J Mol Evol (1989) 29:412–421

Journal of
Molecular Evolution

Strong Functional GC Pressure in a Light-Regulated Maize Gene Encoding Subunit GAPA of Chloroplast Glyceraldehyde-3-Phosphate Dehydrogenase: Implications for the Evolution of GAPA Pseudogenes

Françoise Quigley,[1] Henner Brinkmann,[1] William F. Martin,[2] and Rüdiger Cerff[1]

[1] Laboratoire de Biologie Moléculaire Végétale, CNRS UA 1178, Université Joseph Fourier, F-38041 Grenoble, France
[2] Max-Planck-Institut für Züchtungsforschung, D-5000 Köln 30, FRG

Summary. The light-regulated nuclear gene encoding subunit A of chloroplast glyceraldehyde-3-phosphate dehydrogenase (subunit GAPA, gene *Gpa1*) from maize is extremely G+C rich (67% in codons). The genomic surroundings of this gene have been characterized together with the sequences of two strongly conserved *Gpa* pseudogenes isolated from a genomic maize library by differential cDNA hybridization. The comparisons show that the high G+C content of the maize gene is maintained independently of the surrounding noncoding sequences, which are G+C poor (42%), and only as long as the gene encodes a functional protein. After nonfunctionalization, *Gpa* pseudogenes rapidly loose G+C mainly due to enhanced turnover of CpG and CpXpG methylation sites. These results suggest that the maize *Gpa1* gene is under strong functional GC pressure, due to constraints (CpG island) probably exerted at the transcriptional level. They also indicate that *Gpa* pseudogenes are methylated and that methylation was either the cause or the immediate consequence of their nonfunctionalization. It can be concluded further that the progenitor of pseudogenes 1 and 2 was a second functional *Gpa* gene (*Gpa'*), which, after duplication, accelerated in evolutionary rate due to relaxation of selective constraints. This is in agreement with the neutral theory of evolution. Comparison of *Gpa* intron sequences reveals a gradient of divergence: the more 3′ the position of an intron the more its sequence has diverged between the three *Gpa* genes. A speculative model is presented explaining these observations in terms of a homologous recombination of genes with their reverse-transcribed pre-mRNAs.

Key words: Cytosine methylation — Isochores — CpG islands — Codon bias — Evolutionary rates — Reverse transcription — Homologous recombination — Cytochrome c

Introduction

Previous studies on the codon usage of higher plants (Brinkmann et al. 1987; Niesbach-Klösgen et al. 1987; Salinas et al. 1988) revealed fundamental differences between nuclear genes of monocotyledonous (monocot) and dicotyledonous (dicot) angiosperms. While monocot genes show a broad variation of G+C values at the degenerate third base position of codons (between 40 and 100%), the corresponding values of dicot genes were found to be relatively low, centered around 50% G+C with a rather symmetrical distribution (see Salinas et al. 1988). In view of the rough correlation between codon bias and gene expressivity in vivo, we (Brinkmann et al. 1987) suggested that the high G+C preference of certain inducible monocot genes reflects a *functional* GC pressure due to constraints exerted at the expression level. According to this hypothesis, monocot plants relate third position G+C preference with the tendency of a gene to show strong expression in response to endogenous and exogenous stimuli (for details see Brinkmann et al. 1987).

Offprint requests to: R. Cerff

Alternatively, G+C content and codon bias in genes may be a function of the overall mosaic structure of the genome. Such a relationship has been suggested previously for birds and mammals, where genes are embedded in "isochores" (Bernardi et al. 1985; Aota and Ikemura 1986; Bernardi and Bernardi 1986), long DNA segments (>300 kilobases) with different G+C levels and fairly homogeneous base compositions. This means that the G+C content at the third base position of a particular mammal gene reflects the G+C level of the surrounding isochore, comprising introns, flanking sequences, and spacer DNA.

A similar isochore organization has recently been proposed for the genomes of higher plants by Salinas et al. (1988). These studies seem to suggest that G+C-rich genes of monocots are clustered in G+C-rich isochores, which are absent in dicot genomes. The nuclear gene encoding subunit GAPA of chloroplast glyceraldehyde-3-phosphate dehydrogenase (gene *Gapa1*) is an excellent candidate with which to test this hypothesis. The maize *Gapa1* gene is strongly light regulated (Cerff and Kloppstech 1982) and extremely G+C rich in its coding sequences and in the region surrounding the promoter (67 and 60% G+C, respectively; see Quigley et al. 1988).

In the present work we characterized the genomic surroundings of the functional maize *Gpa1* gene and the sequences of two strongly conserved *Gpa* pseudogenes ($\psi 1$ and $\psi 2$) isolated from a genomic maize library by differential hybridization with specific and nonspecific cDNA fragments. The results show that the maize *Gpa1* gene is surrounded by G+C-poor noncoding sequences and its high G+C content is maintained only as long as the gene encodes a functional protein. After nonfunctionalization, *Gpa* pseudogenes rapidly loose G+C, mainly due to enhanced turnover of CpG and CpXpG methylation sites.

Materials and Methods

Plant Material. The maize plants used for preparation of DNA (genomic library and genomic Southern blots) originated from a genetic stock of P.A. Peterson (Ames, IA) and were grown in Cologne under the accession number 906.

Construction of Genomic Clones. Genomic DNA for cloning and Southern blots was prepared as described (Schwarz-Sommer et al. 1984). For genomic cloning, maize DNA (180 μg) was digested partially with MboI (New England Biolabs) and size fractionated on 0.7% agarose. The 17- to 24-kb fraction was electroeluted and purified by chromatography on Whatman DE 52. Lambda EMBL4 vector arms were prepared by digestion with BamHI and SalI (Frischauf et al. 1983) and subsequent centrifugation through potassium acetate gradients (Maniatis et al. 1982). MboI partials (1 μg) were ligated for 12 h at 16°C to 1.5 μg of EMBL4 vector arms in a volume of 10 μl containing 0.005 U/μl T4 DNA ligase (Boehringer). The ligation mixture was heated at 55°C for 5 min and in vitro packaged according to Hohn (1979). *Escherichia coli* strain K803 (Federoff 1983) was employed as a host.

