A Recently Silenced, Duplicate PgiC Locus in Clarkia

L. D. Gottlieb and V. S. Ford

Section of Evolution and Ecology, Division of Biological Sciences, University of California, Davis

Introduction

Duplicated genes have two possible ultimate fates: they may acquire different functions or, alternatively, one copy may be silenced and remain in the genome as a pseudogene, accumulating mutations until it is no longer recognizable. PgiC has been duplicated in the annual wildflower genus Clarkia (Onagraceae), mostly native to California. PgiC encodes the cytosolic isozyme of phosphoglucose isomerase (PGIC; EC 5.3.1.9), which catalyzes the reversible isomerization of fructose-6-phosphate and glucose-6-phosphate, an essential step in the conversion of triose phosphates to sucrose in the cytosol. Biochemical studies (Higgins and Gottlieb 1984) did not reveal any functional differences between the duplicate PGIC isozymes, but phylogenetic analysis (Gottlieb and Ford 1996) predicted that one of the duplicates had been silenced multiple times.

Electrophoretic surveys of PGIC isozymes in all of the diploid clarkias showed that 17 out of 18 species in five taxonomic sections express two isozymes, whereas the 14 species in three other sections express only one (Gottlieb and Weeden 1979; unpublished data). Analysis of 11 complete nucleotide sequences of PgiCs from species with either one or two PGIC isozymes as well as parts of two pseudogenes demonstrated that a single duplication occurred before the radiation of the extant sections of the genus (Gottlieb and Ford 1996). The duplicate genes, designated PgiC1 and PgiC2, assort independently (Gottlieb 1977; Gottlieb and Weeden 1979; Weeden and Gottlieb 1979). Both genes are expressed in species with two PGIC isozymes, but in species with a single PGIC only PgiC1 is expressed. This result implies that PgiC2 was silenced in the latter species, and the phylogenetic tree shows that PgiC2 was silenced independently in each of the three sections having only one PGIC and also in a single species in a fourth section (fig. 1; Gottlieb and Ford 1996).

The sequenced expressed genes have been described in detail (Thomas, Laudencia-Chingcuano, and Gottlieb 1992; Thomas et al. 1993; Ford, Thomas, and Gottlieb 1995; Gottlieb and Ford 1996). The eight PgiC1s and three PgiC2s, as well as PgiC from Oe. mexicana that was used as an outgroup, have 23 exons and 22 introns in exactly corresponding positions and encode proteins of 568–570 amino acids. The average pairwise nucleotide identity between the PgiC1s and the PgiC2s is 95.9% in exons and 88.0% in introns.

Here we describe the PgiC2 of C. mildrediae, a species with one PGIC isozyme, and compare it with other PGIC genes from Clarkia. We also consider certain constraints that may affect the evolution of duplicated genes, such as PgiC genes that encode subunits of multimeric proteins.

Materials and Methods

Plant Materials

Clarkia mildrediae (Heller) Lewis & Lewis is an outcrossing diploid annual plant distributed in the Yellow Pine Forest of the Sierra Nevada in Plumas and Butte Counties, California. Its taxonomic and reproductive relationships to other species in its section were described by Small (1971a, 1971b). The plants studied here were propagated from a collection by N. F. Weeden (165A) 5.5 miles south of Yellow Creek Bridge over the North Fork of the Feather River, Highway 70, Plumas County.

Cloning Strategy

A Charon 35 library was prepared from genomic DNA of several plants of C. mildrediae by protocols described in Thomas, Laudencia-Chingcuano, and Gottlieb (1992). The library was screened with a 836-nt EcoRI fragment of C. lewisii PgiC2-a, and one clone, M11, was obtained. Sequencing showed that M11 included a PgiC1.
To search for a $\psi$PgiC2, genomic DNA was prepared by a CTAB miniprep protocol (modified from Richards, Reichardt, and Rogers 1994) from seedling leaves of single plants related to those used to make the library. These plants had been made identical-by-descent for alleles encoding single PGIC1 allozymes by self-pollinating heterozygous individuals identified by electrophoresis. The use of identical-by-descent material for PCR experiments to detect fragments of $\psi$PgiC2 avoided coamplification of fragments of two alleles of PGIC1.

