and *UBQ11* genes were isolated. Hybridization of unique DNA fragments flanking the coding region of the *UBQ7* and *12* genes to total RNA gel blots did not detect the presence of corresponding RNAs (data not shown). For *UBQ7*, linkage of 700 bp of the 5′ flanking region to a reporter sequence, encoding β-glucuronidase (GUS), did not produce detectable GUS activity in transgenic tobacco (data not shown). In parallel experiments, readily detectable levels of GUS were produced using gene constructions containing the 5′ flanking regions of *UBQ1* and *UBQ6* (CALLIS *et al.* 1990). Because exhaustive expression studies have not been performed, we can not eliminate the possibility that these variant proteins are synthesized in a highly restricted manner. However, our current evidence suggests that *UBQ7–9* and *UBQ12* most likely represent pseudogenes.

The Arabidopsis ubiquitin-like genes are in contrast to the ubiquitin pseudogenes in humans where several appear to have arisen from processed mRNAs (BAKER and BOARD 1987a,b; COWLAND *et al.* 1988). Another interesting aspect of these genes is that three of the five genes contain amino acid insertions and/or deletions without in-frame stop codons. This differs from the well-characterized maize zein storage protein pseudogenes that contain primarily in-frame stop codons without deletions or insertions in the coding region (KRIDL *et al.* 1984; WANDELT and FEIX 1989).

The absence of expressed ubiquitin-like proteins in Arabidopsis is in contrast to the work in several other systems where expression of ubiquitin-like proteins has been documented. These include an interferon-induced protein in cultured animal cells (HAAS *et al.* 1987), ribosomal proteins in nematodes (JONES and CANDIDO 1993), rat (OLVERA and WOOK 1993) and Dictyostelium (JONES and CANDIDO 1993), radiation-induced protein RAD23 in yeast (WATKINS *et al.* 1993) and baculovirus-encoded protein present in infected lepidopteran cells (GUARINO 1990).

The origin of the introns in the coding region of *UBQ7* is intriguing but unknown. Introns are not found in the coding regions of the Arabidopsis polyubiquitin genes or the other ubiquitin-like genes, but have been identified in the 5′ untranslated regions for the polyubiquitin genes *UBQ3, UBQ10, UBQ11* and *UBQ14* (NORRIS *et al.* 1993; J. CALLIS, unpublished data). Whether an intron is present in the 5′ untranslated regions of fifth polyubiquitin gene, *UBQ4*, is unknown at present. Among all the polyubiquitin genes and pseudogenes sequenced to date, only one other gene in addition to *A. thaliana UBQ7, UbiA* from *Caenorhabditis elegans*, contains introns in the coding region (GRAHAM *et al.* 1989). *C. elegans UbiA* contains 11 ubiquitin repeats, four of which contain an intron interrupting the same position in each repeat. This is in contrast to *UBQ7*, where the two introns are in different positions in each repeat, with neither matching the intron position of *C. elegans UbiA*. The nonconcordance of intron position between the two *UBQ7* repeats reduces the likelihood that this ubiquitin dimer arose by simple duplication of an intron-containing ubiquitin monomer. Identification of *UBQ7* homologs will be necessary to understand the origins of this interesting ubiquitin-like gene.

A model for the evolution of the Arabidopsis polyubiquitin and ubiquitin-like genes must account for the current number of genes, the pattern of intergene nucleotide identity and the diversity of repeat number per gene. Because all eukaryotic organisms contain at least one gene with tandem repeats of the ubiquitin coding region, this gene structure most probably predates speciation of Arabidopsis. Less clear is the timing origin of the multiple polyubiquitin genes in Arabidopsis. Because most eukaryotes contain multiple polyubiquitin genes, they most probably arose from gene duplication events, as has been postulated for the origin of many multigene families (JEFFREYS and HARRIS 1982; MAEDA and SMITHIES 1986). Based on the chromosomal locations of the polyubiquitin genes and sequence divergence, the duplication event that created the two polyubiquitin subtypes could represent a more ancient duplication than the events that gave rise to the multiple members of each subtype. The creation of the two

subtypes could have predated the speciation of Arabidopsis. Alternatively, gene conversion events could homogenize those genes that reside on the same chromosome. Identification of subtypes of polyubiquitin genes in other plant species and especially in close relatives of Arabidopsis may resolve this question.