Identification and Sequencing of Gpa Pseudogenes. Recombinant clones specific for *Gpa* pseudogenes were identified by differential plaque hybridization with a nonspecific GAPA probe (quasi-full-length cDNA pZm57 encoding maize GAPA; see Brinkmann et al. 1987) and a specific 3′ probe previously used to identify the functional *Gpa1* gene (Quigley et al. 1988). Two genomic clones hybridizing to the nonspecific but not to the specific probe were purified and DNA from CsCl-purified phage was isolated and digested with EcoRI. Hybridizing EcoRI fragments were subcloned into pBR322 and Bluescript (Stratagene). Purification of recombinant plasmids was performed as described (Maniatis et al. 1982).

Suitable restriction subfragments of genomic EcoRI fragments were subcloned into phage M13mp10, mp11, mp18, and mp19 and sequenced by the dideoxy chain termination method following the protocol supplied by Amersham.

Southern Hybridizations. Maize DNA (10 μg) was digested to completion with 10 U of the respective restriction enzyme, electrophoresed on 0.8% agarose, and depurinated prior to capillary transfer and UV coupling to Hybond N (Amersham) nylon filters according to the manufacturer's specifications. Filters were hybridized for 24 h at 65°C in 35 ml of 3× SSPE, 0.1% SDS, 0.2% PVP, and 0.2% Ficoll containing 50 μg/ml denatured salmon sperm DNA and 50 ng of a 185-bp Tth3II fragment of cDNA pZm57 (Brinkmann et al. 1987) spanning codons 243–303 in exon IV of the maize *Gpa1* gene (see Fig. 3) and random-prime labeled to a specific activity of 5×10^8 cpm/μg. Filters were washed twice for 20 min in 2× SSPE, 0.1% SDS at 65°C, and were exposed for 72 h at −70°C.

GAPDH Nomenclature. The nomenclature of the plant GAPDH system has been defined as follows to account for the complexity of the system and to comply with the international nomenclature of maize genetics (see also Martinez et al. 1989): Gene products (mRNAs, cDNAs, proteins) encoding or corresponding to subunits A and B of chloroplast GAPDH and subunit C of cytosolic GAPDH are specified as GAPA, GAPB, and GAPC RNAs or proteins, respectively, and products from different members of the same gene family are numbered consecutively, e.g., GAPA1, GAPA2 ..., GAPB1, GAPB2 ..., GAPC1, GAPC2.... The corresponding genes (gene families) are designated: *Gpa1, Gpa2 ..., Gpb1, Gpb2 ..., Gpc1, Gpc2 ...*, respectively. Pseudogenes are termed ψ*Gpa1* ($\psi 1$), ψ*Gpa2* ($\psi 2$), etc.

Computer Analysis. Sequence data were processed on a Multics computer (CICG Grenoble) by using the program developed by Greaves and Ware (University of Bristol, England, unpublished) and on a CII-Honeywell-Bull DPS8 computer of the computer service center CITI2 at Paris by using the program BISANCE.

Results and Discussion

The Noncoding Sequences Surrounding the Functional Gpa1 Gene Are G+C Poor

We have characterized the functional *Gpa1* gene and short stretches of its upstream and downstream flanking sequences previously (Quigley et al. 1988). This sequence (altogether 4060 bases) represents the 5′ part of a 6.4-kb genomic EcoRI fragment. The missing 3′ part (2.3 kb) of this genomic fragment

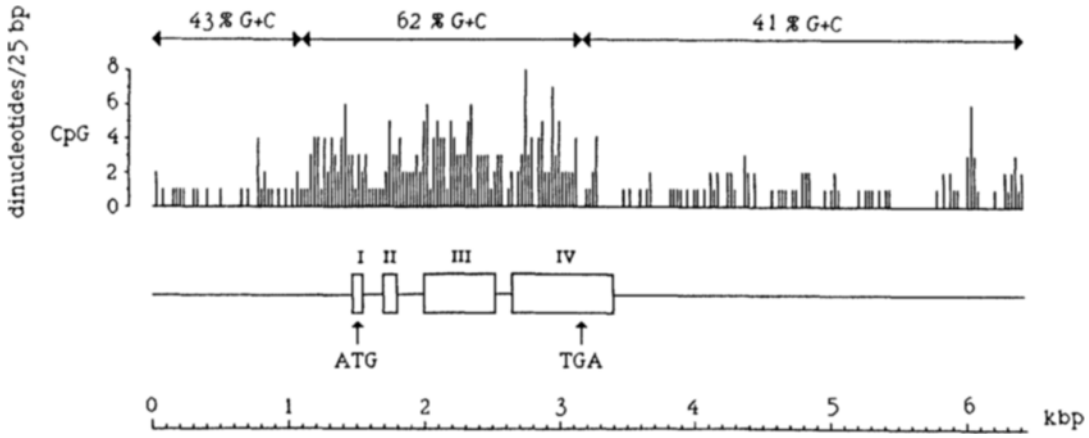

Fig. 1. CpG profile of the functional maize *Gpa1* gene and its flanking sequences. Each vertical line represents the number of CpG doublets per 25 bases. The gene structure is shown below the profile. Exons are indicated by boxes and roman numerals, introns and flanking sequences by continuous lines. Overall G+C values for the gene and its upstream and downstream flanking sequences are indicated above the profile. The precise values in percentages for G+C, CpG observed, and CpG expected are: codons, 67/12.4/11.1; promoter (comprising 390 bases upstream AUG), 60/11.3/8.9; introns, 53/7.4/6.9; 5′ and 3′ flanking sequences (4.3 kb), 42/2.9/4.3.

now has been sequenced. The CpG profile of the entire 6.4-kb region is shown in Fig. 1. It can be seen clearly that the *Gpa1* coding sequences (67% G+C, 12.4% CpG, 11.1% CpG expected) and the region surrounding the promoter (390 bases upstream AUG: 60% G+C, 11.3% CpG, 8.9% CpG expected) are rich in G+C, and CpGs are overrepresented slightly. Introns (53% G+C, 7.4% CpG, 6.9% CpG expected) and especially the 5′ and 3′ flanking sequences (altogether 4.3 kb: 42% G+C, 2.9% CpG, 4.3% CpG expected) are poor in G+C, and CpGs in flanking sequences are underrepresented somewhat. The average G+C content in the maize genome has been reported to be 48% (Salinas et al. 1988). Hence, our sequence data do not support the view that the functional *Gpa1* gene from maize is embedded into a G+C-rich isochore (Salinas et al. 1988). They rather suggest that the G+C enrichment is a local phenomenon maintained independently of the long-range G+C fluctuations in the monocot genome.