PCRs were performed as in Gottlieb and Ford (1996). A number of primer pairs amplified two fragments each from several DNAs. Fragments amplified from DNA of one plant were sequenced. In each case, the larger fragment was identical in sequence to the PGIC1 already cloned (M11) and the smaller fragment resembled a PgiC2. A 1,450-nt PgiC2-like sequence corresponding to the region between exon 13 and exon 21 was assembled from overlapping fragments but attempts to extend the sequence further by PCR were unsuccessful.

The genomic library was rescreened with a 1,760-nt HindIII fragment of PGIC1 of C. mildrediae extending from intron 12 to intron 19, and four positive clones were obtained. Candidate clones of $\psi$PgiC2 were identified by the length of PCR-amplified fragments. The PgiC from one clone, M4, was sequenced. The region between exon 13 and exon 21 proved to be identical in sequence to the PgiC2-like fragments previously obtained by PCR.

DNA Sequencing and Sequence Analysis

Restriction fragments of Charon 35 clones were subcloned into pBluescript SK (Stratagene) for sequencing. Sequencing was performed as in Gottlieb and Ford (1996). Both strands were sequenced. The EMBL number of the PgiC2 of C. mildrediae is X89388. The other Clarkia and Oenothera PgiC sequences (Gottlieb and Ford 1996) are numbered X89384–X89397.

Sequences were aligned by eye, one exon or intron at a time, with some use of the CLUSTAL4 program with user-defined dendrograms (Higgins and Sharp 1989) as described previously (Ford, Thomas, and Gottlieb 1995). Nucleotide substitutions (K) were estimated by the two-parameter method of Kimura (1980). Substitutions per synonymous site ($K_s$) and per nonsynonymous site ($K_a$) were estimated by the original method of Li (counting one third of a two-fold degenerate site as synonymous and two thirds as nonsynonymous; Li, Wu, and Luo 1985; Li 1993; program L193). Relative rate tests were performed using the method of Wu and Li (1985) and data from L193. Divergence rates were also compared by 1D and 2D tests (Tajima 1993).

Results

The 12.2-kb insert in clone M4 includes a 5,039-nt PgiC-like segment identified as part of a PgiC2 by com-
Comparison to other PgiCs from Clarkia (fig. 1; Gottlieb and Ford 1996). It extends from the middle of intron 5 into the 3′-nontranslated region beyond the poly(A) addition signals (fig. 2). A 533-nt segment immediately upstream was sequenced and found not to resemble known PgiCs. An additional 1.7 kb farther upstream on the clone was tested unsuccessfully with sequencing primers corresponding to portions of the 5′ regulatory region and exons 2, 3, and 4. Presumably, the 5′ end of the gene has been substantially modified, deleted, or displaced from the identified region by at least 2.2 kb.

Numerous other defects confirm that M4 carries a \( \psi \)PgiC2. Eighteen “exons” corresponding to exons 6–23 were identified by alignment with expressed PgiCs (fig. 3). Their combined length is 1,189 nt. Nine of the 18 exons have indels: eight deletions totaling 115 nt and three insertions totaling eight nt (figs. 2 and 3). The deletions range from 3 to 52 exon nt each and cause the loss of all or part of 43 codons, including almost all of exon 14. The indels in exons 7, 14, 15, 17, 18, and 23 cause frameshifts; the one in exon 18 results in a stop codon in the reading frame of exon 19. Four deletions (in exons 7, 14, 16, and 22) overlap adjacent introns, causing loss of the splice junction sites. (All introns in all expressed PgiCs examined in Clarkia, Oenothera, and Arabidopsis [Thomas et al. 1993] have consensus 5′-GT ... AG-3′ splice junctions.) Three other splice junction sites have been modified by the substitution of “TT,” “GG,” and “CT” for “GT” at the 5′ end of introns 8, 11, and 13, respectively. These defects preclude mRNA processing and translation of a PGiC.