Although not definitive, several pieces of evidence support that recent duplications and unequal crossing-over events, not gene-conversion events, are responsible for the observed polyubiquitin subtype identity and repeat number differences per gene. Although gene conversion is a well-established mechanism that can homogenize genes and thus give rise to the observed high degree of inter-gene identity, it is less likely to give rise to the distinct pattern of nucleotide identity among the polyubiquitin gene subtypes (terminal repeats, independent of repeat number, share identity) and less likely to result in changes in the size of the coding region (repeat number per gene differences). Second, nucleotide changes between two genes are found dispersed along the entire coding region, rather than in just parts of the coding region; the latter would be more likely in gene-conversion events, which are not required to observe coding region boundaries. Third, if gene conversion were operating to homogenize genes, it is also likely that it would homogenize the closely linked tandem repeats present within a gene. Such distance dependence of repeated sequences has been demonstrated for the 5S rRNA gene family of Neurospora (METSENBERG *et al.* 1985) and the rDNA repeats in wheat (LASSNER and DVORAK 1985). For the Arabidopsis polyubiquitin genes, this is clearly not the case. Ubiquitin repeats within a gene have lower nucleotide identity to each other than they do to specific repeats in other genes at distant loci. It is for these reasons that we favor the hypothesis that more recent duplications and unequal crossing-over events are the mechanism responsible for the high degree of nucleotide identity among the *UBQ10, UBQ11* and *UBQ14* genes and between *UBQ3* and *UBQ4* genes.

Determination of K_s values suggest a relationship between the polyubiquitin *UBQ3/UBQ4* subtype and the ubiquitin-like genes *UBQ8, UBQ9, UBQ12* and *UBQ13*. Relevant to this is the fact that *UBQ3* and *UBQ4* share the same C-terminal amino acids with *UBQ12* and *UBQ13*. In addition, *UBQ9* maps roughly equidistant from *UBQ4* as *UBQ3* does from *UBQ4*. We hypothesize that these genes are the products of duplications from a member of the *UBQ3/UBQ4* polyubiquitin subtype (or derivatives thereof) that have undergone gene inactivations and subsequent sequence divergence.

Two different types of unequal crossing-over events to generate repeat number per gene changes are possible. One is that after duplication of a ubiquitin gene to form two tandemly repeated genes each consisting of multiple coding regions, the two undergo somatic recombination. This results in deletion of the sequence in between and creation of one ubiquitin gene. If the pairing were exact, there would be no change in repeat number between the two progenitors and the resulting single ubiquitin gene. However, if the two ubiquitin genes misalign, such that the homologous repeats are not paired, then the resulting gene will have a different repeat number than either of its progenitors. This event requires a duplication event followed by mispairing, recombination and loss of one of the genes. Such recombination events have been described for the *P* locus (ATHAMA and PETERSON 1991), the 27-kD zein gene family in maize (DAS *et al.* 1990) and marker genes in transgenic tobacco (TOVAR and LICHTENSTEIN 1992). For the creation of the members of each subgroup, this would require multiple gene duplications. Although this mechanism is formally possible, it is difficult to prove without the identification of the duplicated progenitor genes.

Alternatively, another simpler mechanism is duplication of one polyubiquitin gene to form two polyubiquitin genes at two loci, followed by unequal crossing at one of these loci. One possible pathway that illustrates this mechanism is shown in Figure 14. Fixation of one of the products of this unequal crossing-over event will result in one polyubiquitin gene with a high degree of nucleotide identity to another polyubiquitin gene, but with a different repeat number per gene. In addition, this mechanism can account readily for the observed pattern of repeat identity. Unequal exchange has been hypothesized previously for the observed evolutionary changes in the numbers of repeated domains within the coding region of a protein, for example, changes in the size of the C-terminal repeated domain of eukaryotic RNA polymerases II (ALLISON *et al.* 1985), differences in the number of tandem repeats in the third exon of the human salival proline-rich proteins (LYONS *et al.* 1988) and differences in the number of 20-amino acid repeat units in the center of the maize zein 19-kD proteins (PEDERSEN *et al.* 1982). Unequal crossing over has been hypothesized previously to be responsible for changes in the number of human ubiquitin repeats for a human ubiquitin gene (BAKER and BOARD 1989).

