J Mol Evol (1997) 45:499–508

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1997

Evolution of Orthologous Intronless and Intron-Bearing Globin Genes in Two Insect Species

Mary Gruhl,[1] Wen-Yen Kao,[2] Gerald Bergtrom[1]

[1] Department of Biological Sciences, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201, USA
[2] Institute of Molecular Biology, Academia Sinica, Nankang, Taipei 11529, Republic of China

Received: 31 December 1996 / Accepted: 16 May 1997

Abstract. While globin genes *ctt-2*β and *ctt-9.1* in *Chironomus thummi thummi* each have a single intron, all of the other insect globin genes reported so far are intronless. We analyzed four globin genes linked to the two intron-bearing genes in *C. th. thummi.* Three have a single intron at the same position as *ctt-2*β and *ctt-9.1;* the fourth is intronless and lies between intron bearing genes. Finally, in addition to its intron, one gene (*ctt-13RT*) was recently interrupted by retrotransposition. Phylogenetic analyses show that the six genes in *C. th. thummi* share common ancestry with five globin genes in the distantly related species *C. tentans,* and that a 5-gene ancestral cluster predates the divergence of the two species. One gene in the ancestral cluster gave rise to *ctn-ORFB* in *C. tentans,* and duplicated in *C. th. thummi* to create *ctt-11* and *ctt-12.* From parsimonious calculations of evolutionary distances since speciation, *ctt-11, ctt-12,* and *ctn-ORFB* evolved rapidly, while *ctn-ORFE* in *C. tentans* evolved slowly compared to other globin genes in the clusters. While these four globins are under selective pressure, we suggest that most chironomid globin genes were not selected for their unique function. Instead, we propose that high gene copy number itself was selected because conditions favored organisms that could synthesize more hemoglobin. High gene copy number selection to produce more of a useful product may be the basis of forming multigene families, all of whose members initially accumulate neutral substitutions while retaining essential function. Maintenance of a large family of globin genes not only ensured high levels of hemoglobin production, but may have facilitated the extensive divergence of chironomids into as many as 5000 species.

Key words: Insect globin — Maximum parsimony — Molecular evolution — Multigene families — Retroposon — SINE — Neutral evolution — Speciation

Introduction

The larval hemolymph of *Chironomus thummi thummi* (Insecta, Diptera) contains at least 12 hemoglobins (Braun et al. 1968; Bergtrom et al. 1976; Laufer et al. 1976). The globin polypeptides are encoded by an even larger multigene family. The first chironomid globin gene clone to be isolated has two copies each of genes *ctt-3* and *ctt-4* as well as *ctt-E* gene, a novel gene whose product has yet to be demonstrated (Antoine and Neissing 1984; Antoine et al. 1987). Subsequently, *ctt-1, ctt-1A, ctt-6,* eight novel *ctt-7B* gene variants, and four more novel genes designated Y, W, V, and Z were characterized (Saffarini et al. 1985; Trewitt et al. 1987, 1988; Kao et al. 1994; Kao and Bergtrom 1995; Trewitt et al. 1995). Globin genes were also sequenced from *C. th. piger* (a sibling species to *C. th. thummi*), and from *C. tentans,* a more distantly related chironomid (Rozynek et al. 1989, 1991; Hankeln et al. 1991). The observation that the first chironomid globin genes to be characterized are intronless was a surprise, given the absolute conservation of the locations of the outer introns in plant and vertebrate globin genes. Intron loss by an ancestral globin gene,

Correspondence to: G. Bergtrom

possibly involving lateral gene transfer via a cDNA intermediate, was proposed to explain the uninterrupted chironomid globin genes (Lewin 1984; Li and Grauer 1991). However, Kao et al. (1994) discovered two globin genes (*ctt-2*β and *ctt-9.1*[1]) in *C. th. thummi*, each with an intron dividing the coding region of the gene into two roughly equal halves. Central introns had already been found in other invertebrate and in plant globin genes (Stoltzfus and Doolittle 1993; see Kao et al. 1994). Unlike the position of outer introns in the globin genes of vertebrates, plants, and invertebrates (with the exception of the shrimp *Artemia nauplii;* Jellie et al. 1996), the location of the central introns in invertebrate and plant globin genes is neither conserved nor at a position predicted from protein structural analyses by Gō (1981). According to some authors (Rogers 1989; Kwiatowski et al. 1992; Stoltzfus and Doolittle 1993), introns at such discordant locations in genes sharing common ancestry were acquired recently in evolution, rather than by a process of intron sliding as suggested by Gilbert (1987). In considering this possibility, we offered a mechanism whereby an ancestral globin gene acquired a central intron which was then retained by the duplicated *ctt-2*β and *ctt-9.1* genes (Kao et al. 1994). We completed sequencing four more globin genes linked to *ctt-2*β and *ctt-9.1* in *C. th. thummi* to determine if they contain this central (or any other) intron. Of these six genes, one is intronless and another was recently inactivated by transposition. Next, we compared this gene cluster with a cluster of similar but intronless globin genes in *C. tentans* (Rozynek et al. 1991). Orthologies identified by phylogenetic analyses support the existence of an ancestral 5-gene cluster that predates separation of the species. The data indicate that all five descendant genes in *C. tentans* and only one in *C. th. thummi* lost their introns. Since speciation, four of the globin genes show evidence of natural selection, three having evolved more rapidly, and one much more slowly compared to the other genes in the clusters. The absence of clear indications of selective pressure on the other globin genes, and the mutational inactivation of nonorthologous genes in each species is consistent with the selection of high gene copy number as a mechanism favoring abundant hemoglobin synthesis. The implications of these observations are discussed.

Methods

Nomenclature

From comparisons to known *C. th. thummi* globin amino acid sequences (Goodman et al. 1988), three genes reported here are novel, encoding globins whose protein products have not yet been isolated. Building on a systematic approach to naming chironomid globin genes (Kao et al. 1995; Trewitt et al. 1995), these are called *ctt-11, ctt-12,* and *ctt-13* (the numbers 1–10 being previously assigned). The historical use of Roman numerals to name globin polypeptides sequenced at the amino acid level is retained. Several allelic variants of globin genes are reported here and elsewhere. Originally, gene *ctt-9a* was so designated because its inferred amino acid sequence differs from the *CTT-IX* by a single amino acid, $E \rightarrow Q_7$ (Kao et al. 1994). We suggest that putative globin gene alleles be given a numerical rather than an alphabetical suffix. Thus, *ctt-9a,* being the first gene for a *ctt-9* allele, becomes *ctt-9.1* (upon finding it, the gene encoding *ctt-IX* will be named *ctt-9.2*). As noted above, gene and allele designations used here may result in the renaming of some genes reported earlier.