Only two of the 11 indels, a two-codon duplication in exon 9 and a five-codon deletion in exon 23, do not obviously impair gene function. Even these probably occurred after the gene was silenced because functional PgiCs are very highly conserved with respect to exon length. For example, only two single-codon insertions have occurred among 12 expressed PgiCs of Clarkia and Oenothera. The distantly related PgIC of Arabidopsis thaliana has one single-codon insertion relative to Clarkia and nine codons deleted from the terminal exon (Thomas et al. 1993).

Nucleotide Substitutions in Exons

Relative to functional PgiC2 genes, the \( \psi \)PgiC2 has five synonymous and four replacement substitutions in “exons” (fig. 3). Relative-rate tests (Wu and Li 1985) show that exon nucleotides have not diverged faster in the pseudogene than in the expressed PgiC2s or in PgiCl of C. mildrediae (table 1). Similar tests using other reference points (e.g., any PgiCl can be the reference for comparisons of \( \psi \)PgiC2 with expressed PgiC2s) and also Tajima ID and 2D tests gave the same result (not shown). Because a pseudogene is expected to be free of selective constraints on substitutions (Li, Gotojori, and Nei 1981; Miyata and Yasunaga 1981), the absence of faster substitution rates in its “exons” suggests that \( \psi \)PgiC2 was recently silenced. The similarity in overall divergence of \( \psi \)PgiC2 of C. mildrediae and the expressed PgiC genes is shown by the horizontal branch lengths in figure 1. The available fragment of \( \psi \)PgiC2 of C. rostrata (Gottlieb and Ford 1996), also included in figure 1, is too short for significant comparisons.

In most cases, formal relative-rate tests comparing \( K_S \) and \( K_A \) for the pseudogene versus expressed genes are not appropriate because the number of differences is too small. As an alternative method of comparison, all observed substitutions from exons 6 to 23 were assigned to branches of the gene tree (fig. 1). At six sites where two equally parsimonious assignments were possible, the assignment indicating parallel substitutions rather than a substitution and a reversion was used. The branch to \( \psi \)PgiC2 of C. mildrediae has one of the smallest totals among the genes examined. Since there is no evidence of increased substitution rates, the low ratio of synonymous to replacement substitutions in the pseudogene is probably a consequence of the small number of events.

None of the four replacement substitutions introduced in-frame stop codons, but one, GGT(Gly)→TGT(Cys) in exon 11, is “radical” as assessed in Li, Wu, and Luo (1985). For comparison, there are no “radical” and only two “moderately radical” substitutions, Asn→Ile and Ser→Cys, in exons 6–23 of the 11 expressed genes sequenced to date.

The small number of nt substitutions in \( \psi \)PgiC2 cannot be accounted for by gene conversion from PgiCl in the same genome because the pseudogene occupies the appropriate position on the gene tree (fig. 1) and the two genes show no exceptional similarities. They differ at many sites and indels scattered throughout all exons and introns (not shown).
FIG. 3.—Alignment of exons 6-23, grouped by codons, for PgiC1 and \( \psi \)PgiC2 of C. mildrediae (M C1 and MδC2) and PgiC2 of C. lewisi (L C2), C. xantiana (X C2), and C. concinna (C C2). Only nt that differ from PgiC2 of C. lewisi are shown. Vertical lines show positions of introns. Triangles and solid squares mark insertions and deletions. Replacement substitutions in single sequences are lowercase.
Nucleotide Substitutions and Indels in Introns

Substitution in intron sequences of ψPgiC2 was compared with that in PgiC1 of C. mildrediae and the expressed PgiC2s by relative-rate tests (table 1). ψPgiC2 of C. mildrediae has not diverged faster than the expressed PgiC2 genes but has diverged faster than PgiC1 in the same genome. However, the latter comparison probably does not indicate accelerated divergence because tests comparing PgiC1s also show that PgiC1 in C. mildrediae diverged more slowly (Gottlieb and Ford 1996). Similar results were obtained using other reference points and also from Tajima 1D and 2D tests (not shown).

