Molecular Characterization of Duplicate Cytosolic Phosphoglucose Isomerase Genes in *Clarkia* and Comparison to the Single Gene in *Arabidopsis*

B. R. Thomas,* V. S. Ford,* E. Pichersky[†] and L. D. Gottlieb*

*Department of Genetics, University of California, Davis California 95616 and [†]Department of Biology, University of Michigan, Ann Arbor Michigan 48109

Manuscript received March 16, 1993
Accepted for publication July 30, 1993

ABSTRACT

The nucleotide sequence of *PgiC1-a* which encodes a cytosolic isozyme of phosphoglucose isomerase (PGIC; EC 5.3.1.9) in *Clarkia lewisii*, a wildflower native to California, is described and compared to the previously published sequence of the duplicate *PgiC2-a* from the same genome. Both genes have the same structure of 23 exons and 22 introns located in identical positions, and they encode proteins of 569 amino acids. Exon and inferred protein sequences of the two genes are 96.4% and 97.2% identical, respectively. Intron sequences are 88.2% identical. The high nucleotide similarity of the two genes is consistent with previous genetic and biosystematic findings that suggest the duplication arose within *Clarkia*. A partial sequence of *PgiC2-b* was also obtained. It is 99.5% identical to *PgiC2-a* in exons and 99.7% in introns. The nucleotide sequence of the single *PgiC* from *Arabidopsis thaliana* was also determined for comparison to the *Clarkia* genes. The *A. thaliana* *PgiC* has 21 introns located at positions identical to those in *Clarkia PgiC1* and *PgiC2*, but lacks the intron that divides *Clarkia* exons 21 and 22. The *A. thaliana* PGIC protein is shorter, with 560 amino acids, and differs by about 17% from the *Clarkia* PGICs. The *PgiC* in *A. thaliana* was mapped to a site 20 cM from restriction fragment length polymorphism marker 331 on chromosome 5.

PLANT cytosolic phosphoglucose isomerase (PGIC; EC 5.3.1.9) catalyzes the reversible isomerization of fructose-6-phosphate and glucose-6-phosphate, an essential reaction that precedes sucrose biosynthesis in the cytosol. In general, plant cells contain a single cytosolic PGIC (GOTTLIEB 1982) but some species of *Clarkia*, a genus of wildflowers nearly all native to California, have two, the result of a gene duplication that appears to have originated within the genus (GOTTLIEB 1977; GOTTLIEB and WEEDEN 1979). The duplicate genes, designated *PgiC1* and *PgiC2*, assort independently and are found in 12 of the 24 diploid *Clarkia* species (GOTTLIEB 1977; GOTTLIEB and WEEDEN 1979; WEEDEN and GOTTLIEB 1979).

The active form of PGIC is a dimer of like subunits with molecular mass of approximately 60–62 kD (HERBERT and SCHNARRENBARGER 1982; WEEDEN and GOTTLIEB 1982; GOTTLIEB and HIGGINS 1984; HIGGINS and GOTTLIEB 1984). The subunits encoded by the duplicate genes associate *in vivo* in all possible combinations of homo- and heterodimers. The duplicated isozymes have similar specific activities and catalytic properties, although they differ in isoelectric point, some immunological determinants, and the efficiency of subunit reassociation following *in vitro* dissociation (HIGGINS and GOTTLIEB 1984).

The first complete PGI protein sequences deter-

mined were inferred from cDNA sequences in mouse (FAIK *et al.* 1988) and pig (CHAPUT *et al.* 1988). We recently reported the first complete genomic PGI sequence from a multicellular organism, *PgiC2-a* from *Clarkia lewisii* (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992). *PgiC2* has a complex structure of 22 introns and 23 exons encoding a protein of 569 amino acids. The *C. lewisii* PGIC2 protein is 44–46% identical to the inferred amino acid sequences of pig, *Escherichia coli* and *Saccharomyces cerevisiae*.

In this study we report the cloning and sequencing of *PgiC1-a* from *C. lewisii*, as well as a new allele *PgiC2-b*, and discuss their relationship to *PgiC2-a*. For comparison, we also cloned and sequenced *PgiC* from *Arabidopsis thaliana*. This plant was chosen because of the broad interest in its molecular genetics, and suggestions (CHANG and MEYEROWITZ 1986) that its small genome was correlated with reductions in intron number and size.

Plants also have an independent nuclear gene, *PgiP*, encoding a plastid PGI isozyme that catalyzes a reaction preceding starch biosynthesis (SCHNARRENBARGER and OESER 1974; SIMCOX and DENNIS 1978; WEEDEN and GOTTLIEB 1980). A gene sequence for *PgiP* was previously published (TAIT *et al.* 1988), but the origin of the clone is in question and it is now being restudied.

MATERIALS AND METHODS

Plant materials: *Clarkia lewisii* (formerly *C. botata*) is an outcrossing diploid annual plant native to oak woodlands and adjacent habitats in the South Coast Ranges of central California (LEWIS and LEWIS 1955). The plants in this study derived from seeds collected in San Benito County, 7 miles south of Route 156 on Road G1 to Fremont Peak State Park (LDG 795). The allozymes encoded by *PgiC1* (previously known as *Pgi2*) have more anodal electrophoretic mobilities than those encoded by *PgiC2* (previously *Pgi3*).

Cloning strategy for *C. lewisii* *PgiC1*: *PgiC1-a* was obtained from the same genomic library that provided *PgiC2-a* and *PgiC2-b* (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992). The library had been prepared from the DNA of a single plant determined by electrophoresis to be heterozygous at both *PgiC1* and *PgiC2*. During the original study, several fragments were observed on genomic Southern blots that did not correspond either to *PgiC2-a* or *PgiC2-b* and presumably represented the alleles of *PgiC1*. The library was rescreened with a new probe, an 876-nucleotide (nt) *NdeI/EcoRV* fragment of *C. lewisii* *PgiC2-a* extending from the end of exon 18 to the 3'-non-translated region. Rescreening yielded 10 PGI-positive clones. Fragments from each were amplified by polymerase chain reaction (PCR) using 20-nt primers designed from sequences of *C. lewisii* *PgiC2-a* exons 11 and 16. Eight clones gave the 1464-nt fragment predicted from the known *PgiC2-a* sequence, but two clones gave a novel 1220-nt fragment, and this fragment was also amplified from genomic DNA by the same pair of primers. Therefore it was presumed to come from *PgiC1*. Southern hybridization was carried out with DNA from one of the clones, L53, and genomic DNAs from the four double homozygous descendants of the library plant, digested with the five-base cutter *DdeI* and electrophoresed on an agarose gel. The gel was blotted onto Nytran (Schleicher & Schuell), probed with a 2254-nt *StuI/EcoRI* fragment cut from intron 10 to intron 19 of L53 at 65°, and washed at high stringency (0.1 × SSC; 65°). Fragments from L53 were subcloned into pBlueScript SK (Stratagene) for sequencing.