DNA Sequencing and Data Analysis

The generation and screening of a *C. th. thummi* genomic library, the isolation of genomic clone λgb2-1, and the subcloning of overlapping 5-kb *Hind*III/*Hind*III and 5-kb *SalI*/*SalI* fragments containing the *ctt-2*β and *ctt-9.1* genes were described earlier (Kao et al. 1994). Unique oligonucleotide primers were synthesized (BioSynthesis, Inc. Lewisville, TX) based on known sequence data, and used as primers to walk along each fragment until they were completely sequenced in both directions. DNA sequencing was by the dideoxy method of Sanger et al. (1987), and was manual or automated (cycle sequencing with ABI dye-linked dideoxy terminators using an ABI 373 Automated DNA Sequencer; Perkin Elmer, Foster City, CA). The entire 8.9-kb region spanned by the two subclones includes the previously reported *ctt-2*β and *ctt-9.1* genes (GenBank Accession no. AF001292).

DNA sequences were manipulated using the Pustell DNA sequence analysis program (IBI, New Haven, CT). Secretory signal sequences were removed from alignments before comparison because they tend to be less conserved than mature peptide coding DNA. Manually aligned amino acid sequences of mature globins were analyzed by PAUP 3.1 (Swofford 1993) using the branch-and-bound search option (with collapse of zero-length branches) to assure recovery of the most parsimonious trees. PAUP was also used for bootstrap analyses with 1000 replications with the branch-and-bound option (Felsenstein 1985). In this report, clades with bootstrap values of 70% or more were regarded as well supported (Hillis and Bull 1993). The *ctt-6* gene (Kao and Bergtrom 1995), a close relative of the *C. th. thummi* genes sequenced here, was used as the outgroup in most phylogenetic comparisons, though tree topologies are essentially unaltered when the much more distantly related *ctt-1* (Kao et al. 1995) was used as the outgroup (data not shown). MEGA 1.0 (Kumar et al. 1993) was used to compute rates of synonymous and nonsynonymous nucleotide substitutions in coding region DNA by the method Nei and Gojobori (1986) using the Jukes–Cantor correction (1969).

Results and Discussion

Organization and Sequence of Globin Genes in Genomic Clone λgb2-1

A map of the region of genomic clone λgb2-1 containing *ctt-2*β, *ctt-9.1,* and four newly sequenced globin genes is shown in Figure 1. The gene at the 5′ end of the cluster was most similar to *C. th. thummi* Hb VIIA (Kleinschmidt and Braunitzer 1980), but is designated *ctt-7A.1* because it encodes a polypeptide differing from the published amino acid sequence of globin VIIA at two positions ($A_{16} \rightarrow S_{16}$, $S_{25} \rightarrow A_{25}$); barring protein sequenc-

[1] Originally identified as an allele of *ctt-9,* and called *ctt-9a* in Kao et al. 1994; see section on nomenclature.

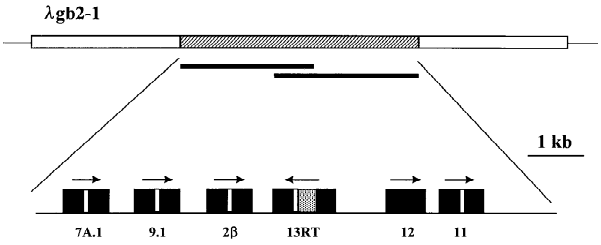

Fig. 1. Genomic clone λgb2-1. Overlapping subcones p2H5 and pIIS5 are the left and right black bars (see Kao et al. 1994). Filled boxes, exons; open boxes, introns; shaded box, retroposon. *Arrows* indicate direction of transcription.

ing errors, *ctt-7A.1* probably encodes an allelic variant of globin VIIA. *ctt-7A.1* contains an intron and has the structural prerequisites of a viable gene, that is, an appropriately placed TATA box, a 16-amino acid signal sequence characteristic of a secreted protein, and appropriately placed stop codon and polyadenylation signal (Fig. 2). The intron contains GT donor and AG acceptor splice sites and is located in the same position as the intron in the *ctt-2*β and *ctt-9.1* genes. In fact, all of the introns in the globin genes in this cluster are located in the same position, supporting the presence of an intron at the same location in their common ancestor.

The *ctt-11* gene, at the 3′ end of the cluster (Fig. 2), has a putative TATA box, 15 triplets encoding a secretory signal sequence, a mature protein coding region comprising exons 1 and 2 separated by the central intron, a stop codon and a polyadenylation signal in the proximal 3′ transcribed, untranslated region. Thus *ctt-11* should also be a viable gene. Comparison of the inferred amino acid sequence of *ctt-11* to other published *C. th. thummi* globin amino acid sequences (Goodman et al. 1988) indicates that this gene encodes a novel globin.

Ctt-12 is a curious gene for several reasons. First, it is intronless. While most globin genes are intronless, we thought that intron-containing globin genes might be maintained by partial gene corrections among members of a linked globin gene cluster. Earlier, we had presented evidence that a high degree of similarity between introns in the *ctt-9.1* and *ctt-2*β genes might in fact be due to a partial gene conversion (Kao et al. 1994). But *ctt-12* lies between two intron-containing genes. *ctt-12* is also unusual in at least one other aspect: there are five serine residues in a single heptapeptide (including four serines in a row), four of which are unique to *ctt-12* among all other chironomid globins. It is noteworthy that the 5′ end of the ''poly-serine'' codon string is only 14 nucleotides away from what, in the other genes in this cluster, would be the 5′ GT exon/intron junction. Was the poly-serine motif generated in some way together with the loss of the central intron in *ctt-12?* And has the poly-serine motif rendered *ctt-12* inactive? Except for the poly-serine motif, *ctt-12* appears to be viable, with a TATA box, a 15-amino acid signal sequence, a coding region, a stop codon, and polyadenylation signal (Fig. 2) Whatever the source of the serine-rich heptapeptide, the location of an apparently transcribable, intronless *ctt-12* between intron-bearing genes leaves open the question of how or why the central introns are retained in some globin genes against the background of intron loss in other globin genes and gene lineages.