We suggested previously (Quigley et al. 1988) that the G+C enrichment in the functional *Gpa1* gene (Fig. 1) may be explained in terms of a combined effect of two separate functional constraints, exerted at both the translational and transcriptional level, respectively, leading to preferential use of G+C-rich codons in addition to CpG clustering in the 5′ part of the gene. Alternatively, it may be argued that the whole G+C-rich region in Fig. 1 is a typical CpG island (see Bird 1987; Antequera and Bird 1988), and since the gene is small, it fits inside. In this case, G+C and CpG clustering would be the cause and codon bias the consequence. Our recent comparison (Martinez et al. 1989) of several maize genes carrying CpG islands of variable sizes indicates that this latter interpretation is probably correct. This analysis showed that the G+C preference at the third base position of codons in a given maize gene is high inside and low outside the CpG island, suggesting that the *average* codon bias of maize genes is determined by the *relative* size and G+C content of the associated CpG island rather than by constraints at the translational level (for details see Martinez et al. 1989). This explanation implies that the lower G+C content in introns relative to exons observed for the *Gpa1* gene (see above) and other G+C-rich maize genes (e.g., the waxy locus gene: 3.7 kb with 13 introns; see Klösgen et al. 1986; Salinas et al. 1988; Martinez et al. 1989) is controlled by intron-specific constraints, possibly at the level of mRNA splicing.

Chloroplast GAPA Is Encoded by a Small Multigene Family in Maize

The number of genes and pseudogenes encoding glycolytic GAPDH in vertebrates varies between a single copy in chicken (Stone et al. 1985), 10–30 copies in human, hare, guinea pig, and hamster, and over 200 copies in mouse and rat (Hanauer and Mandel 1984; Piechaczyck et al. 1984). For cytosolic GAPDH of higher plants (GAPC) a single copy gene has been reported for barley (Chojecki 1986) and for maize three functional copies have been found that are regulated differentially under anaerobic conditions (Martinez et al. 1989; Russell and Sachs 1989).

To enumerate the genes and pseudogenes encoding chloroplast GAPA in maize, we probed Southern blots of maize DNA digested with EcoRI, BamHI, and HindIII (Fig. 2) with a 185-bp Tth3II-fragment of cDNA clone pZm57 (Brinkmann et al. 1987) spanning codons 243–303 in exon IV of the maize *Gpa1* gene (see Fig. 3). Figure 2 shows that there are two or three strongly hybridizing bands

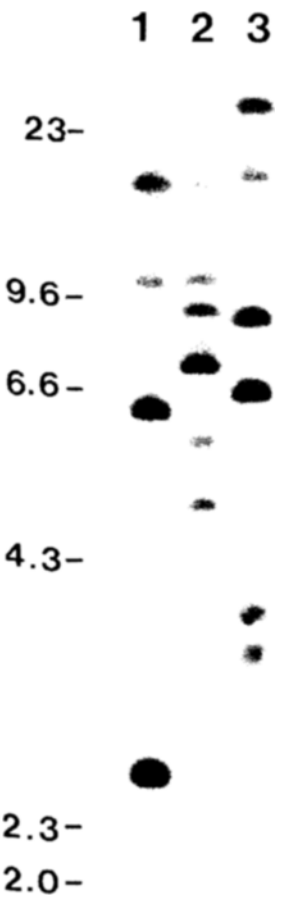

Fig. 2. Counting of *Gpa* genes in maize by Southern blotting. Aliquots of 15 μg of genomic DNA were digested with EcoRI (lane 1), BamHI (lane 2), and HindIII (lane 3). The fragments were separated by electrophoresis on a 0.8% agarose gel and blotted onto a nylon membrane. The filter was probed with a radioactive cDNA fragment and washed at high stringency as described in Materials and Methods. Autoradiography was for 4 days at −70°C with intensifying screen. Positions of molecular weight markers are indicated.

for all three digests. As many as 10 weakly or very weakly hybridizing fragments can be distinguished on the original autoradiogram for BamHI digests (see Fig. 2). We conclude from this that there may be at most two or three active *Gpa* genes in maize, whereas the majority of weakly hybridizing bands may represent more or less diverged pseudogenes (see below).

Isolation and Sequencing of Two Gpa Pseudogenes from Maize

Approximately 10^6 recombinant phage from a maize genomic library were screened with a nick-translated cDNA probe encoding chloroplast GAPDH from maize (clone pZm57; see Brinkmann et al. 1987). Two positive plaques that did not hybridize to the specific 3′ probe previously used to identify the active *Gpa1* gene (Quigley et al. 1988) were purified. DNA was isolated from each of these phage and digested with EcoRI restriction endonuclease. Electrophoresis of the restriction fragments revealed different patterns of the two clones suggesting that they represent different genomic regions. The fragments were transferred to nitrocellulose filters and hybridized with the same probe as used for screening the library. One clone contained a single hybridizing fragment of 5.9 kb carrying pseudogene *ψGpa1* (*ψ1*), while the other contained two hybridizing EcoRI fragments of 5 kb and 3.8 kb carrying the 5′ part (upstream codon −26, see Fig. 3) and 3′ part of pseudogene *ψGpa2* (*ψ2*) respectively. The three fragments were subcloned into pBR322 (*ψ1* and 3′ part of *ψ2*) and Bluescript (5′ part of *ψ2*) and submitted to sequence analysis (see Materials and Methods).