A λ gt11 cDNA library prepared from a descendent (genotype *PgiC1-a PgiC1-a/PgiC2-a PgiC2-a*) of the plant used to make the genomic library was also screened with the *NdeI/EcoRV* probe. Forty-one PGI-positive clones were identified. A fragment of approximately 600 nt was amplified from each of them by PCR using 20-nt primers based on *PgiC2-a* sequences from exon 17 and the 3'-nontranslated region. The amplified fragments were classified into those cut or not cut by *EcoRV*. Since *PgiC2-a* has an *EcoRV* site in the 3'-nontranslated region, clones yielding a fragment not cut by the enzyme were presumed *PgiC1-a* clones. One of these clones, designated LC35, was sequenced.

Cloning strategy for *A. thaliana* *PgiC*: A λ ZapII cDNA library (2–3-kb insert size) prepared from 3-day old dark grown *A. thaliana* (ecotype Columbia) hypocotyls was kindly provided by JOE ECKER. The library was screened with two *C. lewisii* *PgiC2-a* genomic fragments, an *EcoRI* fragment (836 nt) extending from intron 19 into the 3'-nontranslated region, and the *NdeI/EcoRV* fragment described above. Inserts from two positive cDNA clones were recloned into the *EcoRI* site of pBlueScript SK. One, designated A12C, included an open reading frame (ORF) of 1683 nt from an ATG start codon plus a 3'-nontranslated region of 193 nt, for a total of 1876. A shorter clone, A12B, extended 1827 nt from a site 22 bases 3' of the start ATG into the 3'-nontranslated region, and had an identical sequence to A12C.

FIGURE 1.—Southern blot of L53 and genomic DNAs from four genotypes segregating in progeny of the *C. lewisii* genomic library plant: (A) *PgiC1-a PgiC1-a/PgiC2-a PgiC2-a*, (B) *PgiC1-b PgiC1-b/PgiC2-a PgiC2-a*, (C) *PgiC1-a PgiC1-a/PgiC2-b PgiC2-b*, and (D) *PgiC1-b PgiC1-b/PgiC2-b PgiC2-b*. The DNAs were digested with *DdeI* and probed with L53 *StuI/EcoRI* genomic fragment. The blot identifies L53 as a clone of *PgiC1-a*.

An EMBL 3 genomic library, prepared from 4–5-week-old, light grown leaves and stems, also of ecotype Columbia, by Clontech Laboratories, Inc., was kindly provided to us by A. CASHMORE. The library was probed with the entire A12B cDNA fragment. Two positive clones were obtained and shown, by restriction mapping and Southern blotting, to carry DNA from the same genetic locus. Overlapping fragments from clone A4 were subcloned into pBlueScript SK for sequencing.

DNA sequencing: The *Clarkia* cDNA and genomic DNAs were sequenced on both strands from sets of overlapping deletions prepared from plasmid subclones by the exonuclease III method (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992). *PgiC1-a* proved to be located at the 3' end of the L53 insert, terminating at a *Sau3A* site at position 5773 in the middle of exon 23. Consequently, bases 5774–5986 of the reported *PgiC1-a* sequence were obtained from the LC35 cDNA, including 103 nt at the 3' end determined from one strand only. One strand of the *A. thaliana* cDNA and both strands of the genomic clone A4, including the entire coding region, were sequenced from single-stranded DNA templates obtained by M13 phage-assisted rescue from the pBlueScript plasmids. The reported sequence from A4 also includes an additional 96 nt at the 5' end and 73 nt at the 3' end based on data from one strand only. Sequencing was by the dideoxynucleotide method using [32 P]dATP with the Sequenase 2.0 kit from U. S. Biochemical Corp. In addition to vector primers, oligonucleotides (17–21 nt) that hybridize to *PgiC* sequences were synthesized for use as primers. Intron-exon borders for *A. thaliana* *PgiC* and for exons 16 to 23 of *C. lewisii* *PgiC1-a* were determined by comparison to cDNA, the other intron borders of *PgiC1-a* by comparison to *PgiC2-a*. The *C. lewisii* *PgiC1-a* and *PgiC2-b* and *A. thaliana* *PgiC* sequences are reported as accession numbers X71084, X71085 and X69195, respectively, in the EMBL Nucleotide Sequence Database.

Sequence comparisons: Numbers of substitutions per synonymous site (K_s) and per nonsynonymous site (K_A), corrected for multiple substitutions, were estimated by the method of LI, WU and LUO (1985) as improved by LI (1993), using computer program LI93 provided by W.-H. LI. Nucleotide substitutions for noncoding regions were estimated by the two-parameter method of KIMURA (1980) using the computer program DNADIST by JOYCE MILLER, provided by M. NEI.

RESULTS

Identification of *C. lewisii* *PgiC1-a*: Two lines of evidence demonstrated that genomic clone L53 carried *PgiC1-a*. First, corresponding exons from its coding region exactly matched the sequence of 822 nt (exon 16 to poly(A) tract) of cDNA clone LC35. Second, Southern hybridization using the *StuI/EcoRI* L53 genomic probe revealed a perfect size match between a *DdeI* 1.9-kb fragment in restriction digests of L53 and the genomes of two of the four double homozygous descendents of the library plant: *PgiC1-a PgiC1-a/PgiC2-a PgiC2-a* and *PgiC1-a PgiC1-a/PgiC2-b PgiC2-b* (Figure 1). Because only *PgiC1-a* was common to these two genotypes, analogous to the case for *PgiC2-a* in the previous study (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992), L53 was a clone of *PgiC1-a*.

The 1.4-kb *DdeI* fragment in the two other genomic lanes of the blot shown in Figure 1 presumably represents *PgiC1-b*, but it has not yet been cloned. The original film also showed two short (679 and 847 nt) *PgiC2 DdeI* fragments not visible in the figure. Because they are mostly intron sequences, they hybridized weakly with the *PgiC1-a* probe under the high stringency wash conditions.

Structure and sequence of *PgiC1-a*: *PgiC1-a* has 23 exons and 22 introns (Figure 2) and encodes a protein of 569 amino acids, the same as *PgiC2-a* (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992). The molecular mass is 62,689. The nucleotide and inferred amino acid sequences of *PgiC1-a* and *PgiC2-a* are 96.4% and 97.2% identical, respectively. The exons of both genes have identical lengths and no insertion/deletion differences. Exon lengths range from 43 (exon 7) to 156 (exon 5) with an average length of 74.3 nt. All intron borders show the consensus 5'-GT . . . AG-3' dinucleotides typically found in eukaryotes.