The fourth gene, *ctt-13RT* (RT refers to the retroposon inserted in the gene; see below), lies between *ctt-2*β and *ctt-12*, is opposite in transcriptional orientation to the other genes in the cluster, contains an intron and encodes yet another novel globin gene (Fig. 2). However, exon 1 of *ctt-13RT* is interrupted by 457 bp that are neither intronic nor globin coding DNA (Fig. 2). The first thirteen nucleotides of this extra DNA are repeated in the same direction immediately after the other end of the interruption. The repeated motifs are characteristic of the direct repeats flanking a transposable element. A BLAST search of GenBank and EMBL data bases failed to find any DNA sequences similar to the interruption in *ctt-13RT*, but did reveal two short appropriately spaced motifs strongly similar to the A-box and B-box internal promoter sites found in transcripts made by RNA polymerase III (Fig. 2). Together with the presence of a poly d(A)d(T) stretch at the 3′ end of the fragment, these observations indicate that the insert is a SINE-like retroposon derived by reverse transcription of an RNA polymerase III-catalyzed RNA. Though it cannot encode a viable globin polypeptide, the *ctt-13RT* gene itself has a TATA box, 15 triplets encoding a signal peptide, a stop codon, a polyadenylation signal, and an intron with GT and AG donor and acceptor splice sites. Further, alignment of the intron in *ctt-13RT* with the 64–68 nucleotide introns of the other globin genes reveals several well-conserved shorter motifs including one most likely to be the branch site for lariat formation during splicing (Fig. 3). Therefore, transposition appears to have been very recent, a conclusion also supported by data from comparative sequence analyses discussed below. Southern blots of restriction digested *C. th. thummi* genomic DNA hybridized with probes specific for the transposon indicate extensive transposition of the element (data not shown), confirming its mobility.

Inactivation of ctt-13RT *in* C. th. thummi *and* ctn-ORFB *in* C. tentans *Is Recent*

Although *ctt-13RT* (with its retroposon) and *ctn-ORFB* (truncated at the 5′ and 3′ ends) are nonviable genes, both have normal reading frames (no internal stop codons, characteristically conserved amino acids; Goodman et al. 1988) and intact transcription control signals (Rozynek et al. 1991; this report), suggesting that they were recently inactivated. To confirm that the mutations in *ctt-13RT* and *ctn-ORFB* were recent, the rates of nonsynonymous and synonymous mutations (d_N and d_S, respectively) between these genes and their presumably