Pseudogenes 1 and 2 Do Not Encode Functional Proteins

In Fig. 3, pseudogenes 1 and 2 (*ψ1* and *ψ2*) are aligned with the functional *Gpa1* gene previously published (Quigley et al. 1988). Though highly homologous to the *Gpa1* gene, *ψ1* and *ψ2* clearly represent pseudogenes that do not encode functional proteins. In *ψ1*, a 9-bp deletion in exon III (codons 106–109) together with an 8-bp insertion in exon IV (after codon 319; see Fig. 3) destroy the proper function of the mature protein. Interestingly, the 8-bp insertion in *ψ1* is typical of a footprint of the transposable Ac/Ds system in maize (Saedler and Nevers 1985). In *ψ2*, a stop codon at triplet position 128 eliminates the entire catalytic domain (residues 148–333; see Biesecker et al. 1977) of any putative translation product emanating from *ψ2*.

Sequence Divergence in Introns of Maize Gpa Genes Follows a Gradient

The alignment of Fig. 3 reveals a high degree of sequence conservation in the promoter regions of *ψ1* and *ψ2*. The TATA box is conserved completely. The presumptive CATT box region, comprising the 13-mer CAGCCATTCGTCG in the functional *Gpa1* gene (Quigley et al. 1988) is conserved to a large extent in *ψ1* and *ψ2*, which clearly supports its functional significance. Interestingly, this 13-mer also is conserved in the promoter of the chicken GAPDH gene where it shows 11 matches (Stone et al. 1985). Pseudogenes 1 and 2 share an extensive deletion upstream of the TATA box, which was probably present in the common progenitor, a functional *Gpa* gene (*Gpa′*, see below). Hence, this region may not be essential for transcriptional regulation. The three genes show no significant sequence similarity upstream of the CATT box region (see Fig. 3).

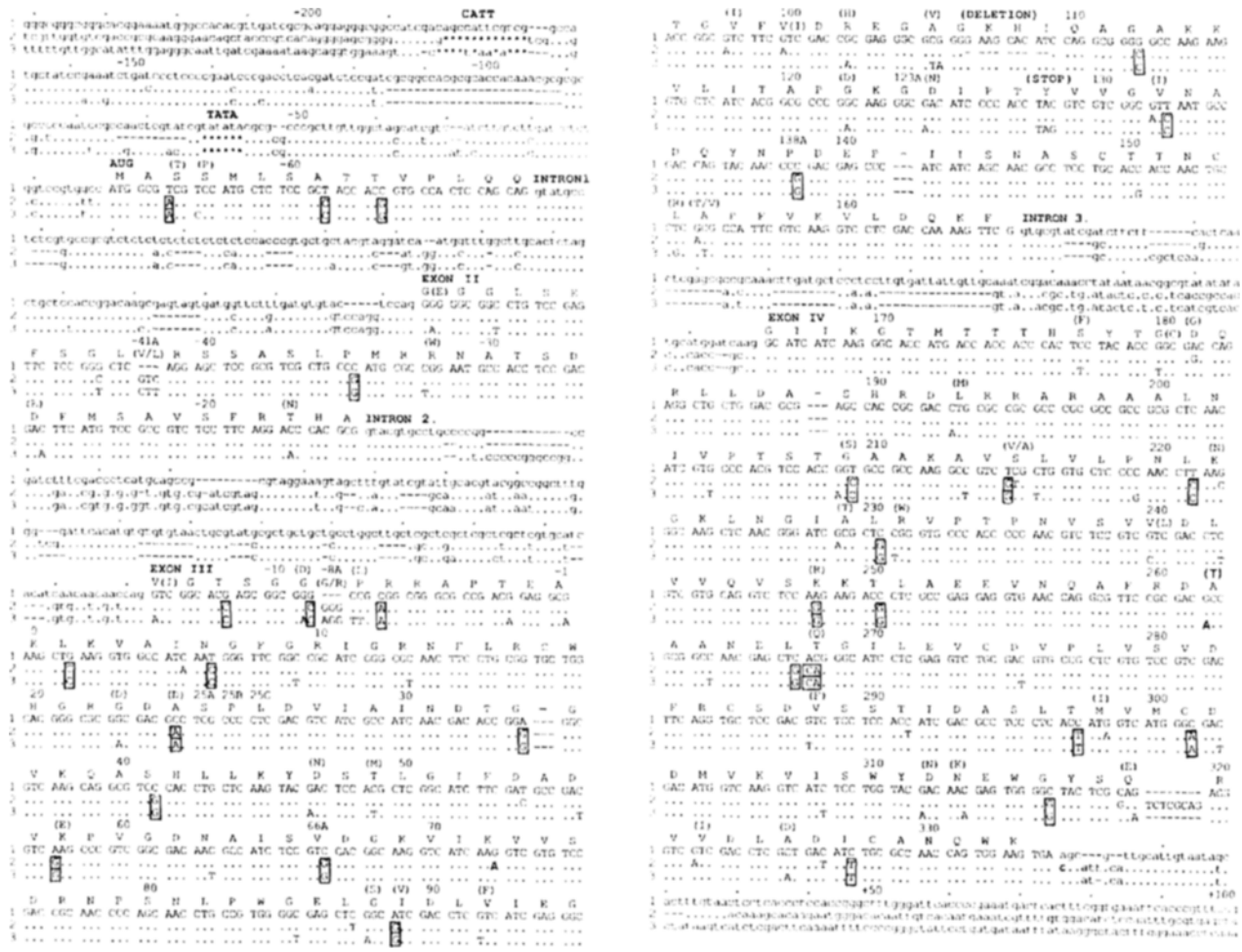

Fig. 3. Sequence alignments of pseudogenes 1 and 2 (lines 2 and 3) with the functional *Gpa1* gene (line 1). Insertions and deletions were introduced in the promoter regions and introns to maximize homology. Identical mutations in coding sequences of pseudogenes 1 and 2 (shared differences) are boxed, and boxes for replacement substitutions are subdivided. The deduced amino acid sequence of the functional *Gpa1* gene and corresponding substitutions in pseudogenes 1 and 2 (shown in parentheses) are aligned above the nucleotide sequences. Codons are numbered according to the standard numerical order. CATT and TATA boxes (******), beginning of exons and introns, and positions of mutations disrupting the proteins are marked with bold letters above the sequences.