The 3,564-nt intron sequence has 21 insertions totaling 949 nt and 10 deletions (314 nt) in addition to the four deletions (58 intron nt) that overlap exons. Most of these indels are similar to those in expressed PgiCIs (Ford, Thomas, and Gottlieb 1995; Gottlieb and Ford 1996). All but two of the insertions are ≤6 nt, and most are duplications of adjacent sequences. An exceptional 857-nt insertion in intron 10 is described below. Three deletions in intron 16 total more than 200 nt but four expressed PgiCIs have similarly large deletions in that intron (Gottlieb and Ford 1996). The other seven deletions range from 2 to 26 nt.

Similar counts of indels in corresponding regions of other genes vary from 15 to 44 (not shown). Such counts are not reliable indicators of the number of indel events that have occurred because there are some regions of uncertain alignment, not all observed indels are necessarily independent, and there is no method for correcting for multiple events in the same position. The number of indels observed for ψPgiC2 (31) is large but not out of range.

The 857-nt insertion in intron 10 is the largest now known in any PgiC. It has a stem-loop structure like a transposable element, including perfect 9-nt direct repeats (TATTATTTT) outside a stem of imperfect 105-nt inverted repeats and a 646-nt loop. The inverted repeats and two regions of the loop are similar to corresponding regions of a previously described 513-nt stem-loop structure in intron 11 of PgiC2 of C. lewissii (Thomas, Laudencia-Chingcuanco, and Gottlieb 1992). Within the loop is a second pair of imperfect inverted repeat sequences (length ~143 nt) not found in the structure in C. lewissii. In consequence of this insertion, intron 10 of ψPgiC2 has 1,329 nt, 63% more than the largest intron in any expressed PgiC. Generally, intron 10 ranges from 400 to 500 nt; most PgiC introns in Clarkia are less than 200 nt.

Discussion

Clarkia mildrediae is found as a small number of populations in only two counties in California. Electrophoretic surveys of 330 individuals in 10 of its populations showed only one PGIC isozyme (Gottlieb and Weeden 1979; unpublished data), subsequently determined to be encoded by PgiC1 (Gottlieb and Ford 1996). Since the PgiC duplication occurred prior to the divergence of all extant species in Clarkia, PgiC2 should also be present in the C. mildrediae genome. The discovery of a ψPgiC2 fulfills this prediction and shows that the absence of a second PGIC isozyme in this species does not reflect failure to detect PGIC2 because of reduced activity or overlap with PGIC1 after electrophoresis. The ψPgiC2 appears to be one of the first examples of a specific duplicated locus regularly expressed in a group of related plant species that has been silenced in one of them.
diæ, i.e., a null mutation has been fixed in the species. Most of the more than 60 unprocessed nuclear pseudogenes from higher plants that have been published are members of multigene families. In these examples it is not generally known whether the pseudogene represents a silenced locus currently or formerly expressed in related species, a silenced allele at a locus expressed in other individuals or cultivars, or an imperfect gene duplication that was nonfunctional from its origin. At least a dozen processed pseudogenes have also been published, but these of course do not represent silenced loci.

Recent Silencing and Rapid Accumulation of Indels

The absence of faster nucleotide substitution rates in the “exons” of the ψPgiC2 indicates that it was silenced recently and, consequently, most of the nine substitutions probably occurred while the gene was still functional. Only the G→T causing the “radical” Gly→Cys replacement in exon 11 is unlikely to have occurred during that time. In contrast, all or most of the 11 indels affecting “exons” occurred coincident with or subsequent to the silencing. Thus, it appears that more indels than substitutions have accumulated in “exons” of the silenced gene. The comparable analysis cannot be made for “introns” because there is no good method to estimate the number of indels that occurred after the gene was silenced.

Compared to other plant pseudogenes, the accumulation of indels may have been unusually rapid in “exons” of ψPgiC2 of C. mildrediae. For example, a truncated tRNA5" differs from 11 functional genes by one deletion and three substitutions at otherwise invariant coding positions (Fuchs, Beier, and Beier 1992). A maize U3 snRNA pseudogene differs from a functional copy by five indels and 19 nt substitutions in the comparable coding region; at least eight of the substitutions likely occurred after silencing because the positions are invariant among maize, wheat, and Arabidopsis (Leader et al. 1994). The “exons” of maize GAPA ψ/ apparently acquired two indels and 18 substitutions, and ψ2 no indels and 56 substitutions, since gene inactivation (Quigley et al. 1989). Most published comparisons of a plant pseudogene and a functional gene do not provide any estimate of the number of mutations after silencing.