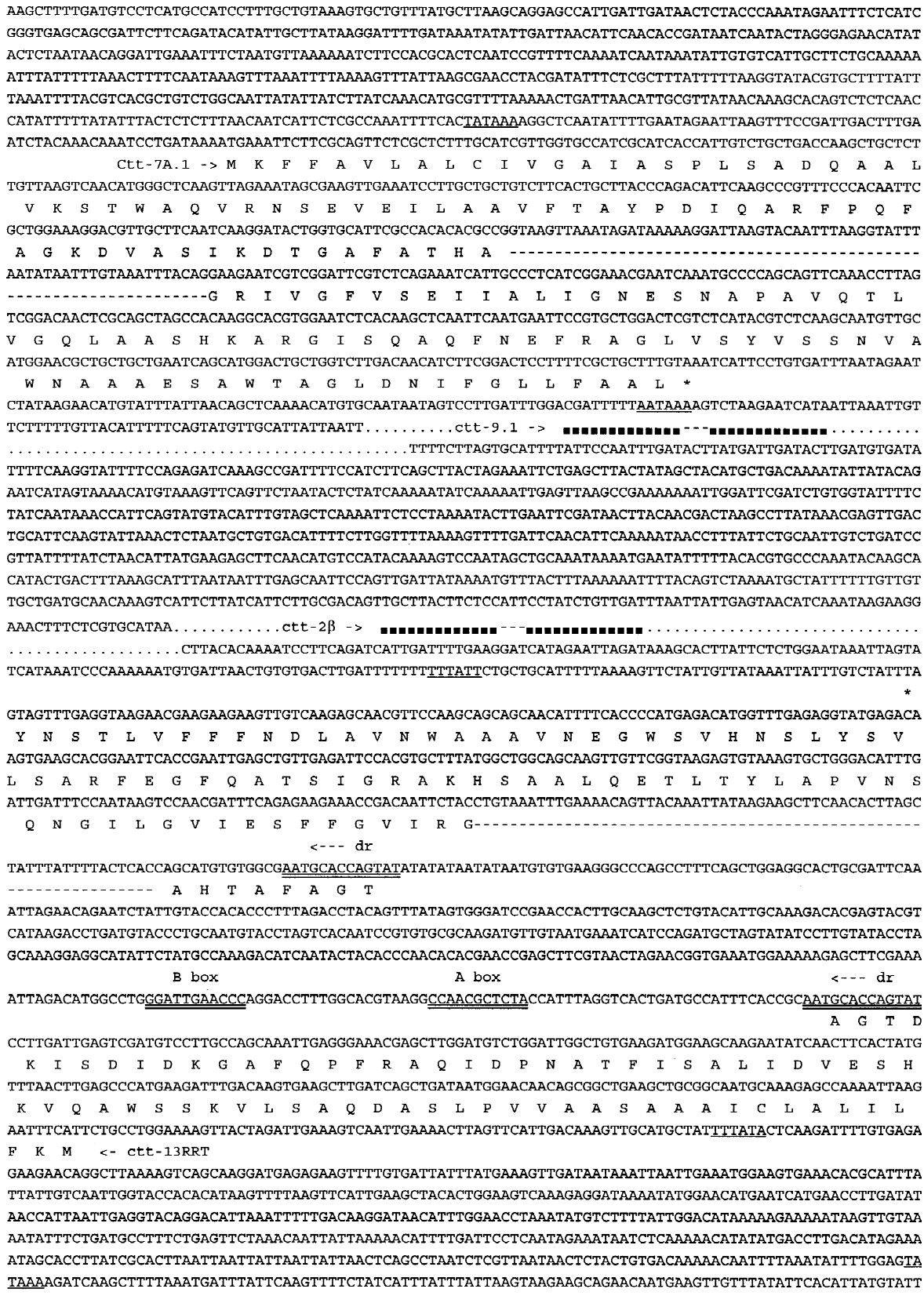

Fig. 2. Sequence of region indicated in Figure 1. Previously sequenced DNA, including regions containing the *ctt-2*β and *ctt-9.1* genes (Kao et al. 1994) are omitted to save space, though the coding regions of the two genes are diagrammed; resulting breaks in the sequence are indicated by boldface base numbers. TATA boxes and AATAAA polyadenylation recognition sequences are underlined. Dashed lines between amino acid sequences are introns. Asterisks indicate stop codons. Short direct repeats (< - - dr, double underlined) flank a transposable element inserted in the coding region of *ctt-13RT*. Appropriately spaced A and B promoter motifs (A box, B box, double underlined) indicate that the element is a retroposon derived by reverse transcription and mobilization of a transcript generated by RNA polymerase III.

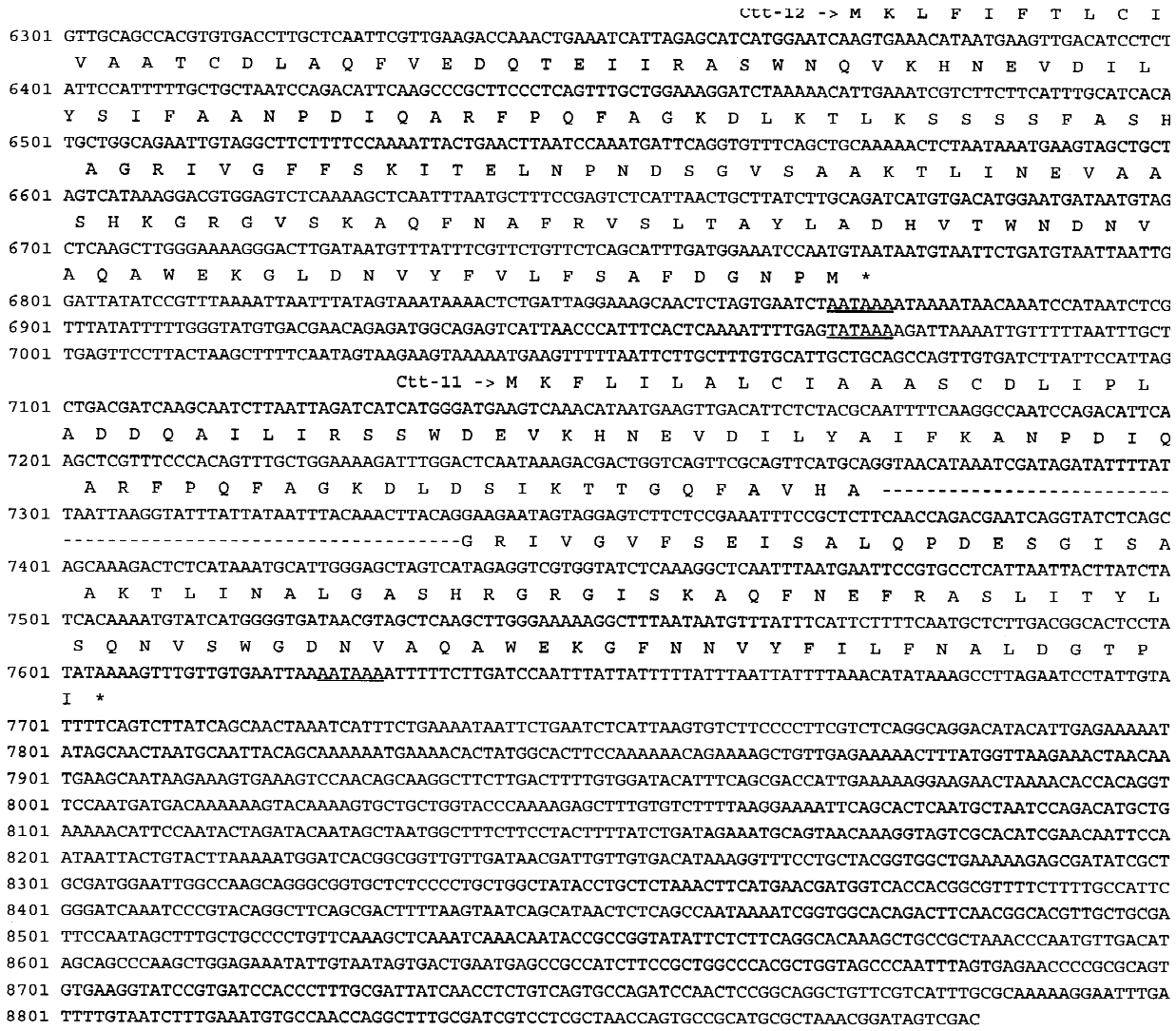

Fig. 2. Continued.

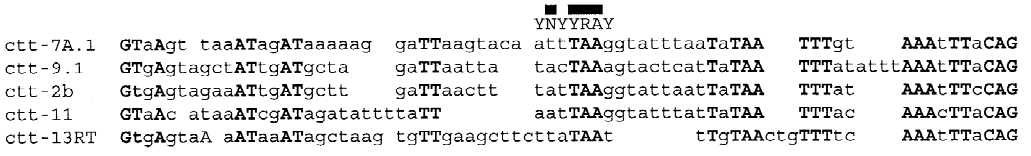

Fig. 3. Alignment of introns of the *C. th. thummi* cluster globin genes showing that essential structural features of the intron in the nonviable *ctt-13RT* gene are still intact. Identical sequences, including GT donor and AG acceptor sites, are in upper case boldface. A consensus sequence including the branch site is shown over the intron sequences (Perlman et al. 1991); dark boxes over the consensus sequence indicate conforming nucleotides in the globin gene introns, and an asterisk indicates the likely branch site.