A comparison of the three intron sequences in Fig. 3 reveals a gradient of divergence: the more 3′ the position of an intron, the more its sequence has diverged between the three *Gpa* genes. As shown in Table 1A, this gradient is observed for all three pairwise comparisons and, in addition, has a similar steepness in each case: *ψ1/Gpa1* (18.6, 29.9, 52.9%), *ψ2/Gpa1* (21.2, 34.4, 54.7%), and *ψ1/ψ2* (4.3, 6.7, 9.9%). Therefore, this pattern has been generated independently in different genes, suggesting that it may be due to a basic, as yet unknown, mechanism rather than to coincidence. This 5′ → 3′ sequence divergence polarity in *Gpa* introns may be explained, if one accepts that reverse transcription and homologous recombination (gene conversion) occur in higher plants (for processed pseudogenes and retroposons in plants see Drouin and Dover 1987; Voytas and Ausubel 1988; Grandbastien et al. 1989). Reverse transcription always starts at the 3′ end and rarely extends completely to the 5′ end of the RNA. Consequently, repetitive reverse transcriptions of a pre-mRNA followed by homologous recombination (gene conversion) may be expected to generate the observed pattern: the more 3′ the position of an intron, the more errors from reverse transcription could accumulate, whereas most errors in exons would be eliminated by selection. There are several interrupted oligo(A) tracts downstream of the stop codon of *ψ1* (not shown in Fig. 3) that may or may not correspond to remnants of poly(A) tails of former RNA intermediates: A_4CA_4, A_4TA_3, A_7CTA_3TACAA at positions +176, +202, and +313, respectively. Pseudogene 2 has an interrupted oligo(T) tract ($T_4CTTCT_7CCTTACGT_6$) at position +150 starting 13 bases downstream of a conventional polyadenylation signal AATAAA (not shown in Fig. 3).

The present observation may be related to our previous finding (Martinez et al. 1989) that the

Table 1. Sequence comparisons between the functional *Gpa1* gene and two related pseudogenes ($\psi 1$ and $\psi 2$)

A) Noncoding sequences (promoter + introns): differences in %[a]

	Promoter	Intron 1	Intron 2	Intron 3	All introns: Observed	All introns: Corrected
$\psi 1/Gpa1$	11.5	18.6	29.9	52.9	31.6	43
$\psi 2/Gpa1$	14.6	21.2	34.4	54.7	34.9	47
$\psi 1/\psi 2$	12.2	4.3	6.7	9.9	6.6	8

B) Coding sequences: shared differences in $\psi 1$ and $\psi 2$

	31 nucleotide substitutions: Replacement	Silent
$\psi 1, 2/Gpa1$	8 (5 G+C)	23 (19 G+C)

C) Coding sequences: unique differences in $\psi 1$ and $\psi 2$

	74 nucleotide substitutions: Replacement	Silent	% total change (1218 sites)
$\psi 1/Gpa'$	11 (8 A+T)	7 (5 A+T)	1.5%
$\psi 2/Gpa'$	34 (29 A+T)	22 (20 A+T)	4.6%

[a] Each insertion and deletion was scored as a single mutational event independent of size

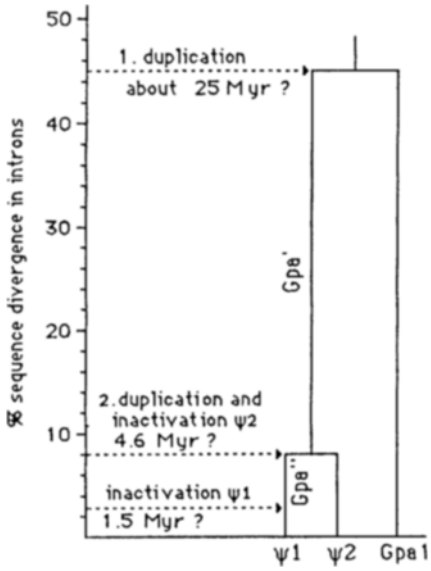

Fig. 4. Evolution of *Gpa* pseudogenes $\psi 1$ and $\psi 2$. The tree is based on the intron sequence differences for the three pairwise comparisons $\psi 1/Gpa1$, $\psi 2/Gpa1$, and $\psi 1/\psi 2$ as shown in Table 1A and explained in the text. Estimations of dates for duplications and nonfunctionalizations of genes are based on the assumption that introns and pseudogenes evolve at about the same rate and that *Gpa* pseudogenes from maize change about two times faster than the average animal pseudogene (4.85×10^{-9}; see text and Li et al. 1985). *Gpa'* and *Gpa''* are successive variants of a second functional *Gpa* gene active in the past (see text).

placement of introns in the maize gene encoding cytosolic GAPDH (subunit GAPC, gene *Gpc1*) relative to the chicken gene follows a gradient: the more 3′ their position, the more strongly *Gpc1* introns are displaced relative to the chicken introns. It seems possible, therefore, that the two observations, sequence gradient and placement polarity of introns, represent the short- and long-time effects, respectively, of the same genetic mechanism [that is, the homologous recombination of genes with their reverse-transcribed (modified) mRNA precursors (see Martinez et al. 1989)].

The average values of intron sequence divergence, corrected for multiple mutations at the same site (Jukes and Cantor 1969), clearly suggest that the two pseudogenes (8% difference) are more related to each other than either is to the functional *Gpa1* gene (43% and 47% difference for $\psi 1$ and $\psi 2$, respectively; see Table 1A). This suggests that the evolution of $\psi 1$ and $\psi 2$ involved at least two consecutive duplication events as depicted by the evolutionary scheme in Fig. 4.