Several studies present alignments of two or more pseudogenes that share significant indels or in-frame stop codons, suggesting that they diverged after the silencing of their common ancestor and, thus, all differences between pseudogenes occurred after silencing (e.g., Heim et al. 1989; Anderson 1991; Longhurst et al. 1994). In each case substitutions greatly outnumber indels in both coding and noncoding regions. None of these comparisons take into account multiple events in the same sites, but corrected data would likely increase the degree to which substitutions outnumber indels. We do not know any example other than the ψPgiC2 of C. mildrediae in which a plant nuclear pseudogene has more indels than substitutions.

Studies of animal pseudogenes also report fewer indels than substitutions (e.g., Miyamoto, Slightom, and Goodman 1987; Graur, Shuali, and Li 1989; Brown, Aquadro, and Anderson 1990).

If ψPgiC2 of C. mildrediae is exceptional in its accumulation of indels, the large number of introns in PgiC may be implicated. For example, the repetitive and AT-rich sequences typical of PgiC introns may provide many targets. Also, a number of large insertions and many short tandem duplications in PgiC introns suggest transposon activity. The transposon-like insert in intron 10 of ψPgiC2, which has appeared repeatedly in PgiC2 but never in PgiC1 (Thomas, Laudencia-Chingcuano, and Gottlieb 1992; unpublished data) is an apparent example.

The recent silencing of PgiC2 of C. mildrediae was unexpected because the other five diploid species in its section also have only a single PGIC, and thus it appears either that species divergence was very rapid or that a second gene was silenced in the section. The species in section Myxocarpa are very similar morphologically but extremely diverged in chromosome structure and number (Small 1971b). Their divergence involved two aneuploid chromosome reductions, one from n = 7 (C. borealis, C. mildrediae, C. stellata) to n = 6 (C. mosquinii) and a second from n = 6 to n = 5 (C. australis, C. virgata; Small 1971a, 1971b). The section also includes C. rhomboidea, a geographically widespread allotetraploid hybrid of one of the n = 5 species and C. mildrediae (Mosquin 1964; Small 1971b) that could not have originated prior to the origin of its n = 5 parent. If PgiC2 was silenced before the divergence of these species, the considerable chromosomal changes must have been very rapid. Alternatively, if C. mildrediae diverged at the base of the section but its PgiC2 was silenced only recently, a second silencing may have occurred in the lineage with the aneuploid reductions. The latter possibility would bring the number of independent instances of PgiC silencing in Clarkia to at least five.

Duplicated Genes Encoding Dimeric Enzymes Like PGIC May Have a Distinctive Pattern of Molecular Evolution

Since the duplication of PgiC, Clarkia has diverged to form 32 diploid species, and two functional PgiCs have persisted in 17 of them. Yet each section with an expressed PgiC2 is closely related to a section in which PgiC2 has been silenced (fig. 1). Thus, PgiC2 appears to have remained functional but inessential in the ancestral lineages.

We propose that PgiC2 may have been preserved in these lineages by selection against mutants causing defective PGI1-PGI2 heterodimers. It is usually postulated that after gene duplication, one copy is free to diverge as long as the other copy continues to perform its original function (Nei and Roychoudhury 1973; Li 1980; Walsh 1995). Nonfunctional alleles are generally treated as neutral or nearly neutral in effect. But Hughes and Hughes (1993) noted that a mutation affecting the structure of a duplicated protein may be deleterious if the altered product interacts adversely with other proteins. For a gene like PgiC that encodes subunits that associate at random to form homo- and heterodimers,
mutant products from one duplicate locus may interact directly and adversely with the products of the other. For example, if both loci are initially transcribed and translated at the same rate as before the duplication, the total production of enzyme is expected to double. But if an allele, say $PgiC2-m$, has mutations that reduce catalytic activity, three quarters of the PGIC dimers produced by a $PgiC1/PgiC2-m$ double homozygote would carry at least one copy of the defective monomer. The result could be reduction of total enzyme activity below that conferred by the single ancestral locus. If a subsequent mutation, $PgiC2-m2$, reduced the number of PGIC2 monomers, overall activity could actually increase because there would be more PGIC1 homodimers. Experimental evidence of such deleterious subunit interactions has recently been described for Cu/Zn superoxide dismutase in Drosophila (Phillips et al. 1995).