active paralogues were compared by the method of Nei and Gojobori (1986) using the Jukes–Cantor Correction option in MEGA (Table 1). The d_N/d_S ratio for pseudogenes or rapidly decaying genes (i.e., unselected DNA) should be higher than for viable genes because in nonviable genes both synonymous and nonsynonymous sites are free from selective constraint. The average d_N/d_S ratios for pairwise comparisons of *ctt-13RT* with all of the globin genes in the *thummi* and *tentans* clusters (0.274) is within the range of 0.172–0.317 found for all pairwise combinations of putatively viable genes in the two clusters. The average d_N/d_S ratio for *ctn-ORFB* is 0.174, within the range 0.169–0.286 for viable genes. Quite the opposite to expectations for a disabled gene, the remarkably low d_N/d_S ratio for *ctn-ORFB* implies that the gene is undergoing purifying selection (i.e., eliminating nonsynonymous base substitutions)! Thus, *ctt-13RT* and *ctn-ORFB* must have been inactivated re-

Table 1. Rates of synonymous (d_S) and nonsynonymous (d_N) substitutions per site: pairwise comparisons between genes in the *C. th. thummi* and *C. tentans* globin gene clusters, including inactive *ctt-13RT* and *ctn-ORFB* genes

Alignment	All genes: d_N	All genes: d_S	All genes: Average d_N/d_S	d_N/d_S range, all pairwise possibilities	d_N/d_S, *ctt-13RT* versus all others	d_N/d_S, *ctn-ORFB* versus all others
ctt-13RT[a]	0.30 ± 0.02	1.15 ± 0.12	0.261	0.172–0.317	0.274	NA
ctn-ORFB[a]	0.25 ± 0.02	1.44 ± 0.10	0.176	0.169–0.286	NA	0.174

[a] To confirm that the transposition that inactivated *ctt-13RT* was recent, full-length coding region DNA sequences were compared. To confirm that the mutations that inactivated *ctn-ORFB* by truncation are recent, it was compared only to globin sequences that were trimmed to match the length of *ctn-ORFB*

cently in each species. Apart from the mutations that immediately disrupt their viability, these genes seem not to have suffered other significant changes.

GC-rich DNA in an AT-rich Intergenic DNA Background: Coding Sequence Remnants?

The average A + T content of globin locus intergenic DNA in *C. th. thummi*, *C. th. piger*, *C. tentans*, and *Kiefferulus cornishi* (a distantly related New Zealand chironomid) is unusually high (71%–74%; Table 2), consistent with the high overall genomic A + T content determined for *C. th. thummi* (Wobus 1975; Schmidt et al. 1980). The A + T% of *C. th. thummi* globin gene introns is even higher (78%). In contrast, typical coding region A + T content ranges from 51% to 57%. Regions of high G + C content, easily discerned against the A + T-rich background, are, therefore, candidates for open reading frames (including globin genes). One G + C-rich sequence in *C. th. thummi* begins 250 nucleotides beyond the stop codon of *ctt-11* and extends more than 1 kbp to the end of the gene cluster. This region, however, shows no similarity to any chironomid globins in pairwise comparisons (data not shown). Likewise, a 200+ bp high G + C region in *C. tentans* extending from about nt 1540–nt 1765 (Rozynek et al. 1991) shows no similarity to globin genes. Both regions have only short open reading frames, and a BLAST search failed to match either the *thummi* or *tentans* regions with any sequences in the GenBank or EMBL databases. Apparently, these G + C-rich regions are neither decadent chironomid globin genes nor remnants of other known genes. The regions may be 1) corruptions of genes not yet entered into world data bases, 2) decaying transposons, or 3) conserved for some other, e.g., structural purpose. In *C. th. thummi*, the presence of many runs of oligo(dA) and oligo(dT) within the G + C-rich region suggests that the region is conforming to a more intergenic character, and argues against sequence conservation for some structural or other purpose.

Evolution of the Globin Gene Clusters in C. th. thummi *and* C. tentans

Orthologous Globin Genes

A more than 80% similarity between amino acid sequences inferred from *C. th. thummi* and *C. tentans* globin genes supports the orthologous *7A.1/ORFE, 9.1/ORFD,* and *2*β*/ORFC* gene pairs suggested by similar, earlier studies (Rozynek et al. 1991; Kao et al. 1994), and also suggests the orthology of *ctt-13RT* and *ctn-ORFA.* The relationship between *ctn-ORFB* and either *ctt-11* and *ctt-12* was not obvious from the same percent similarity matrix (data not shown), but was clarified by an interspecific phylogenetic analysis of mature globin sequences encoded by the six *C. th. thummi* and five *C. tentans* globin genes.

Table 2. Average A + T% in globin gene DNA and in intergenic DNA

Organism	Exon I	Exon II	Intron	Intronless gene coding DNA	Intergenic DNA
C. th. thummi[a]	54.86	56.56	78.28	55.85[e,f,g,h]	72.74
C. th. piger[b]	—	—	—	55.73	74.28
C. tentans[c]	—	—	—	52.60	70.77
K. cornishi[d]	—	—	—	50.93	71.69
Average	54.86	56.56	78.28	53.73	72.37

[a] Kao et al. 1994
[b] Hankeln et al. 1991
[c] Rozynek et al. 1991
[d] Chen et al. 1995
[e] Kao and Bergtrom 1995
[f] Kao et al. 1995
[g] Trewitt et al. 1987
[h] Trewitt et al. 1995

The single branch-and-bound tree that resulted from the maximum parsimony phylogenetic analysis by PAUP (Fig. 4A) confirms orthologous *ctn-ORFE/ctt-7A.1, ctn-ORFD/ctt-9.1 ctn-ORFC/ctt-2*β, and *ctn-ORFA/ctt-13RT* gene pairs, but did not resolve the relationship between *ctn-ORFB* and *ctt-11* or *ctt-12.* The phylogeny obtained with bootstrapping (1000 replications; Fig. 4B) strongly supported these same five clades, recovering the branches in at least 78% of the replicates. All branches with less than 70% support were collapsed to polytomy in this analysis. The especially robust 84% support of the clade including *ctt-11, ctt-12,* and ctn-ORFB indicates that the *ctt-11* and *ctt-12* genes arose from a duplication event after species divergence, and that the correct orthology is between the *ctt-11/ctt-12* ancestor in *C. th. thummi* and the ancestor to *ctn-ORFB* in *C. tentans.* We