The Progenitor of $\psi 1$ and $\psi 2$ Was a Second Functional Gpa Gene (Gpa') with an Accelerated Evolutionary Rate

Within the coding region, $\psi 1$ and $\psi 2$ share 31 nucleotide substitutions in identical positions relative to the functional *Gpa1* gene (boxed nucleotides in Fig. 3; see Table 1B). In addition, there are two single codon insertions at identical positions (however, different in sequence) in the region encoding the transit peptide in $\psi 1$ and $\psi 2$ (positions −41A and −8A; see Fig. 3). These shared differences reflect the common history of $\psi 1$ and $\psi 2$ (see Fig. 4). The last common progenitor can be reconstructed within the coding region by substituting boxed nucleotides for their homologues in the *Gpa1* sequence and by inserting codons −41A and −8A of $\psi 1$, the least diverged of the two pseudogenes (see below). From the pattern of substitutional events that occurred in this common progenitor encoding *Gpa'* (see Fig. 4) we can conclude that it was almost certainly an active gene. Evidence for this is found along several lines. A high degree of functional constraint for the protein product encoded by the pseudogene progenitor is indicated by the predominance of 23 silent over 8 replacement substitutions relative to *Gpa1* (Table 1B). For both silent and replacement substitutions, the strong preference for high G+C content characteristic of the active *Gpa1* gene (see Fig. 1) is clearly present in the pseudogene progenitor, reflecting continued selective pressure at the level of synonymous substitutions. Both insertions in the transit peptide relative to *Gpa1* (−41A and −8A) maintain the reading frame, and essential nucleotides at splicing junctions have been conserved. Furthermore, none of the events that have rendered $\psi 1$

and *ψ2* incapable of proper function (deletion, insertion, and internal stop codon; see above) is common to either pseudogene, strongly suggesting that they became nonfunctional independently subsequent to duplication of their functional progenitor encoding *Gpa'* (see Fig. 4).

As shown in Figs. 3 and 4, *Gpa1* and *Gpa'* accumulated seven replacement substitutions within the mature subunit (337 codons) during an estimated period of 20 million years (Myr) (see below). Alignment of the maize GAPA' amino acid sequence with all other GAPA and GAPB sequences known to the present day (pea, tobacco, mustard, maize; not shown, see Brinkmann et al. 1989) shows that three out of the seven replacement mutations occurred in sites (lys 58 → glu, ile 89 → val, lys 248 → arg; see Fig. 3) occupied by identical amino acids in all other GAPA and GAPB sequences (including maize GAPA also is different (see Brinkmann et al. sites (codons 25, 215, and 268; see Fig. 3) where maize GAPA is different also (see Brinkmann et al. 1989). This indicates that at least five (three plus two) of the seven replacement substitutions occurred in *Gpa'* and only two in *Gpa1*. Hence, *Gpa'* changed at least two to three times faster than its counterpart *Gpa1* during the period between the first and second gene duplication.

The GAPA'-specific changes lysine 248 → arginine and lysine 58 → glutamic acid are particularly interesting because they indicate that GAPA' was probably a defective enzyme. This may be concluded because lysine 248 is conserved strictly in the GAPDH polypeptides of all organisms except in that of the extreme thermophile *Thermus aquaticus*, where it is replaced by arginine as in GAPA'. Lysine 58 is conserved in most GAPDH sequences and has been replaced by glutamic acid only in the *E. coli* enzyme (for sequence comparisons see Martin and Cerff 1986).

Acceleration of evolutionary rate following gene duplication has been reported previously for the β-globin gene family of goat (β^A, β^C, γ) by Li and Gojobori (1983) and for the duplicate genes DC3 and DC4 encoding cytochrome c in *Drosophila melanogaster* by Wu et al. (1986). In the latter case it has been shown that DC3 evolves about three times faster than DC4 and, in addition, that DC3 is expressed only weakly compared to DC4 (see Limbach and Wu 1985). This indicates that the cause of rate acceleration of DC3 is relaxation of selective constraints rather than advantageous mutations (see Wu et al. 1986). The present findings, showing that genes with an accelerated evolutionary rate may be defective and eventually become pseudogenes, support this neutralist view. It also may be speculated that in the past *Gpa'* and *Gpa"* were, and, in the present, *Gpa"'* (see below) possibly still is expressed only weakly in maize plants. However, if this is true, our original hypothesis that a high G+C content and strong gene expression are correlated (see above and Brinkmann et al. 1987) probably would need to be modified. It may be, for instance, that once a CpG island of a strongly expressed gene is saturated in G+C, it would maintain this status after duplication also under conditions of low expression, unless the gene is completely nonfunctionalized (see below).

Estimation of Rates and Nonfunctionalization Dates of Gpa Pseudogenes

Substitutions unique to *ψ1* and *ψ2* relative to *Gpa'* (see alignment in Fig. 3) reflect the course of independent evolution of the two pseudogenes after duplication of the *Gpa'* progenitor (second duplication in Fig. 4). As shown in Table 1C, the majority of these 74 unique differences are replacement substitutions consisting mainly of A+T. This indicates that most of these mutational events occurred after nonfunctionalization of the two pseudogenes in the absence of selective pressure on replacement sites as well as in the absence of GC pressure on codon usage. Pseudogenes 1 and 2 show 1.5% and 4.6% total sequence divergence relative to the *Gpa'* progenitor (see Table 1C). Hence, the *ψ2* coding sequences accumulated about half as many mutations (4.6%) as the intron sequences in both pseudogenes since their separation (about 8%; see Table 1A). Assuming that pseudogenes and introns diverge at about the same rate (Li et al. 1985), this would mean that *ψ2* was inactivated during or shortly after the second duplication, while *ψ1* became nonfunctional more recently, possibly after a third duplication of the *Gpa"* gene (see Fig. 4) leading to *ψ1* and a *Gpa"'* gene still functional in present-day maize plants. The dates of nonfunctionalization (see Fig. 4) are difficult to assess because the evolutionary rates of plant nuclear pseudogenes are unknown (Wolfe et al. 1987). However, if one assumes that *Gpa* pseudogenes from maize, because of their high degree of methylation (see below and Table 2), change about two times faster than the average animal pseudogene (4.85×10^{-9} per base and year; see Li et al. 1985), the present percentage values would transform into 1.5 Myr for inactivation of *ψ1*, 4.6 Myr for inactivation of *ψ2* and the second duplication, and about 25 Myr for the first duplication.