In conclusion, we have presented a description of the complete ubiquitin gene family from *A. thaliana* ecotype Columbia and a working hypothesis to explain the origin and evolution of the Arabidopsis polyubiquitin and ubiquitin-like genes. Gene duplications coupled with unequal crossing-over events can account for the present diversity of polyubiquitin genes and the present duplicated structure of *UBQ8*. Alternatively, changes in ubiquitin genes could be the result of gene conversion events. Whether these genes represent the products of ancient or more recent duplications will require additional analysis, for example, the isolation of ubiquitin genes from species closely related to Arabidopsis. Interspecies comparisons identifying paralogous and or-

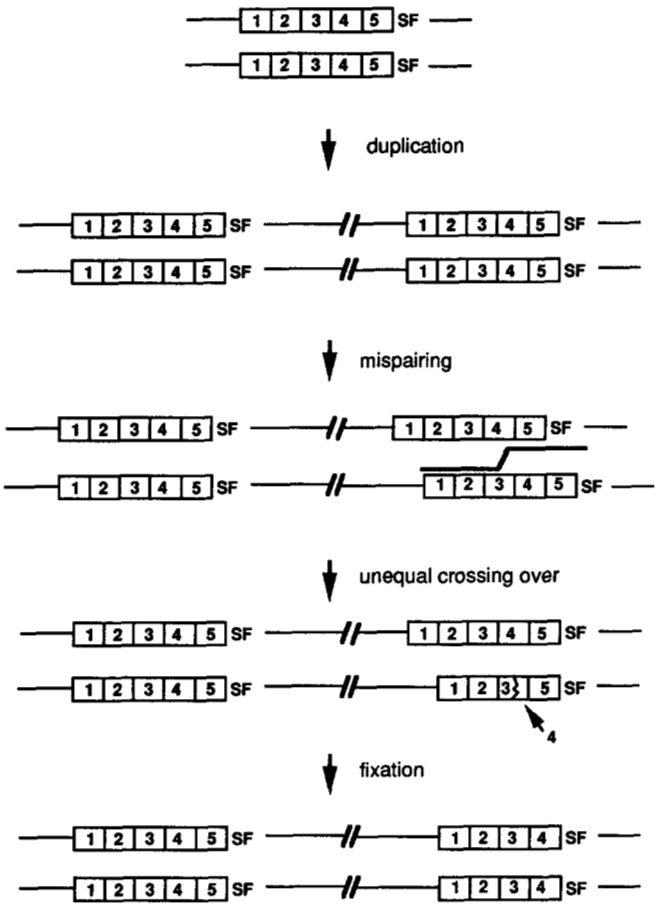

FIGURE 14.—Duplication of ubiquitin genes and unequal crossing between allelic Arabidopsis polyubiquitin genes. Diagram of possible events resulting in the present number of polyubiquitin genes and present repeat number differences between genes. Both homologous chromosomes are shown as thin black lines. The present structure of the *UBQ4* gene is shown as an example. Duplication of this gene results in an additional ubiquitin gene at an unknown distance from the progenitor gene (indicated by a break in the line). One locus undergoes misaligned pairing, and a resulting unequal crossing-over event generates one allele with additional repeats and one with less. Fixation of one allele in the population results in multiple ubiquitin genes with different repeat numbers.

thologous genes should help resolve outstanding questions about the origin and evolution of this important and interesting gene family.