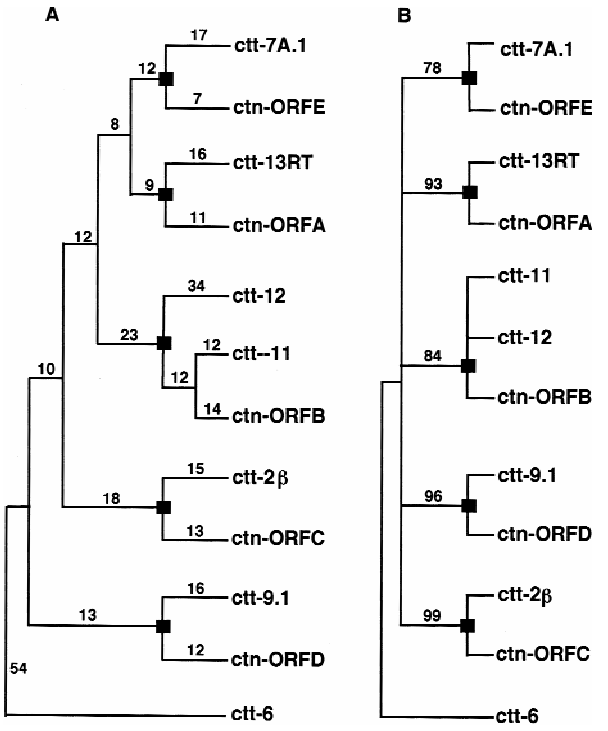

Fig. 4. Maximum parsimony analysis of full-length, manually aligned amino acid sequences inferred from gene coding regions using PAUP (Swofford 1993). Because secretory signal sequences tend to vary, only mature globin polypeptides were aligned. Gaps in alignments were retained, including those representing the missing amino acids in the truncated *ctn-ORFB* gene. *ctt-6* served as the outgroup sequence. Filled boxes at nodes ancestral to orthologous gene lineages also mark the relative time of the *thummi/tentans* species divergence. **(A)** The single tree generated by branch-and-bound analysis, with relative distances shown above the branches (an heuristic search generates the same tree). **(B)** Branch-and-bound with bootstrapping (1000 replicates); the tree indicates the same familial groups as in **(A)** (i.e., *ctt-7A.1/ctn-ORFE, ctt-13RT/ctn-ORFA, ctt-11/ctt-12/ctn-ORFB, ctt-9.1/ctn-ORFD,* and *ctt-2β/ctn-ORFC*), strongly supporting the lineages leading to each group; percentages of bootstrap replicates supporting the lineages are shown above the branches. Tree topologies generated for the same data set by alternate arguments within PAUP, and using *ctt-6.3* or *ctt-1A* (data not shown) as the outgroup were similar. Likewise, inclusion or exclusion of the inactive *ctt-13RT* and *ctn-ORFB* genes in interspecific analyses also resulted in similar tree topology (data not shown). Intraspecific analyses were also done (not shown). These treat paralogous genes as if they arose only in the species being analyzed and, therefore, do not separate the evolution of orthologous genes from divergence of ancestral gene family members before speciation. Nevertheless, branch-and-bound trees for each species had different topologies confirming the conclusion that members of the ancestral globin gene cluster followed different paths after speciation; bootstrap analyses support the *ctt-11/ctt-12* lineage in *C. th. thummi*, while all other lineages in each cluster collapse to polytomy, as expected if five lineages predate the *tentans/thummi* split.

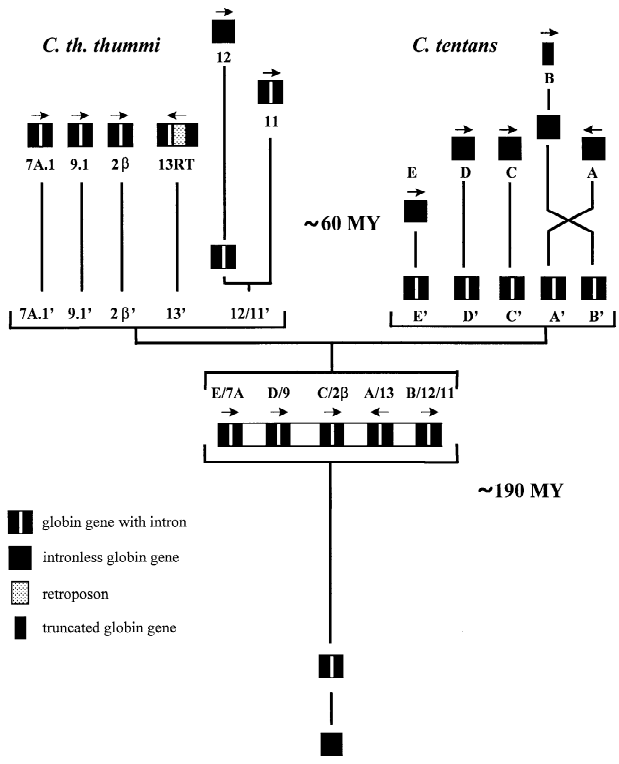

Fig. 5. Proposed evolutionary history of genes in orthologous clusters in *C. th. thummi* and *C. tentans.* Evolutionary times are from Goodman et al. (1988) and Kao et al. (1994). Abbreviated gene designations: 7A′, 9′, 2β′, 13′, 12′, 11′ are immediate ancestors to the genes in *C. th. thummi;* E′, D′, C′, B′, and A′ are immediate ancestors to the genes in *C. tentans.* Because the polytomous origin of five lineages is well-supported (Fig. 4B), five ancestral genes (*E/7A, D/9, C/2*β, *A/13,* and *B/12/11*) are shown predating species divergence. Loss of the intron from *ctt-12* is shown after gene duplication of a *ctt-11/12* ancestor (an event that occurred shortly after speciation). Recent retrotransposition into *ctt-13′,* creating *ctt-13RT,* is indicated. The recent inactivation of *ORFB* by truncation is illustrated, and a reciprocal translocation of *ctn-ORFA′* and *ctn-ORFB′* is proposed in *C. tentans* to explain their transcriptional orientation and that of their orthologues in *C. th. thummi.* The differential accumulation of unambiguous amino acid substitutions by different globins is indicated by (unscaled) differences in branch length. The acquisition of the central intron by an ancestral gene that then duplicated to create the ancestral five-gene cluster (first suggested by Kao et al. 1994) is indicated.

conclude that the five globin genes in *C. tentans* and the six genes in *C. th. thummi* are descended from five genes (filled squares in Fig. 4) already present in the common ancestor to the two species. While the discovery of additional genes in either cluster could reinforce or force revision of these relationships, the well-supported lineages shown here probably reflect true orthologies. The fact that the more basal lineages found in the branch-and-bound analysis collapse to polytomy in the bootstrap tree suggests that these basal lineages originated at or about the same time. This is further supported by intraspecific phylogenetic analyses that confirm the polytomous origin of the six *C. th. thummi* genes in one tree, and the five *C. tentans* genes in another (not shown). We designate the five ancestral genes existing prior to the separation of *C. tentans* and *C. th. thummi* as *E/7A, D/9, C/2*β, *A/13RT,* and *B/12/11* (Fig. 5).