Such considerations suggest that, at least initially, mutations at either locus that modify enzyme function would likely be deleterious, whereas mutants that reduce interaction between the duplicate products may be selectively neutral or perhaps favored. Gene silencing in such a case is apt to be initiated by mutations that, for example, reduce or prevent transcription or translation or completely inactivate the product rather than by mutations that modify activity. Similarly, acquisition of a new function may be initiated by a change in regulation, permitting the duplicated loci to be expressed at different times or in different tissues (e.g., ADH in Drosophila [Sullivan et al. 1994] and many others). However, this route to divergence is perhaps unlikely for a gene like $PgiC$ because its product is required in all living plant cells. Alternatively, a mutation preventing or reducing formation of intergenic dimers could initiate silencing or set the stage for independent divergence of duplicate loci.

Li (1980) noted that if one duplicate locus comes to be expressed at a lower level, subsequent gene loss at the unchanged locus would likely be harmful. An early mutation in $PgiC2$ that reduced transcription or subunit dimerization could have made $PgiC1$ essential and set the stage for the multiple silencings of $PgiC2$. Experiments in C. xanthiana (Higgins and Gottlieb 1984; Jones, Pichersky, and Gottlieb 1986; Kruckeberg et al. 1989) showed that there is about 50% less PGIC2 activity than PGIC1 activity, although both isozymes have the same specific activity, and that dissociated PGIC2 monomers do not reassociate completely. Densitometric analyses show that PGIC2 makes a smaller contribution than does PGIC1 to total activity in other Clarkia species as well (Higgins and Gottlieb 1984). $PgiC2$ of C. xanthiana and C. lewisi but not C. concinna have a deletion in the probable TATA box (Gottlieb and Ford 1996). The deletion may reduce transcription but the matter has not yet been studied.

Presumably, the longer duplicate loci continue to share the same function, the more likely it is that both will be selected to achieve an optimal total enzyme activity. PGIC levels are about the same in representative species with one or both isozymes (Gottlieb and Higgins 1984), suggesting that an optimal level has been selected. In C. xanthiana, the wild-type level of PGIC activity is necessary to maintain wild-type levels of relevant metabolites, avoid shifts in the product/substrate ratio away from equilibrium and limit changes in fluxes to sucrose or starch (Kruckeberg et al. 1989). The two loci in this species are known to be independently regulated because an EMS-induced loss of activity at either locus causes a corresponding reduction in total activity (Jones, Pichersky, and Gottlieb 1986). Thus, the combined output of both loci is now required to attain the wild-type level.

If, after the $PgiC$ duplication, the reduction of PGIC activity to an optimal level was slow relative to the divergence of species in Clarkia, silencing of $PgiC2$ may have been possible, or favored, in species with excess total PGIC activity. Alternatively, if an optimal level of PGIC was attained relatively rapidly, then each silencing of $PgiC2$ must have been associated with a compensatory increase in activity of PGIC1. It is probable that the relative activities of the duplicate products are continuously adjusted against each other over time until the silencing of one restores the ancestral single-locus condition.

Although the silencing of a duplicate gene has been a matter of theoretical interest for 60 years (reviewed in Walsh 1995), most of our knowledge of silenced genes at the molecular level is based on processed pseudogenes and a very few well-studied systems of linked duplicate families, such as the globins and Adh. The $PgiCs$ of Clarkia present a somewhat different picture from these, in regard to both the persistence of duplicate loci without functional divergence and the prevalence of indels over substitutions in the “exons” of the pseudogene in C. mildrediae. Experimental studies of a greater variety of genes are needed to develop a more general understanding of the processes of gene silencing.

Acknowledgments

This research was supported by National Science Foundation grant BSR91-00831 to L.D.G. We thank Eran Pichersky for technical advice and discussion.

LITERATURE CITED


BARBARA A. SCHAAI, reviewing editor

Accepted October 23, 1996