Nucleotide and amino acid differences between the duplicate genes are summarized in Table 1. Nucleotide differences are not distributed homogeneously over the exons (heterogeneity χ^2 32.74, d.f. = 21). In particular, exons 8 and 23 have 8.8% and 9.8% nt substitutions, respectively. Two of the 16 amino acid replacements, methionine for arginine in exon 2 and serine for arginine in exon 5 (Fig. 3), reduce the positive charge of the PGIC1-A protein and are consistent with its faster electrophoretic mobility toward the anode than PGIC2-A (see Fig. 3 in THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992).

A putative TATA box appears in *PgiC1-a* at position 111 and a probable CAATT sequence at position 83 (Fig. 2). The CAATT sequence occupies the same position as in *PgiC2-a*, but the TATA sequence of *PgiC1-a* is about 20 nt closer than in *PgiC2-a*.

A poly(A) addition signal AAATAA is at position

5955 in *PgiC1-a* (Fig. 2), a location homologous to that in *PgiC2-a* although closer to the stop codon because of insertion/deletion differences. The LC35 cDNA of *PgiC1-a* included a poly(A) tract at position 5987, 26 nt from the signal.

Codon usage of *PgiC1-a* is similar to that in *PgiC2-a*, with the same codons used commonly or rarely. *PgiC1-a* makes use of 60 codons, one less than *PgiC2-a* (ACG, a rarely used codon for threonine, does not appear in *PgiC1-a*).

Most introns of *PgiC1-a* include a stop codon in all three reading frames. Comparison of the duplicate genes (Table 2, Fig. 4) shows that, although only three introns have no insertion/deletion differences, the introns of the two genes are similar in length, with the exceptions of introns 11 and 12. In *PgiC2-a*, intron 11 contains 610 nt and includes a very large stem-loop structure with 102 nt pairs in the stem, of which 93 are complements, and a loop of 299 nt (Fig. 5). The stem-loop is bounded by a perfect 9-nt direct repeat (TATATATTT). In contrast, intron 11 in *PgiC1-a* is 129 nt long and lacks the entire stem-loop structure. However, it does have sequences that are reminiscent of the two direct repeats, an ATA-TATTT at position 3117 and a TATATGTTC at 3190 (Fig. 2), but the short sequence between them has no resemblance to the stem-loop of *PgiC2-a*. Intron 12 of *PgiC1-a* is more than two times longer than its counterpart.

As expected, the introns of the two genes exhibit substantially greater divergence than the exons (Table 1). The estimated number of substitutions per site is greater than the estimated number of substitutions per synonymous site in exons. Base substitutions are distributed homogeneously among the 22 introns.

Structure and sequence of *PgiC2-b*: The genomic clone L3 (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992) included about 55% of the coding region of *PgiC2-b*. A total of 3249 nt was sequenced, comprising the final 6 nt of intron 10 through the translation stop codon in exon 23 and 220 nt into the 3'-nontranslated region (Fig. 4). On the basis of this partial sequence, *PgiC2-b* is nearly identical to its allele *PgiC2-a* (Table 1).

In the comparable introns, the two alleles differ by 7-nt substitutions (0.3%), and five insertion/deletions. In the 3'-nontranslated region, they differ by 3-nt substitutions and one insertion/deletion (Table 3).

In the 13 exons that can be compared, there are 5-nt substitutions, all resulting in amino acid replacements (Tables 1 and 3). In three of these, *PgiC2-a* has the same amino acid as *PgiC1-a*, and at the other two, *PgiC2-b* is like *PgiC1-a*. Two of the replacements, alanine for glutamate and glutamine for glutamate, both in exon 19, reduce the negative charge on the PGIC2-B protein and are consistent with its relatively

FIGURE 2.—Genomic sequence of *C. lewisii* *PgiCl-a* including 23 exons and 22 introns. The derived amino acid sequence of the exons is shown in one-letter code. The positions of putative CAATT, TATA and AAATAA signals are boxed.

The relatively large number of base substitutions in exons, and the fact that all cause amino acid replacements, is surprising. If the number of substitutions per synonymous site is predicted to be similar to the