The Role of Positive Selection and Neutral Evolution in Forming Chironomid Globin Genes

The *ctt-7B* loci in *C. th. piger* (Hankeln et al. 1991) and *C. th. thummi* (Trewitt et al. 1995) are well charac-

terized. Further, comparative analyses of the two loci provide ample evidence that, over the short time separating these species (5–10 million years), many of the differences in the globin sequences within and between species are neutral (Trewitt et al. 1995). Though *C. th. thummi* and *C. th. piger* are reproductively isolated (i.e., as adults), they can share the same larval habitat, further supporting the neutrality of differences between homologous globins. *C. th. thummi* and *C. tentans* in contrast are more diverged in both time and habitat. In the 60 million years following their separation, *C. th. thummi* adapted to larval life in a shallow benthos, while larvae of larger *C. tentans* adapted to life in deeper, less oxygenated waters. One would reasonably expect molecular adaptations of hemoglobin to play a major role in this process. Though distances along basal lineages are not meaningful given the polytomous origin of the ancestral cluster prior to speciation (cf. Fig. 4A,B), the distances since speciation may be instructive.

For example, distances from *ctt-7A.1* and *ctn-ORFE* to their common *E/7A* ancestor are 17 and 7, respectively (a 2.4-fold difference). Since *ctt-7A.1* has evolved from its ancestor to about the same degree as most of the other genes (note distances ranging from 11 to 16 for descendants of *A/13, D/9,* and *C/2*β), it appears that *ctn-ORFE* has evolved from *E/7A* more slowly than *ctt-7A.1* (i.e., since the *thummi/tentans* split). The functional basis for overall evolutionary constraint coupled with limited change in *ctn-ORFE* (compared to its orthologue in *C. th. thummi*) is unknown. In contrast to the slow rate of evolution of *ctn-ORFE,* distances from their common *B/11/12* ancestor to *ctt-11* (24), *ctn-ORFB* (26), and especially *ctt-12* (34) are unusually large (Fig. 4A). For *ctn-ORFB,* this faster evolutionary rate could be due to its inactivation by truncation, after which rapid changes could have occurred in the absence of selective pressure since inactivation. However, there is no other evidence of a loss of ''globin'' identity by the gene (e.g., loss of critical amino acids conserved in all chironomid globins), nor does the d_N/d_S ratio of *ctn-ORFB* support significant decay expected for a ''dead'' gene. As noted above, the ratio for *ctn-ORFB* is, if anything, lower than that of viable genes in the cluster. The *ctt-12* gene might have been inactivated by the loss of its intron and an accompanying mutation(s) generating the unusual serine-rich heptapeptide in the vicinity of where an ancestral intron must have been. But *ctt-12* otherwise has all the hallmark signals of a viable gene. Thus the faster evolutionary rates of these globins (as well as *ctt-11*) shortly after the *thummi/tentans* split might yet prove to be the result of diversifying or positive selection of as yet unknown structural and functional advantage(s). In sum, the evolutionary rates of the globin genes in each species since the *thummi/tentans* split suggest three distinct paths: 1) constraint (*ctn-ORFE* versus *ctt-7A.1*); 2) rapid evolution (*ctn-ORFB, ctt-11,* and *ctt-12*); and 3) neither rapid nor slow evolution (all other apparently viable globin genes).

Figure 5 summarizes the evolution of the *C. tentans* and *C. th. thummi* gene clusters. Though not drawn to scale, branch lengths suggest the different evolutionary rates experienced by the globin genes since species divergence. The earlier phylogenetic analyses by Goodman et al. (1988) postulated two successive duplications of an ancestral gene over the course of about 200–250 million years to generate the three globins known at the time. From the polytomous origin of the *E/7A, D/9, C/2*β, *A/13RT,* and *B/11/12* genes demonstrated here, it is more likely that the ancestral gene cluster arose nearly simultaneously, sometime before the *thummi/tentans* divergence. Figure 5 also reiterates the possibility that the ancestor to the five gene cluster acquired (rather than inherited) a central intron (Kao et al. 1994). This assumption and the subsequent intron losses are consistent with a ''most parsimonious'' reconstruction of chironomid globin gene evolution, in which most of the genes are, in fact, intronless, and in which we assume intron loss to be more common than gain. Thus, the loss of an intron by *ctt-12* in *C. th. thummi* and from each of the five *C. tentans* globin genes after speciation is the most acceptable explanation for the existing gene structures. A reciprocal translocation of *ctn-ORFB* and *ctn-ORFA* is postulated to explain their opposite locations but common transcriptional orientation compared to the *ctt-13RT* and *ctt-11/12* ancestors in *C. th. thummi* (the translocation may of course have occurred in the latter).

Gene Copy Number Selection: A Mechanism Explaining the Formation of Multigene Families

Multigene families begin as clustered genes that form from episodic unequal cross-overs between family members. When synthesis of identical products was most important, duplicated genes have tended to homogenization, as with the ribosomal RNA genes for example. In contrast, *chironomids* (and other organisms) have tolerated considerable divergence in globin genes. The 12 hemoglobins present during the fourth instar of *C. th. thummi* include the products of three genes in the cluster presented in this report. Could differences in the structure of these hemoglobins and the likely products of still more globin genes truly be the result of fine tuning positive selection of each globin? We suggest instead that positive selection has operated to ensure adequate hemoglobin synthesis by creating and maintaining a large number of expressible globin genes. Because newly duplicated globin genes increased individual fitness, they were rapidly fixed in chironomid populations. In the presence of so many redundant globin genes, there would be less stringent selection pressure on individual globins, allowing most to accumulate neutral substitutions that do not disrupt essential oxygen binding function. We call this phenomenon *gene copy number selection,* without which the duplicated genes would be truly redundant and subject to evolutionary decay.