Apparent Methylation of Gpa Pseudogenes in Symmetrical Sites CpG and CpXpG

A close inspection of the unique base substitutions in *ψ1* and *ψ2* shows that 48 of the 74 unique changes occur in CpG and CpXpG (positions 1 and 3) methylation sites and are mainly C → T and G → A transitions (for which there are 19 and 20 respec-

Table 2. Unique differences in pseudogenes 1 and 2 ($\psi 1+2$) are mainly C → T and G → A transitions occurring predominantly in CpG and CpXpG methylation sites

Number and type of changes (*Gpa'* → $\psi 1+2$)		Changes in CpG[a]	Changes in CpXpG[a]	Changes in Non-CpG/CpXpG
	26 C → T	12.5	6.5	7 [4 in CpC]
	6 C → A	1	1	4
	4 C → G	1	1	2
	25 G → A	16	4	5 [5 in GpG]
	5 G → T	3	0	2
	4 G → C	2	0	2
	2 T → C	0	0	2
	1 T → G	0	0	1
	1 A → G	0	0	1
Total changes:	74	35.5	12.5	26
Total sites[a]:	1218	290	178	750
% changes:	6.1	12.2	7	3.5
Rate factors[b]:	2.3	4.5	2.6	1.3

[a] Overlapping sites, positions 1 and 3 in triplets CpGpG and CpCpG, respectively, and mutations therein were scored 0.5 for either CpG and CpXpG

[b] The rate factors represent the total mutation rates in units "normal mutation rate" (percent changes in each column divided by 2.7% "normal" change, independent of methylation; see text)

tively; see Table 2, columns 2 and 3). This leaves little doubt that the present pseudogenes are methylated at symmetrical sites of CpG and CpXpG (Gruenbaum et al. 1981; Antequera and Bird 1988) and that the deamination of 5-methylcytosine to thymine in these sites on both DNA strands is the major cause of the selective loss of G+C in $\psi 1$ and $\psi 2$. As shown in Table 2, 12.2% and 7.0% of the potential methylation sites in CpGs and CpXpGs, respectively, have been turned over compared to only a 3.5% total change in the rest of the sequence. Interestingly, C → T and G → A transitions also are overrepresented outside CpGs and CpXpGs, and 9 out of the total 12 transitions occurred in CpC or GpG doublets, respectively (see Table 2, column 4). This may be due to partial methylation of CpC doublets on both DNA strands, resulting in a methylation pattern that is neither symmetrical nor inheritable.

Although deamination of 5-methylcytosine seems to be the major driving force for the preferential elimination of G+C in *Gpa* pseudogenes, it may not be the only cause. As shown in Table 2 (column 1), there are 10 C → A/G and 9 G → T/C substitutions, suggesting that 5 C → T and 4 or 5 G → A transitions may be "normal" mutations not due to 5-methylcytosine deamination, assuming that all six "normal" C/G substitutions are random. This means that there are 20 "normal" G/C → A/T substitutions compared to only 4 A/T → G/C substitutions (see Table 2, column 1). Because the *Gpa'* coding sequence has 67% G+C, one-third of all normal G/C ↔ A/T changes would be expected to be A/T → G/C substitutions instead of only the one-sixth observed. This raises the intriguing question of whether there is a methylation-independent AT mutation pressure present in the maize nuclear genome possibly involving a biased DNA polymerase or DNA repair system (see Sueoka 1988). Because the total G+C content of the maize genome is 48% (Salinas et al. 1988), this mutation pressure would not reflect a general mechanism but would have to be specific for G+C-rich (nonfunctional) sequences.

Based on the above calculations there are 33 "normal" substitutions (15 C → G/A/T, 14 G → C/A/T, 3 T → C/G, and 1 A → G changes; see above and Table 2, column 1) in 1218 sites of $\psi 1$ and $\psi 2$, corresponding to 2.7% methylation-independent changes. Hence, the methylation-dependent turnover of CpG (12.2% change) and CpXpG (7% change) would be 4.5- and 2.6-fold higher, respectively, than the background mutation rate (see rate factors in Table 2).

Conclusions

The present work demonstrates that the high G+C content of the maize *Gpa1* gene is maintained independently of the surrounding noncoding sequences, which are G+C poor (Fig. 1), and only as long as the gene encodes a functional protein. This suggests that the maize *Gpa1* gene is under strong *functional* GC pressure leading to a local G+C enrichment (CpG island; see also Martinez et al. 1989) that is independent of the long-range G+C fluctuations reported for monocot genomes (Salinas et al. 1988).

After nonfunctionalization, *Gpa* pseudogenes rapidly loose G+C mainly due to enhanced 5 mC-mutability in CpG and CpXpG methylation sites. The corresponding turnover rates are about five times (CpGs) and three times (CpXpGs) faster than the background mutation rate. Hence, *Gpa* pseudogenes are methylated, and methylation was either the cause or the immediate consequence of their nonfunctionalization. In either case, cytosine methylation would have a permanent silencing effect on gene transcription (for review see Cedar 1988) and, in addition, would be an efficient means to homogenize compositional discontinuities in nontranscribed DNA regions.

Pseudogene-specific CpG erosion by mutability of 5-methylcytosine also occurs in vertebrates, as shown for the human α-globin pseudogene (Bird et al. 1987). This raises the intriguing question of how the regulation apparatus of the nuclear genome from vertebrates and higher plants distinguishes between

G+C-rich pseudogenes to be silenced by cytosine methylation and their functional counterparts that are to be allowed to approach saturation in G+C content and, hence, CpG and CpXpG methylation sites.

Intron sequences of *Gpa* genes reveal a gradient of divergence with a similar steepness in the 5′ → 3′ orientation for all three pairwise gene comparisons. The generation of this polarity may be explained by a gene conversion model suggesting that there is homologous recombination of genes with their reverse-transcribed mRNA precursors.

Another interesting outcome from this study is the observation that the progenitor of *Gpa* pseudogenes was a second functional *Gpa* gene (*Gpa′*) with an accelerated evolutionary rate. The corresponding polypeptide GAPA′ has changed in three amino acid positions that are conserved strictly in all other GAPA and GAPB sequences from higher plants, two of which also are conserved in the GAPDH proteins from most other organisms. This suggests that GAPA′, although active for a long time (~20 Myr, see Fig. 4), has functioned less efficiently than GAPA, supporting the neutralist view that rate acceleration after duplication is more likely to be due to relaxation of functional constraints than to advantageous mutations.