overall substitution rate in introns (0.003 substitution/site), and even if an overestimate of the number of synonymous sites is obtained by counting all two- and fourfold degenerate sites [terminology of Li (1993), a total of 341 sites], then only one synonymous substitution is expected (0.003×341). No amino acid replacements are expected because nonsynonymous

```

3101 ctatttttcttaatccatataatttaattaaataattttgcattgattgtttgagctgttccctcgctctctcctgttggtactaataattatattgttcac
3201 atctatattattgtggcagtcgttagctgtgttgagtggttgcctgttatcccttcagtatggctttgctgtgttgagagaagtagtaaacctatcatttc
    C S A V G V L P L S L Q Y G F A V V E K
3301 gcacatcatatcaattttcttctgttttccatgtcaatcacccaatcgctttcatgttcattttatttttttcttgggaagtccaaaagataaagaa
3401 tatgttttggaatatatacatgtataaactatataattctcttggaacaccccttctgcagtgaaatagtaaacaaattattaacaaacaccatagagg
3501 agtgtagacatttttggagtaaatgcctttagtaaatgcatgtagagagcaggagaattgcttgcctctctcgaaaagtatatgttcatatatactgc
3601 tgactattatggagttgactaatttttttaataagcggtcatagtaaatgttgtaattctttatgcagATTCTTGCAAGGGGCTCAGCAGATTGAT
    F L Q G A H S I D
3701 CAGCATTCTCTTCTGCGCCATTGAGAAAAATATACCTgtaagtccattgatacacacttgtagtaagtaagtcttggatggatgcacaagttgaa
    Q H F S S A P F E K N I P
3801 tgtttagtattttcagccataccctctttataagtttgatgacaaattgaattgattgaatctttaatacagGTACTTCTAGGTTTGTGAGTGTATG
    V L L G L L S V W
3901 GAATGTATCATTCTTGGATACCTGCAAGGgtgagttcacatgatagcttgaggcattatctcctttcttttagtagtctttaaaccactgaagcaatc
    N V S F L G Y P A R
4001 aattctttctagGCCATCTTACCATACTCCCAAGCACTGGAGAAATGGCTCCACATATACAACAGgtaagctagtgatcacatttctttattagtaata
    A I L P Y S Q A L E K L A P H I Q Q
4101 attagggtcttggaaactggaatgaagttggttttgcacctcttcttgatgtgactttacagGTCAGCATGGAGAGCAACGGCAAGGGTGTGTCATAGA
    V S M E S N G K G V S I D
4201 TGGTTTGCCTCTTCCCTTTGAGTCTGGTGAATGACTTTGGTGAACCAAGGTACAAATGGGCAGCAGATTTTACCAGCTGATTACCAGgtaacttcc
    G L P L P F E S G E I D F G E P G T N G Q H S F Y Q L I H Q
4301 ttactatctcaggtagacacgttttacttctgccatatttatccttgaaattattcttatttcattccctaataataggaataccttctattcgattatttccc
4401 gtatgtgcatgttggaactgtgtccattttattttaattttgaactaaaacttcaaaaaaaaaaaaaatttcttgcagcaaaaaaactactgcaagatc
4501 tggacctttcaccatgctcacaaaggcaattgcgagttgtattctgttttaggcaactcctttctcatcctatatgtgattaggactactactcaaattc
4601 acggttcaatgattccttcaaatgttcagctgcacaaatatttcttcaataacaatttagttgaggtgaattaagcaagtcactaagtaagttt
4701 ctttgtaacacacatgcatcttgaataacagGGCGGTGTCATTCCCTGCGAATTTATTGGGGTCGTGAAGAGCCAGCAACCTGTTTACCTCAAAGgtgaa
    G R V I P C D F I G V V K S Q Q P V Y L K G
4801 aaattgcagctagctatatttctaagagcttttatatttctacagctctgcatgagttattgtcacttttattgtagGTGAAGTGGTTAATAACCATGA
    E V V N N H D
4901 TGAAGTGAATGCCAATTTTTTTGCCCAGCCTGATGCTTTGGCATATGCCAAGgtaaatcctaattccatgtgcaatagctttcaaatccgctctttcatg
    E L M S N F F A Q P D A L A Y G K
5001 aaagttaagtgaatcacaaatgttgttgggttttagACCCAGAAAGAGCTAAAGAAGGAGAAATGTGTCAGAGCATCTTATCTCTACAAGgtaatactgtct
    T P E E L K K E N V S E H L I P H K
5101 ctgatttgggaattcaacatcttcttatttgttgaagctaataatgaatgtgtgatttatgcagACATTCACTGGTAACAGGCCATCTATTAGCAAT
    T F T G N R P S I S I
5201 CTGCTTCCAACATTGGATGCTTACCGTATGGACAGgtctgttgttctgtctcaatgattactaccgatagtttctttaaagaatatagttgtgttaa
    L L P T L D A Y R I G Q
5301 tctctatttgggtattactgtttatttctggactgtcagCTGTTGGCAATTTATGAACATAGAGTTCAGTGAAGGTTTTCGTATGGGGCATCAACTCA
    L L A I Y E H R V A V Q G F V W G I N S
5401 TTTGACCACTGGGGAGTGGAGTTAGGAAAGgtgaagtgccagccattcaaatgatttttagctttcatttgtattagcttataaatctgtgtgttatt
    F D Q W G V E L G K
5501 gttgcagTCTCTGGCAACACAAGTCCGAAAGCAACTCCATGGATCTCGAGTCAAGGGAGAACCTGTGGAGGAGGGTTTCACTTCAGCACTAAAACCTTG
    S L A T Q V R K Q L H G S R V K G E P V E E G F N F S T K T L
5601 TTAACAAGATACCTCCAGgtattctatttatctgtgtaaaccttaaacacccccctactagcttttcttctgctgaggtggcacatcagatgattctaa
    L T R Y L Q
5701 tcaatgtagttaattaagcttaattcttctctcatctcatgtacagGCAACTTCCGATGTCCTGCGATCTTCCACCTACTGCCTAACATATGAT
    A T S D V P A D P S T L L P N I #
5801 taggctgcttttgcatttttgaagacggatggttgttaccgtagagttgtcatcttcttttgccttaattgctcttcttactcttgactctctctgt
5901 ttgtatctgaaatcccatacccttgggttgtatgaattcttctaataataataataataatgaagacaagagattatt

```

FIGURE 2.—Continued.

substitutions are usually much less frequent than synonymous ones.

PgiC2-b lacks a 50-nt sequence in its 3'-nontranslated region that is present in both *PgiC2-a* and in *PgiC1-a*, suggesting that its absence in *PgiC2-b* results from a recent deletion. This difference between the two alleles provided a critical advantage because an *EcoRV* site within the sequence is the only known

restriction site that distinguishes them and was used in Southern hybridization tests (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992).

The other large length difference between the alleles occurs in intron 11 and involves the previously described stem-loop structure. *PgiC2-b* exhibits an additional 32 nt in the central portion of the loop (Figure 5). Sixteen nucleotides are a perfect duplica-

TABLE 1

Summary of some pairwise comparisons among *C. lewisii* *PgiC1-a*, *PgiC2-a*, *PgiC2-b*, and *A. thaliana* *PgiC*

	<i>PgiC1-a</i> <i>PgiC2-a</i>	<i>PgiC2-a</i> <i>PgiC2-b</i>	<i>PgiC1-a</i> <i>PgiC</i>	<i>PgiC2-a</i> <i>PgiC</i>
Exons:				
Comparable nt	1707	951	1677	1677
Differences	62 [3.6%]	5 [0.5%]	402 [24.0%]	409 [24.4%]
K_s	0.102 (0.017)	0 (0)	1.413 (0.161)	1.463 (0.165)
K_A	0.013 (0.003)	0.007 (0.003)	0.108 (0.010)	0.109 (0.010)
Comparable aa	569	317	559	559
Differences	16 [2.8%]	5 [1.6%]	95 [17.0%]	99 [17.7%]
Introns:				
Comparable nt	2998	2041		
Differences	355 [11.8%]	7 [0.3%]		
Subst./site	0.129 (0.007)	0.003 (0.001)		

Standard errors are in parentheses, percentages in brackets.

<i>C. lew</i>	<i>PgiC1-a</i>	MASP-ALISE TEAMKDLKAH LEGIKMTHLR ELMGDTERCQ SMMVEFDNIF LDYSRQQASP	60
<i>C. lew</i>	<i>PgiC2-a</i>	..T.-..... ..R....	
<i>C. lew</i>	<i>PgiC2-b</i>	...ST...CDG. V.D..K.... D..S.AN... ..M...GLLR.TV	
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	DTISKLYKLA DAAHLKQKID RMYNGDHINT TENRSVLHVA LRAPRNAIC SDGKNVVPDV	120
<i>C. lew</i>	<i>PgiC2-a</i>	...N..... E.....S	
<i>C. lew</i>	<i>PgiC2-b</i>	E.MD..LN.. K.SQ.TE..S ..F..E...S	
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	WNVLDKIKDF SDSVRNGSWI GATGKELKDV IAVGIGGSFL GPLFVHTALQ TDPEASKNAR	180
<i>C. lew</i>	<i>PgiC2-a</i>	ER.....V	
<i>C. lew</i>	<i>PgiC2-b</i>	E. ..KI.S...VP.... ..I.....	
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	GRELRFLANV DPIDVARNIS GLNPETTLVV VVSKTFTTAE TMLNARTLRE WISSALGPSA	240
<i>C. lew</i>	<i>PgiC2-a</i>	A.....	
<i>C. lew</i>	<i>PgiC2-b</i>	Q.....I ..V.....	
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	VAKHMAVST NLPLVEKFGI DPNNAFAFWD WVGGRYSVCS AVGVLPPLSLQ YGFAVVEKFL	300
<i>C. lew</i>	<i>PgiC2-a</i>		
<i>C. lew</i>	<i>PgiC2-b</i>		
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	QGAHSIDQHF SSAPFEKNIP VLLGLLSVMN VSFLGYPARA ILPYSQALEK LAPHIQQVSM	360
<i>C. lew</i>	<i>PgiC2-a</i>	...N.....	
<i>C. lew</i>	<i>PgiC2-b</i>	K..S..... Q.T.....	
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	ESNGKGVSID GLPLPFESGE IDFGPGTNG QHSFYQLIHQ GRVIPCDFIG VVKSQQPVYL	420
<i>C. lew</i>	<i>PgiC2-a</i>		
<i>C. lew</i>	<i>PgiC2-b</i>		
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	KGEVNNHDE LMSNFFAQPD ALAYGKTPEE LKKENVSEHL IPHKTFTGNR PSISILLPTL	480
<i>C. lew</i>	<i>PgiC2-a</i>		
<i>C. lew</i>	<i>PgiC2-b</i>	AQ	
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	DAYRIGQLLA IYEHRAVAVQG FVWGINSFDQ WGVELGKSLA TQVRKQLHGS RVKGEPVEEG	540
<i>C. lew</i>	<i>PgiC2-a</i>		
<i>C. lew</i>	<i>PgiC2-b</i>		
<i>A. tha</i>	<i>PgiC</i>		
<i>C. lew</i>	<i>PgiC1-a</i>	FNSTKTLT RYLQATSDVP ADPSTLLPNI	
<i>C. lew</i>	<i>PgiC2-a</i>	T....	
<i>C. lew</i>	<i>PgiC2-b</i>	T....	
<i>A. tha</i>	<i>PgiC</i>		