We thank Drs. NIGEL CRAWFORD and HARRY KLEE for their gifts of genomic libraries, Dr. PEGGY HATFIELD, with whom the cDNA library was made and L. MEDRANO, Drs. T. BLEEKER and E. MEYEROWITZ for the mapping data. We also thank the Monsanto Company for synthesis of the oligonucleotides. Expert technical assistance was provided by SANDRA E. MEYER. We thank Drs. J. DOYLE, V. FORD, C. GASSER, L. GOTTLIEB, D. IRWIN, C. LANGLEY and V. WALBOT for helpful discussions. We also are indebted to Dr. J. DOYLE for assistance with tree analysis and Dr. V. FORD for Ks analysis. Support was provided by grants from the United States Department of Agriculture NRICGP (91-37301-6290) to R.D.V. and National Science Foundation Grants (DCB 90-05062 and DCB 93-06759) and a National Science Foundation Presidential Young Investigator Award (NSF 91-58453) to J.C.

Note added in proof: A cDNA corresponding to *UBQ7* has been identified as an Arabidopsis expressed sequence tag (GbGe310) from a flower bud cDNA library indicating that *UBQ7* is expressed in this organ at this developmental stage.

LITERATURE CITED

AGARWAL, M. L., and C. A. CULLIS, 1991 The ubiquitin-encoding multigene family of flax, *Linum usitatissimum*. Gene **99:** 69–75.

ALLISON, L. A., M. MOYLE, M. SHALES and C. J. INGLES, 1985 Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell **42:** 599–610.

ATHAMA, P., and T. PETERSON, 1991 Ac induces homologous recombination at the maize P locus. Genetics **128:** 163–173.

BAKER, R. T., and P. G. BOARD, 1987a The human ubiquitin gene family: structure of a gene and pseudogenes from the Ub B subfamily. Nucleic Acids Res. **15:** 443–463.

BAKER, R. T., and P. G. BOARD, 1987b Nucleotide sequence of a human ubiquitin UbB processed pseudogene. Nucleic Acids Res. **15:** 4352.

BAKER, R. T., and P. G. BOARD, 1989 Unequal crossover generates variation in ubiquitin coding unit number at the human UBC polyubiquitin locus. Am. J. Hum. Genet. **44:** 534–542.

BAKER, R. T., J. W. TOBIAS and A. VARSHAVSKY, 1992 Ubiquitin-specific proteases of *Saccharomyces cerevisiae*. J. Biol. Chem. **267:** 23364–23375.

BINET, M. N., A. STEINMETZ and L. H. TESSIER, 1989 The primary structure of sunflower ubiquitin. Nucleic Acids Res. **17:** 2119.

BINET, M., J.-H. WEIL and L.-H. TESSIER, 1991 Structure and expression of sunflower ubiquitin genes. Plant Mol. Biol. **17:** 395–407.

BOND, U., and M. J. SCHLESINGER, 1985 Ubiquitin is a heat shock protein in chicken embryo fibroblasts. Mol. Cell. Biol. **5:** 949–956.

BURKE, T., J. CALLIS and R. D. VIERSTRA, 1988 Characterization of a polyubiquitin gene from *Arabidopsis thaliana*. Mol. Gen. Genet. **213:** 435–443.

CALLIS, J., L. POLLMANN, J. SHANKLIN, M. WETTERN and R. D. VIERSTRA, 1989 Sequence of a cDNA from *Chlamydomonas renhardii* encoding a ubiquitin 52 amino acid extension protein. Nucleic Acids Res. **17:** 8377.

CALLIS, J., J. RAASCH and R. VIERSTRA, 1990 Ubiquitin extension proteins of *Arabidopsis thaliana:* structure, localization and expression of their promoters in transgenic tobacco. J. Biol. Chem. **265:** 12486–12493.

CALLIS, J., and R. D. VIERSTRA, 1989 Ubiquitin and ubiquitin genes in higher plants. Oxf. Surv. Plant Mol. Cell Biol. **6:** 1–30.

CHANG, C., J. L. BOWMAN, A. W. DEJOHN, E. S. LANDER and E. M. MEYEROWITZ, 1988 Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **85:** 6856–6860.

CHRISTENSEN, A., and P. QUAIL, 1989 Sequence analysis and transcriptional regulation by heat shock of polyubiquitin transcripts from maize. Plant Mol. Biol. **12:** 619–632.