Acknowledgments. The excellent technical assistance of M. Rocipon is acknowledged gratefully. This work was funded by grants from the Centre National de la Recherche Scientifique (UA 1178), the Ministère de la Recherche et Technologie (MRT, program: "Génétique et Physiologie des Végétaux Supérieurs"), and the Ministère de l'Education Nationale (specific program: "Essor des Biotechnologies"). H. Brinkmann is a recipient of a grant from the Deutsche Forschungsgemeinschaft.

References

Antequera F, Bird AP (1988) Unmethylated CpG islands associated with genes in higher plant DNA. EMBO J 7:2295–2299

Aota S, Ikemura T (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res 14:6345–6355

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Biesecker G, Harris JI, Thierry JC, Walker JE, Wonacott AJ (1977) Sequence and structure of D-glyceraldehyde 3-phosphate dehydrogenase from *Bacillus stearothermophilus*. Nature (Lond) 266:328–333

Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus. Trends Genet 3:342–347

Bird AP, Taggart MH, Nicholls RD, Higgs DR (1987) Non-methylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin pseudogene. EMBO J 6:999–1004

Brinkmann H, Martinez P, Quigley F, Martin W, Cerff R (1987) Endosymbiotic origin and codon bias of the nuclear gene for chloroplast glyceraldehyde 3-phosphate dehydrogenase from maize. J Mol Evol 26:320–328

Brinkmann H, Cerff R, Salomon M, Soll J (1989) Cloning and sequence analysis of cDNAs encoding the cytosolic precursors of subunits GapA and GapB of chloroplast glyceraldehyde-3-phosphate dehydrogenases from pea and spinich. Plant Mol Biol 13:81–94

Cedar H (1988) DNA methylation and gene activity. Cell 53: 3–4

Cerff R, Kloppstech K (1982) Structural diversity and differential light control of mRNAs coding for angiosperm glyceraldehyde-3-phosphate dehydrogenases. Proc Natl Acad Sci USA 79:7624–7628

Chojecki J (1986) Identification and characterization of a cDNA clone for cytosolic glyceraldehyde-3-phosphate dehydrogenase in barley. Carlsberg Res Commun 51:203–210

Drouin G, Dover GA (1987) A plant processed pseudogene. Nature 328:557–558

Federoff N (1983) Plant Mol Biol Rep 1:27–29

Frischauf A-M, Lehrbach H, Poustka A, Murray N (1983) J Mol Biol 170:827–842

Grandbastien M-A, Spielmann A, Caboche M (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature 337:376–380

Gruenbaum Y, Naveh-Many T, Cedar H, Razin A (1981) Sequence specificity of methylation in higher plant DNA. Nature 292:860–862

Hanauer A, Mandel JL (1984) The glyceraldehyde 3-phosphate dehydrogenase gene family: structure of a human cDNA and of an X chromosome linked pseudogene; amazing complexity of the gene family in mouse. EMBO J 3:2627–2633

Hohn B (1979) In: Colowick SP, Kaplan NO (eds) Methods in enzymology 68:299–309

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

Klösgen RB, Gierl A, Schwarz-Sommer ZS, Saedler H (1986) Molecular analysis of the *waxy* locus of *Zea mays*. Mol Gen Genet 203:237–244

Li WH, Gojobori T (1983) Rapid evolution of goat and sheep globin genes following gene duplication. Mol Biol Evol 1:94–108

Li WH, Luo CC, Wu CI (1985) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 1–94

Limbach KJ, Wu R (1985) Characterization of two *Drosophila melanogaster* cytochrome c genes and their transcripts. Nucleic Acids Res 13:631–644

Maniatis T, Fritsch EF, Sambrook G (1982) Synthesis of cDNA. In: Molecular cloning. A laboratory manual. Cold Spring Harbor Laboratory Press, New York, pp 211–246

Martin W, Cerff R (1986) Prokaryotic features of a nucleus-encoded enzyme. Eur J Biochem 159:323–331

Martinez P, Martin WF, Cerff R (1989) Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase (GAPC) from maize. J Mol Biol 208:551–566

Niesbach-Klösgen U, Barzen E, Bernhardt J, Rohde W, Schwarz-Sommer ZS, Reif HJ, Wienand U, Saedler H (1987) Chalcone synthase genes in plants: a tool to study evolutionary relationships. J Mol Evol 26:213–225

Piechaczyk M, Blanchard JM, Sabouty SR-E, Dani C, Marty L, Jeanteur P (1984) Unusual abundance of vertebrate 3-phosphate dehydrogenase pseudogenes. Nature 312:469–471

Quigley F, Martin WF, Cerff R (1988) Intron conservation across the prokaryote–eukaryote boundary: structure of the

nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. Proc Natl Acad Sci USA 85:2672–2676

Russel D, Sachs M (1989) Differential expression and sequence analysis of the maize glyceraldehyde-3-phosphate dehydrogenase gene family. The Plant Cell 1:793–803

Saedler H, Nevers P (1985) Transposition in plants: a molecular model. EMBO J 4:585–590

Salinas J, Matassi G, Montero LM, Bernardi G (1988) Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res 16:4269–4285

Schwarz-Sommer ZS, Gierl A, Klösgen RB, Wienand U, Peterson PA, Saedler H (1984) The Spm (En) transposable element controls the excision of a 2 kb DNA insert at the *wx-m8* allele of *Zea mays*. EMBO J 3:1021–1028

Stone EM, Rothblum KN, Alevy MC, Kuo TM, Schwartz RJ (1985) Complete sequence of the chicken glyceraldehyde-3-phosphate dehydrogenase gene. Proc Natl Acad Sci USA 82: 1628–1632

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Voytas DF, Ausubel FM (1988) A copia-like transposable element family in *Arabidopsis thaliana*. Nature 336:242–244

Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc Natl Acad Sci USA 84:9054–9058

Wu CI, Li WH, Shen JJ, Scarpulla RC, Limbach KJ, Wu R (1986) Evolution of cytochrome c genes and pseudogenes. J Mol Evol 23:61–75

Received January 16, 1989/Revised April 27, 1989

Note Added in Proof. The present sequence data will appear in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases under the accession numbers X15406 (*ψGpa1*, 5022 bp), X15407 (*ψGpa2*, 8839 bp) and X15408 (functional gene *Gpa1*, 6414 bp).