FIGURE 3.—Aligned amino acid sequences inferred from *C. lewisii* *PgiC1-a*, *PgiC2-a* and *PgiC2-b* and *A. thaliana* *PgiC*. A dash indicates a site at which an insertion or deletion is inferred. A dot indicates an amino acid identity to *PgiC1-a*. A triangle identifies the active site histidine at position 98 and the active site lysine at position 517.

TABLE 2
Intron lengths

Intron	<i>C. Lewisii</i>			<i>A. thaliana</i>
	<i>PgiC2-a</i>	<i>PgiC2-b</i>	<i>PgiC1-a</i>	<i>PgiC</i>
1	82		82	94
2	144		151	157
3	87		87	83
4	389		374	184
5	94		94	81
6	113		168	98
7	223		240	277
8	81		83	91
9	318		361	180
10	439		430	277
11	610	642	129	76
12	189	189	392	280
13	97	97	135	197
14	86	86	81	89
15	95	95	96	118
16	440	435	441	115
17	86	86	82	92
18	73	72	82	95
19	88	88	79	97
20	95	95	104	99
21	91	91	77	0
22	92	92	130	79
Total	4012		3898	2859

tion of a 16-nt sequence in *PgiC2-a* (positions 3318–3333 in Figure 6 of THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB (1992). A short nucleotide segment is found between the duplicated stretches in *PgiC2-b*.

***A. thaliana PgiC*:** The PGIC protein has 560 amino acids and a molecular mass of 61,804. Comparison of the A4 genomic and A12C cDNA sequences identified 22 exons and 21 introns (Figure 4). The cDNA sequence completely matches the sequence derived from corresponding exons. All of the intron borders show consensus 5'-GT . . . AG-3' dinucleotides.

Codon usage in the *A. thaliana PgiC* is similar to that reported in dicot genes (CAMPBELL and GOWRI 1990). Only CCG (arginine) does not occur.

Comparison of *PgiC* with the exon sequences of *C. lewisii PgiC1* and *PgiC2* shows only three insertion/deletion differences: (1) a single codon in position 5 from the start codon that is absent from the *Clarkia* genes; (2) a single codon in position 538 in the *Clarkia* genes that is absent from the *A. thaliana* sequence; and (3) nine codons at the 3' end of the *Clarkia* genes that are absent in the *A. thaliana PgiC* (see inferred amino acid translation in Figure 3). Apart from these differences, all of the *PgiC1* and *PgiC2* exons of *C. lewisii* and *PgiC* exons of *A. thaliana* have the same lengths, except that exons 21 and 22 of *C. lewisii* correspond to exon 21 of *A. thaliana* (Figure 4). Thus, intron 21 of *C. lewisii PgiC* is absent from *A. thaliana PgiC*. Overall, the *A. thaliana PgiC* introns contain

2859 nt, 28.7% fewer than does *PgiC2-a*. The strong resemblance to the *C. lewisii* genes, both in gene structure and in inferred amino acid sequence (below) confirm the identity of the *A. thaliana PgiC*.

The exon sequences of the *A. thaliana* and *C. lewisii* genes are compared in Table 1. They are identical at about 76% of their nucleotides and 83% of amino acids. The estimate of substitutions per synonymous site is greater than one, meaning that every synonymous site is expected to have experienced a nucleotide substitution during the divergence of the two species. The intron sequences of the two genes are so highly diverged that they are not comparable.

Of the 99 amino acid differences between *A. thaliana* PGIC and *C. lewisii* PGIC2-A, 45 are encoded by exons 1–5 and another 25 by exons 20–22. Thus, 71% of the amino acid differences are clustered at the protein ends. A greater incidence of amino acid differentiation at both ends also characterizes the PGI proteins from pig, *E. coli* and yeast (THOMAS, LAUDENCIA-CHINGCUANCO and GOTTLIEB 1992). The most highly conserved protein sequences are encoded by exon 11 and exons 14–17.

Southern hybridization and restriction fragment length polymorphism (RFLP) mapping, kindly performed by M. McGRATH, using the A12B cDNA as a probe against genomic DNAs isolated from the Columbia and Landsberg ecotypes, demonstrated that the *A. thaliana* genome contains only a single copy of *PgiC*. DNAs from both ecotypes were cut by seven different enzymes and the wash following hybridization was carried out at high stringency ($0.1 \times \text{SSC}$; 65°). A single *Bgl*II fragment in each genome hybridized to the probe (Figure 6). One, two or three fragments from digests with the other enzymes hybridized in both genomes (Figure 6). These simple hybridization patterns are consistent with the restriction sites predicted by the sequence from the genomic clone. On the basis of a *Hind*III polymorphism, *PgiC* was mapped to a position on chromosome 5, 20 cM from marker 331 in the vicinity of sAT2105 on the RFLP linkage map (CHANG *et al.* 1988).