CHRISTENSEN, A. H., R. A. SHARROCK and R. H. QUAIL, 1992 Maize polyubiquitin genes: structure, thermal perturbation of expression and transcript splicing, and promoter activity following transfer to protoplasts by electroporation. Plant Mol. Biol. **18:** 675–689.

COWLAND, J. B., O. WIBORG and J. VUUST, 1988 Human ubiquitiin genes: one member of the UbB gene family is a tetrameric non-processed pseudogene. FEBS Lett. **231:** 187–191.

DAS, O. P., S. LEVIS-MINZI, M. KOURY, M. BENNER and J. MESSING, 1990 Somatic rearrangement contributing to genetic diversity in maize. Proc. Natl. Acad. Sci. USA **87:** 7809–7813.

DELLAPORTA, S., 1983 A plant minipreparation: version II. Plant Mol. Biol. Rep. **1:** 19–21.

DEVEREUX, J., P. HAERBERLI and O. SMITHIES, 1984 A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:** 387–395.

FEINBERG, A. P., and B. VOGELSTEIN, 1983 A technique for radiolabeling DNA restriction fragments to a high specific activity. Anal. Biochem. **132:** 6–13.

FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22:** 521–565.

FELSENSTEIN, J., 1989 Phylip-Phylogeny interference package. Cladistics **5:** 164–166.

FINLEY, D., B. BARTEL and A. VARSHAVSKY, 1989 The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome function. Nature **338:** 394–401.

FINLEY, D., and V. CHAU, 1991 Ubiquitination. Annu. Rev. Cell Biol. **7:** 25–69.

GARBINO, J. E., D. R. ROCKHOLD and W. R. BELKNAP, 1992 Expression of stress-responsive ubiquitin genes in potato tubers. Plant Mol. Biol. **20:** 235–244.

GENSCHIK, P., Y. PARMENTIER, A. DURR, J. MARBACH, M.-C. CRIQUI *et al.*, 1992 Ubiquitin genes are differentially regulated in protoplast-derived cultures of *Nicotiana sylvestris* and in response to various stresses. Plant Mol. Biol. **20:** 897–910.

GRAHAM, R. W., D. JONES and P. M. CANDIDO, 1989 UbiA, the major polyubiquitin locus in *Caenorhadbitis elegans*, has unusual structural features and is constitutively expressed. Mol. Cell. Biol. **9:** 268–277.

GUARINO, L. A., 1990 Identification of a viral gene encoding a ubiquitin-like protein. Proc. Natl. Acad. Sci. USA **87:** 409–413.

HAAS, A. L., P. AHRENS, P. BRIGHT and H. ANKEL, 1987 Interferon induces a 15-kilodalton protein exhibiting marked homology to ubiquitin. J. Biol. Chem. **262:** 11315–11323.

HANLEY, B. A., and M. A. SCHULER, 1988 Plant intron sequences: evidence for distinct groups of introns. Nucleic Acids Res. **16:** 7159–7176.

HATFIELD, P., J. CALLIS and R. VIERSTRA, 1990 Cloning of ubiquitin activating enzyme from wheat and expression of a functional protein in *Escherichia coli*. J. Biol. Chem. **265:** 15813–15817.

HAUGE, B. M., S. M. HANLEY, S. CARTINHOUR, J. M. CHERRY, H. M. GOODMAN *et al.*, 1993 An integrated genetic/RFLP map of the Arabidopsis genome. Plant J. **3:** 745–754.

HEIDECKER, G., S. CHAUDHURI and J. MESSING, 1991 Highly clustered zein gene sequences reveal evolutionary history or the multigene family. Genomics **10:** 719–732.

HERSHKO, A., and A. CIECHANOVER, 1992 The ubiquitin system. Annu. Rev. Biochem. **61:** 761–807.

JEFFREYS, A. J., and S. HARRIS, 1982 Processes of gene duplication. Nature **296:** 9–10.

JONES, D., and E. P. M. CANDIDO, 1993 Novel ubiquitin-like ribosomal protein fusion genes from the nematodes *Caonorhabditis elegans* and *Caenorhabditis briggsae*. J. Biol. Chem. **268:** 19545–19551.