DISCUSSION

The duplicate *PgiC1* and *PgiC2* genes of *Clarkia lewisii* have the same structure of 23 exons interrupted by 22 introns in identical positions. The exon sequences are 96.4% identical and the proteins are alike at 553 of 569 amino acids (97.2%). Few other duplicate genes in plants have been sequenced and, in general, they are much more diverged. For example, *Adh1* and *Adh2* in maize differ at 17.8% of their exon nucleotides and 12.6% of their amino acids (DENNIS *et al.* 1985), and the more recently duplicated *Adh2* and *Adh3* in barley differ at 11% of their exon nucleotides (TRICK *et al.* 1988). Both *Clarkia* genes differ

FIGURE 4.—Intron-exon diagram drawn to scale of *PgiC* genes from *C. lewisii* and *A. thaliana*. Exons are shown as stippled boxes and introns are numbered. Intron 21 of the *C. lewisii PgiC* genes is absent from the *A. thaliana PgiC*.

FIGURE 5.—Scale diagram of stem-loop structure in intron 11 in *C. lewisii PgiC2-a* (top) and *PgiC2-b* (bottom). The flanking perfect direct repeats of nine nt are shown with triangles. The three pairs of direct repeats within the stem are designated by letters, and the positions of additional direct repeats in the loop are shown as boxes.

equally from the single *PgiC* gene of *A. thaliana*, with 76% nucleotide identity. The *A. thaliana* protein is slightly smaller, with 560 amino acids, but is 83% identical to the *Clarkia* proteins.

The high similarity of *PgiC1* and *PgiC2* is consistent with the duplication having arisen within *Clarkia*. This suggestion was originally proposed (GOTTLIEB 1977; GOTTLIEB and WEEDEN 1979) when an electrophoretic survey of all diploid *Clarkia* species and related genera in the family Onagraceae revealed duplicated PGIC isozymes in 12 species that had been assigned to taxonomically advanced sections, whereas 12 species in other sections, including the putative *Clarkia*

ancestral stock (LEWIS and LEWIS 1955), as well as all related genera had a single PGIC.

The great similarity of the two alleles of *PgiC2* is of interest because it appears probable that none of the base substitutions occurring during their divergence have been obscured by "multiple hits" at the same site. For example, the number of base differences observed in introns (seven) equals the number of substitutions predicted from the Kimura estimate ($2041 \times 0.003 = 6.123$, Table 3). Similarly, the number of differences in exons (five) equals the predicted number of substitutions ($L_0K_0 + L_1K_1 + L_1K_1 = 5.002$, from the LI93 program). Overall, the sequenced nontranslated re-

TABLE 3

Summary of nucleotide and amino acid substitutions and insertion/deletion differences between *C. lewisii* *PgiC2-a* and *PgiC2-b*

Exon 12	1-base substitution (Ala in <i>PgiC2-a</i> , Ser in <i>PgiC2-b</i>)
Exon 13	1-base substitution (Asn in <i>PgiC2-a</i> , Ser in <i>PgiC2-b</i>)
Exon 19	2-base substitutions (Glu in <i>PgiC2-a</i> , Ala in <i>PgiC2-b</i>) (Glu in <i>PgiC2-a</i> , Gln in <i>PgiC2-b</i>)
Exon 20	1-base substitution (Cys in <i>PgiC2-a</i> , Ser in <i>PgiC2-b</i>)
Intron 11	1 addition ^a in <i>PgiC2-b</i> (32 nt)
Intron 12	1-base substitution
Intron 16	2-base substitutions 2 additions ^a in <i>PgiC2-a</i> (1 nt, 4 nt)
Intron 18	2-base substitutions 1 addition ^a in <i>PgiC2-a</i> (2 nt), 1 in <i>PgiC2-b</i> (1 nt)
Intron 19	1-base substitution
Intron 22	1-base substitution
3' non-trans	3-base substitutions 1 insertion in <i>PgiC2-a</i> (50 nt)

^a An addition is a sequence which is present in one allele but not the other, regardless of whether the difference resulted from an insertion or deletion.

FIGURE 6.—Southern blot of *A. thaliana* ecotypes Columbia (C) and Landsberg (L) genomic DNAs digested with different restriction enzymes and probed with *A. thaliana* cDNA A12B3. Lanes 1, 2: *Bgl*II; lanes 3, 4: *Dra*I; lanes 5, 6: *Eco*RI; lanes 7, 8: *Eco*RV; lanes 9, 10: *Hind*III; lanes 11, 12: *Sac*I; lanes 13, 14: *Xba*I.

gions differ by six observed insertion/deletions and 10 observed base substitutions, that is, insertion/deletion events appear to be about half as frequent as base substitutions. No comparable data for segregating alleles from a natural plant population are known to us.

That the proportion of base differences between the alleles is slightly higher in exons than in introns, and that all five exon differences result in amino acid replacements is surprising. If nucleotide substitutions in exons were fewer than expected, codon bias might provide an explanation. However, all 61 codons are used in *PgiC2-a* and, in the half gene region for which we have data for both alleles, the same 55 codons occur in each (data not shown).

The *A. thaliana* *PgiC* has 21 introns located at identical positions to those in Clarkia *PgiC1* and *PgiC2*, but lacks a single intron that separates Clarkia exons 21 and 22. Until *PgiC* sequences are obtained from other genera, we do not know whether this difference results from the gain of an intron in Clarkia or a loss in *Arabidopsis*. Clarkia intron 21 is not exceptional in length (Table 2) although it has been the site of four or five insertion/deletions distinguishing *PgiC1* and *PgiC2*. The flanking exons, exons 21 (90 nt) and 22 (111 nt) of Clarkia are longer than average (74.2 nt) but the corresponding single Arabidopsis exon (198 nt) is by far the largest in the gene. Only two other exons exceed 100 nt in length, exon 5 (156 nt) and exon 16 (129 nt). Thus it is the Arabidopsis *PgiC* that appears to have an exceptional feature.

The close similarity of intron number in the Arabidopsis and Clarkia *PgiC* indicates that the early suggestion (CHANG and MEYEROWITZ 1986) that *A. thaliana* genes might exhibit fewer introns than characteristic of homologous genes in other plants is not general. The few-intron hypothesis was based primarily on the finding that *Adh* has six introns in *A. thaliana* but nine in maize. More recent genomic sequence analysis of other genes encoding enzymes of primary metabolism in *A. thaliana* and other plants reveals a near equality of intron number. For example, the cytosolic enolase gene has 12 introns in tomato and 11 in *A. thaliana* (VAN DER STRAETEN *et al.* 1991), the cytosolic aldolase gene has a single intron in maize (DENNIS *et al.* 1988) and three in *A. thaliana* (CHOPRA, DOLFERUS and JACOBS 1990), and the sucrose synthase gene in maize (WERR *et al.* 1985) and in *A. thaliana* (CHOPRA *et al.* 1992) both have 14 introns in their coding regions. Similarity in intron number has also been found in the chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenase genes of *A. thaliana* (SHIH, HEINRICH and GOODMAN 1991) and other plants (LIAUD, ZHANG and CERFF 1990; MARTINEZ, MARTIN and CERFF 1989; SHIH, HEINRICH and GOODMAN 1988).