JONNALAGADDA, S., T. R. BUTT, B. R. MONIA, C. K. MIRABELLI, L. GOTLIB *et al.*, 1989 Multiple (aNH-ubiquitin) protein endoproteases in cells. J. Biol. Chem. **264:** 10637–10642.

KAWALLECK, P., I. E. SOMSSICH, M. FELDBRUGGE, K. HAHLBROCK and B. WEISSHAAR, 1993 Polyubiquitin gene expression and structural properties of the *ubi*4–2 gene in *Petroselium crispum*. Plant Mol. Biol. **21:** 673–684.

KREBBERS, E., J. SEURINCK, L. HERDIES, A. CASHMORE and M. TIMKO, 1988 Four genes in two diverged subfamilies encode the ribulose-1,5-bisphosphate carboxylase small subunit polypeptides of *Arabidopsis thaliana*. Plant Mol. Biol. **11:** 745–759.

KRIDL, J. C., J. VIEIRA, I. RUBENSTEIN and J. MESSING, 1984 Nucleotide sequence analysis of a zein genomic clone with a short open reading frame. Gene **28:** 113–118.

KUMAR, S., K. TAMURA and M. NEI, 1993 MEGA: Molecular evolutionary genetics analysis, version 1.01. The Pennsylvania State University, University Park, PA.

LANDER, E. S., and P. GREEN, 1987 Construction of multi-locus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

LANDER, E. S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. DALY *et al.*, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174–181.

LASSNER, M., and J. DVORAK, 1985 Preferential homogenization between adjacent and alternate subrepeats in wheat rDNA. Nucleic Acids Res. **14:** 5499–5512.

LI, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

LI, W.-H., C.-I. WU and C.-C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

LYONS, K. M., J. H. STEIN and O. SMITHIES, 1988 Length polymorphisms in human proline rich protein genes generated by intragenic unequal crossing-over. Genetics **120:** 267–278.

MAEDA, N., and N. SMITHIES, 1986 The evolution of multigene families: human haptoglobin genes. Annu. Rev. Gen. **20:** 81–108.

MAYER, A. N., and K. D. Wilkinson, 1989 Detection, resolution and nomenclature of multiple ubquitin carboxyl-terminal esterases from bovine calf thymus. Biochemistry **28:** 166–172.

MCLEAN, M., W. V. BAIRD, A. GERATS and R. MEAGHER, 1988 Determination of copy number and linkage relationships among five actin gene subfamilies in *Petunia hybrida*. Plant Mol. Biol. **11:** 663–672.

METSENBERG, R. L., J. N. STEVENS, E. V. SELKER and E. MORZYCKA-WROBLEWSKA, 1985 Identification and chromosomal distribution of 5S rRNA genes in Neurospora crassa. Proc. Natl. Acad. Sci. USA **82:** 2067–2071.

NEI, M., and N. SAITOU, 1987 The neighbor-joining method—a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

NORRIS, S., S. MEYER and J. CALLIS, 1993 The intron of Arabidopsis polyubiquitin genes is conserved in location and is a quantitative determinant of chimeric gene expression. Plant Mol. Biol. **21:** 895–906.

OHMACHI, T., R. GIORDA, D. R. SHAW and H. L. ENNIS, 1989 Molecular organization of developmentally regulated *Dictyostelium discoideum* ubiquitin cDNAs. Biochemistry **28:** 5226–5231.

OLVERA, J., and I. G. WOOK, 1993 The carboxyl extension of a ubiquitin-like protein is rat ribosomal protein S30. J. Biol. Chem. **268:** 17967–17974.

OZKAYNAK, E., D. FINLEY, M. J. SOLOMON and A. VARSHVASKY, 1987 The yeast ubiquitin genes: A family of natural gene fusions. EMBO J. **6:** 1429–1439.

PEDERSEN, D., J. DEVEREUX, D. R. WILSON, E. SHELDON and B. A. LARKINS, 1982 Cloning and sequence analysis reveal structural variation among related zein genes in maize. Cell **29:** 1015–1026.