In each of these comparisons, intron position is generally highly conserved. However, intron length in the *A. thaliana* genes is usually significantly shorter as initially pointed out (CHANG and MEYEROWITZ 1986). For *PgiC*, in both *A. thaliana* and *C. lewisii*, the majority of introns are less than 100 nt in length, but all *A. thaliana* introns are less than 300 nt, while six *C. lewisii* introns may be larger than 300 nt, ranging up to 642 (Table 2).

Clarkia is an excellent genus for studies of gene evolution because the phylogenetic relationships among many of its species, and the grouping of species into taxonomic sections, are well established as a result of biosystematic research (LEWIS and LEWIS 1955; LEWIS 1973), isozyme analysis (GOTTLIEB 1974; ODR-

ZYKOSKI and GOTTLIEB 1984; SOLTIS, SOLTIS and GOTTLIEB 1987), and restriction enzyme analysis of chloroplast DNA (SYTSMA and GOTTLIEB 1986a,b; SYTSMA, SMITH and GOTTLIEB 1990). The discovery that Clarkias in some sections had two PGICs led to the suggestion that those sections form a phylogenetic group in whose common ancestor a single gene duplication event occurred. However, the uniqueness of the duplication event remains to be confirmed. Also, if the duplication occurred early in the history of the genus, some species that now display a single PGIC may have two *PgiC* genes with one lost or silenced. This possibility is being studied by Southern hybridization using *PgiC* cDNA probes on digests of genomic DNA from Clarkias with a single PGIC isozyme.

The mechanism of origin of many gene duplications in plants is uncertain. Most pairs of duplicate genes encoding enzymes in plants assort independently (GOTTLIEB 1982; PICHESKY and GOTTLIEB 1983; BERNATSKY and TANKSLEY 1986; TANKSLEY *et al.* 1988; TANKSLEY and PICHESKY 1988), unlike the case in animals, and this is also true for *PgiC1* and *PgiC2*. Thus, their origin probably does not involve unequal crossing over since this normally results in tandem duplicates. Reciprocal translocation may set up conditions that lead to duplications. If plants that differ for overlapping reciprocal translocations hybridize, the progeny will include individuals with duplicate chromosomal segments (from the overlap region) as demonstrated by BURNHAM (1962) in progenies from crosses between maize translocation stocks. In self-compatible plants like Clarkias, duplicate chromosomal segments that might occur by such a process could become homozygous by self-pollination. On this model, the ancestral *PgiC* locus was translocated to a nonhomologous chromosome giving rise to two lineages. The region marked by *PgiC* was included within an overlap between reciprocal translocations. Then, a cross between the lineages resulted in the duplication.

Reciprocal translocations resulting in gross chromosomal repatterning even between closely related species are frequent in Clarkia (LEWIS 1973). Recently we showed by an indirect method that *PgiC* is on different chromosomes in *C. amoena* and *C. rubicunda*, two species in the primitive *Rhodanthos* section of the genus that have a single PGIC isozyme (GOTTLIEB 1993). However, translocations appear not to be responsible for duplications in the well mapped tomato genome. Duplication may also occur via transposition and integration or by reintegration into the genome of short fragments by heteroduplex formation and ligation (PICHESKY 1990). If the origin of the Clarkia *PgiC* duplication involved hybridization of diverged lineages, then we expect that in the species with the duplication one gene will resemble the single *PgiC* of one lineage and the other gene will resemble the single

PgiC of the other lineage. If hybridization was not a factor, or if hybridization followed soon after a translocation event, then the two duplicate genes will resemble each other more closely than they will any *PgiC* genes of non-duplicate species. In addition to comparative studies of the *PgiC* genes, we are examining flanking DNA to determine the extent of the duplicated region. This examination of the origin of a gene duplication is made possible by the availability of a number of closely related species and a good phylogenetic framework, suggesting which species to examine to find genes that have greater or lesser degrees of divergence and that are closer to ancestral or derived lineages. These first results from *C. lewisii* suggest that *PgiC* will be appropriate for gene sequence comparisons in Clarkia, enabling us to determine the mode of origin of the duplication and where on the phylogenetic tree it occurred.

M. McGRATH carried out the Southern hybridization and RFLP mapping of *A. thaliana PgiC*, and we are grateful to him. This research was supported by National Science Foundation grants BSR 90-06458 and BSR 91-06831 to L.D.G.

LITERATURE CITED

- BERNATSKY, R., and S. D. TANKSLEY, 1986 Majority of random cDNA clones correspond to single loci in the tomato genome. *Mol. Gen. Genet.* **203**: 8–14.
- BURNHAM, C., 1962 *Discussions in Cytogenetics*. Burgess, Minneapolis.
- CAMPBELL, W. H., and G. GOWRI, 1990 Codon usage in higher plants, green algae and cyanobacteria. *Plant Physiol.* **92**: 1–11.
- CHANG, C., and E. M. MEYEROWITZ, 1986 Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. *Proc. Natl. Acad. Sci. USA* **83**: 1408–1412.
- 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.
- CHAPUT, M., V. CLAES, D. PORTELLE, I. CLUDTS, A. CRAVADOR, A. BURNIE, H. GRAS and A. TARTAR, 1988 The neurotrophic factor neuroleukin is 90% homologous with phosphohexose isomerase. *Nature* **332**: 454–455.
- CHOPRA, S., R. DOLFERUS and M. JACOBS, 1990 Cloning and sequencing of the *Arabidopsis* aldolase gene. *Plant Mol. Biol.* **15**: 517–520.
- CHOPRA, S., J. DEL-FAVORO, R. DOLFERUS and M. JACOBS, 1992 Sucrose synthase of *Arabidopsis thaliana*: genomic cloning and sequence characterization. *Plant Mol. Biol.* **18**: 131–134.
- DENNIS, E. S., M. M. SACHS, W. L. GERLACH, E. J. FINNEGAN and W. J. PEACOCK, 1985 Molecular analysis of the alcohol dehydrogenase 2 (*Adh2*) gene of maize. *Nucleic Acids Res.* **13**: 727–743.
- DENNIS, E. S., W. L. GERLACH, J. C. WALKER, M. LAVIN and W. J. PEACOCK, 1988 Anaerobically regulated aldolase gene of maize: a chimeric origin? *J. Mol. Biol.* **202**: 759–767.
- FAIK, P., J. I. WALKER, A. A. M. REDMILL and M. J. MORGAN, 1988 Mouse glucose-6-phosphate isomerase and neuroleukin have identical 3' sequences. *Nature* **332**: 455–456.
- GOTTLIEB, L. D., 1974 Genetic confirmation of the origin of *Clarkia lingulata*. *Evolution* **28**: 244–250.
- GOTTLIEB, L. D., 1977 Evidence for duplication and divergence

- of the structural gene for phosphoglucose isomerase in diploid species of *Clarkia*. *Genetics* **86**: 289–308.
- GOTTLIEB, L. D., 1982 Conservation and duplication of isozymes in plants. *Science* **216**: 373–380.
- GOTTLIEB, L. D., 1993 A simple method to test genetic allelism in nearly sterile interspecific plant hybrids. *Syst. Bot.* **18**: 145–149.
- GOTTLIEB, L. D., and R. C. HIGGINS, 1984 Evidence from subunit molecular weight suggests hybridization was the source of the phosphoglucose isomerase gene duplication in *Clarkia*. *Theor. Appl. Genet.* **68**: 369–373.
- GOTTLIEB, L. D., and N. F. WEEDEN, 1979 Gene duplication and phylogeny in *Clarkia*. *Evolution* **33**: 1024–1039.
- HERBERT, M., and C. SCHNARRENBERGER, 1982 Purification and subunit structure of glucosephosphate isomerase 2 from spinach leaves and immunochemical comparison with other isomerases. *Arch. Biochem. Biophys.* **217**: 452–459.
- HIGGINS, R. C., and L. D. GOTTLIEB, 1984 Subunit hybridization and immunological studies of duplicated phosphoglucose isomerase isozymes. *Biochem. Genet.* **22**: 957–979.
- KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111–120.
- LEWIS, H., 1973 The origin of diploid neospecies in *Clarkia*. *Am. Nat.* **107**: 161–170.
- LEWIS, H., and M. E. LEWIS, 1955 The genus *Clarkia*. *Univ. Calif. Publ. Bot.* **20**: 241–392.
- 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.
- LIAUD, M. F., D. X. ZHANG and R. CERFF, 1990 Differential intron loss and endosymbiotic transfer of chloroplast glyceraldehyde-3-phosphate dehydrogenase genes to the nucleus. *Proc. Natl. Acad. Sci. USA* **87**: 8918–8922.
- MARTINEZ, P., W. MARTIN and R. CERFF, 1989 Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase from maize. *J. Mol. Biol.* **208**: 551–565.
- ORDZYKOSKI, I. J., and L. D. GOTTLIEB, 1984 Duplication of genes coding 6-phosphogluconate dehydrogenase in *Clarkia* (Onagraceae) and their phylogenetic implications. *Syst. Bot.* **9**: 479–489.
- PICHERSKY, E., 1990 Nomad DNA—a model for movement and duplication of DNA sequences in plant genomes. *Plant Mol. Biol.* **15**: 437–448.
- PICHERSKY, E. and L. D. GOTTLIEB, 1983 Evidence for duplication of the structural genes coding plastid and cytosolic isozymes of triose phosphate isomerase in diploid species of *Clarkia*. *Genetics* **105**: 421–436.
- SCHNARRENBERGER, C., and A. OESER, 1974 Two isozymes of glucose-6-phosphate isomerase from spinach leaves and their intracellular compartmentation. *Eur. J. Biochem.* **45**: 77–82.
- SHIH, M. C., P. HEINRICH and H. M. GOODMAN, 1988 Intron existence predated the divergence of eukaryotes and prokaryotes. *Science* **242**: 1164–1166.
- SHIH, M. C., P. HEINRICH and H. M. GOODMAN, 1991 Cloning and chromosomal mapping of nuclear genes encoding chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenase from *Arabidopsis thaliana*. *Gene* **104**: 133–138.
- SIMCOX, P. D., and D. T. DENNIS, 1978 Isoenzymes of the glycolytic and pentose phosphate pathways in proplastids from the developing endosperm of *Ricinus communis*. *Plant Physiol.* **61**: 871–877.
- SOLTIS, P. S., D. E. SOLTIS and L. D. GOTTLIEB, 1987 Phosphoglucose isomerase gene duplications in *Clarkia* (Onagraceae) and their phylogenetic implications. *Evolution* **41**: 667–671.
- SYTSMA, K. J., and L. D. GOTTLIEB, 1986a Chloroplast DNA evidence for the origin of the genus *Heterogaura* from a species of *Clarkia* (Onagraceae). *Proc. Natl. Acad. Sci. USA* **83**: 5554–5557.
- SYTSMA, K. J., and L. D. GOTTLIEB, 1986b Chloroplast DNA evolution and phylogenetic relationships in *Clarkia* sect. *Peripetasma* (Onagraceae). *Evolution* **40**: 1248–1261.
- SYTSMA, K. J., J. F. SMITH and L. D. GOTTLIEB, 1990 Phylogenetics in *Clarkia* (Onagraceae): restriction site mapping of chloroplast DNA. *Syst. Bot.* **15**: 280–295.
- TAIT, R. C., B. E. FROMAN, D. L. LAUDENCIA-CHINGCUANCO and L. D. GOTTLIEB, 1988 Plant phosphoglucose isomerase genes lack introns and are expressed in *Escherichia coli*. *Plant Mol. Biol.* **11**: 381–388.
- TANKSLEY, S. D., and E. PICHERSKY, 1988 Organization and evolution of sequences in the plant nuclear genome, pp. 55–83 in *Plant Evolutionary Biology*, edited by L. D. GOTTLIEB and S. K. JAIN. Chapman & Hall, London.
- TANKSLEY, S. D., R. BERNATSKY, N. L. LAPITAN and J. P. PRINCE, 1988 Conservation of gene repertoire but not gene order in pepper and tomato. *Proc. Natl. Acad. Sci. USA* **85**: 6419–6423.
- THOMAS, B. R., D. LAUDENCIA-CHINGCUANCO and L. D. GOTTLIEB, 1992 Molecular analysis of the plant gene encoding cytosolic phosphoglucose isomerase. *Plant Mol. Biol.* **19**: 745–757.
- TRICK, M., E. S. DENNIS, K. J. R. EDWARDS and W. J. PEACOCK, 1988 Molecular analysis of the alcohol dehydrogenase gene family in barley. *Plant Mol. Biol.* **11**: 147–160.
- VAN DER STRAETEN, D., R. A. RODRIGUES-POUSADA, H. M. GOODMAN and M. V. MONTAGU, 1991 Plant enolase: gene structure, expression and evolution. *Plant Cell* **3**: 719–735.
- WEEDEN, N. F., and L. D. GOTTLIEB, 1979 Distinguishing allozymes and isozymes of phosphoglucose isomerases by electrophoretic comparison of pollen and somatic tissues. *Biochem. Genet.* **17**: 287–296.
- WEEDEN, N. F., and L. D. GOTTLIEB, 1980 The genetics of chloroplast enzymes. *J. Hered.* **71**: 392–396.
- WEEDEN, N. F., and L. D. GOTTLIEB, 1982 Dissociation, reassociation and purification of plastid and cytosolic phosphoglucose isomerase isozymes. *Plant Physiol.* **69**: 717–723.
- WERR, W., W. B. FROMMER, C. MAAS and P. STARLINGER, 1985 Structure of the sucrose synthase gene on chromosome 9 of *Zea mays*. *EMBO J.* **4**: 1373–1380.