PICHERSKY, E., R. BERNATZKY, S. TANKSLEY, B. BREIDENBACH, A. KAUSCH *et al.*, 1985 Molecular characterization and genetic mapping of two clusters of genes encoding chlorophyll a/b-binding proteins in *Lycopersicon esculentum* (tomato). Gene **40:** 247–258.

RIGBY, P. W., M. DIECKMAN, C. RHOADES and P. BERG, 1977 Labeling deoyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113:** 237–251.

SANDERS, P. R., J. A. WINTER, A. R. BARNASON, S. J. ROGERS and R. T. FRALEY, 1987 Comparison of cauliflower mosaic virus 35S and nopaline synthase promoters in transgenic plants. Nucleic Acids Res. **15:** 1543–1558.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SCHLESINGER, M. J., and U. BOND, 1987 Ubiquitin genes. Oxf. Surv. Eukaryotic Genes **4:** 77–89.

SHARP, P. M., and W.-H. LI, 1987 Molecular evolution of ubiquitin genes. Trends Ecol. Evol. **2:** 328–332.

SUGITA, M., and W. GRUISSEM, 1987 Developmental, organ-specific, and light-dependent expression of the tomato ribulose-1,5-bisphosphate carboxylase small subunit gene family. Proc. Natl. Acad. Sci USA **84:** 7104–7108.

SUGITA, M., T. MANZARA, E. PICHERSKY, A. CASHMORE and W. GRUISSEM, 1987 Genomic organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from tomato. Mol. Gen. Genet. **209:** 247–256.

SUN, C.-W., and J. CALLIS, 1993 Recent stable insertion of mitochondrial DNA into an Arabidopsis polyubiquitin gene by non-homologous recombination. Plant Cell **5:** 97–107.

SWOFFORD, D. L., 1990 PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0. Illinois Natural History Survey, Champaign, IL.

TACCIOLI, G. E., E. GROTEWOLD, G. O. AISEMBERG and N. D. JUDEWICZ, 1989 Ubiquitin expression in *Neurospora crassa:* cloning and sequencing of a polyubiquitin gene. Nucleic Acids Res. **17:** 6153–6165.

THEIN, S. L., and R. B. WALLACE, 1986 The use of synthetic oligonucleotides as specific hybridization probes in the diagnosis of genetic disorders, pp.33–50 in *Human Genetic Diseases: A Practical Approach*, edited by K. E. DAVIS. IRL Press, Herndon, VA.

TOVAR, J., and C. LICHTENSTEIN, 1992 Somatic and meiotic chromosomal recombination between inverted duplications in transgenic tobacco plants. Plant Cell **4:** 319–332.

VIEIRA, J., and J. MESSING, 1987 Production of single stranded plasmid. Methods Enzymol. **153:** 3–11.

VIERSTRA, R. D., S. LANGAN and E. SCHALLER, 1986 Complete amino acid sequence of ubiquitin from the higher plant *Avena sativa.* Biochemistry **25:** 3105–3108.

VIJAY-KUMAR, S., C. E. BUGG, K. D. WILKINSON, R. D. VIERSTRA, P. HATFIELD *et al.*, 1987 Comparison of the three-dimensional structures of human, yeast, and oat ubiquitin. J. Biol. Chem. **262:** 6396–6399.

WANDELT, C., and G. FEIX, 1989 Sequence of a 21-kd zein gene from maize containing an in-frame stop codon. Nucleic Acids Res. **17:** 2354.

WATKINS, J. F., P. SUNG, L. PRAKASH and S. PRAKASH, 1993 The *Saccharomyces cerevisiae* DNA repair gene RAD23 encodes a nuclear protein containing a ubiquitin-like domain required for biological function. Mol. Cell. Biol. **13:** 7757–7765.

WIBORG, O., M. S. PEDERSEN, A. WIND, L. E. BERGLUND, K. A. MARKCKER *et al.*, 1985 The human ubiquitin multigene family: some genes contain multiple directly repeated ubiquitin coding sequences. EMBO J. **4:** 755–759.

Communicating editor: W. F. SHERIDAN