Evidence for gene copy number selection is perhaps

strongest in the 7B loci of *C. th. thummi* and *C. th. piger*. Most of these genes encode polypeptides differing by only a few amino acids (Hankeln et al. 1991; Trewitt et al. 1995). In each species, the locus has experienced very recent gene duplication and loss. In *C. th. thummi*, for example, *ctt-7B5* and *ctt-7B9* differ by one nucleotide and encode identical globins (Trewitt et al. 1995), while a *ctt-9/5* chimera, formed by fusion of the *ctt-7B9* and *ctt-7B5* genes and the loss of an intervening gene, has the same coding region nucleotide sequence as *ctt-7B5* (Kao and Bergtrom 1995). Clearly, events involving *ctt-7B5* and *ctt-7B9* were very recent. Gene duplications, fusions, deletions, and most amino acid substitutions in the *7B* gene clusters are probably neutral, tolerated as long as net oxygen binding capacity is not affected. The simplest explanation for the state of the *7B* genes is that new copies were retained and spread because they conferred a beneficial increase in the oxygen binding capacity of the hemolymph (Trewitt et al. 1995). Gene copy number selection at the *ctt-7B* locus (the stable maintenance of globin genes of similar function and essentially neutral divergence) could be seen as an alternative to the selection of stronger promoters enhancing the expression of fewer genes to ensure high larval hemoglobin concentrations.

Although gene conversion may have corrected parts of the *ctt-9.1* and *ctt-2*β genes (Kao et al. 1994), recombinational activity is more limited at the loci reported here than at the *7B* loci. This is undoubtedly due to longer divergence times and greater sequence differences among paralogous genes. The genes in *C. th. thummi* and *C. tentans* do not show the degree of subfamily expansion and contraction seen in the *7B* gene clusters. Further, the conservation of *ctn-ORFE* in *C. tentans* and the rapid change in *ctt-11* and *ctt-12* in *C. th. thummi* indicate that some globins have evolved adaptively. Nevertheless, recent mutations inactivating *ctn-ORFB* and *ctt-13RT* apparently have not harmed the organisms, and six of the 11 genes in the two clusters show comparable rates of evolution. Is it possible that these six globins are accumulating neutral substitutions under minimal selective pressure, and are like the *7B* genes, being retained to ensure ample production of the related hemoglobins encoded within each cluster? We cannot answer this question yet, but it is hard to imagine a unique functional adaptation for so many hemoglobins produced at the same time.

We suggest that after selection of a large number of structurally redundant but functionally useful globin genes, some became the raw material for Darwinian selection of individual globins during adaptation and subsequent speciation. Kimura (1991) termed the accumulation of neutral change among duplicated genes *preadaption*, where preadapted duplicate genes become substrates for positive selection during bursts of speciation. One prediction of preadaptation is that the rate of amino acid substitutions should be higher as new gene duplicates accumulate neutral changes, and slower when new family members undergo positive selection. Thus, Sidow (1992) found that there were four fold fewer amino acid substitutions in the last 500 million years (since the emergence of jawed vertebrates) than in the first 100 million years among the *Wnt* (wingless) genes that regulate embryogenesis. Iwabe et al. (1996) found early rapid evolution in 25 gene families, followed by slower change as ''preadapted'' genes were selected for tissue-specific expression/function.

Since the globin genes studied here are much younger than the gene families studied by Sidow and Iwabe, it may not be possible to detect the changing rates of globin gene evolution. Could we be witnessing the tenuous beginnings of a preadaptive accumulation of neutral amino acid substitutions in the *7B* cluster (tenuous because many *ctt-7B* genes are so similar), and the same process slightly more advanced in the orthologous *C. th. thummi* and *C. tentans* globin gene clusters. Perhaps, while Darwinian selection is evident mainly in the negative selection of *ctn-ORFE* and possibly in the diversifying selection of *ctn-ORFB*, *ctt-12*, and/or *ctt-11*, other globins in the *thummi* and *tentans* clusters are ''banking'' essentially neutral change, Gene copy number selection in the chironomid globin multigene family supports Kimura's hypothesis of preadaptation by neutral divergence (1991) in asserting that many residues in globin genes are safely substitutable with no major consequence to globin function. Selection because ''more is better'' may be the best explanation for the maintenance of the large active chironomid globin gene family. Perhaps the vertebrate globin gene family began with the maintenance of duplicated globin genes that initially conferred the advantage of extra oxygen binding capacity. Spreading rapidly through an ancestral population, some divergence in structure, but not in essential function, would be tolerated. Only later would vertebrate hemoglobins be selected for heterodimer/tetramer formation and differential expression in development, with a consequent slowdown in sequence divergence rates among paralogous genes. In contrast, the chironomid globin multigene family may still be an evolutionary incubator of neutral, preadaptive change. With >5000 chironomid species living in diverse habitats (Oliver, 1971), the hypothesis that many preexisting functionally redundant globin genes were the raw material for Darwinian selection during speciation can be tested by expanding the analyses applied to *C. th. thummi* and *C. tentans* to orthologous globin gene families in other chironomids.

Acknowledgments. We thank Drs. Sara Hoot, and Ruth Phillips and Mr. Patrick M. Trewitt for helpful discussions and critical reading of the manuscript. This work was supported in part by an NSF grant (DCB-8904772) and by the Graduate School of the University of Milwaukee.

References

Antoine M, Niessing J (1984) Intron-less globin genes in the insect *Chironomus thummi, thummi.* Nature 310:795–798

Antoine M, Erbil C, Munch E, Schnell S, Niessing J (1987) Genomic organization and primary structure of five homologous pairs of intron-less genes encoding secretory globins from the insect *Chironomus thummi thummi.* Gene 56:41–51

Bergtrom G, Laufer H, Rogers R (1976) Fat body: a site of hemoglobin synthesis in *Chironomus thummi* (Diptera). J Cell Biol 69:264–274

Braun VV, Crichton RR, Braunitzer G (1968) Über monomere und dimere Insekten-Hämoglobine (*Chironomus thummi*). Z Physiol Chem Hoppe Seyler 349:197–210

Chen Z-Z, Martin J, Lee BTO (1995) Hemoglobins of *Kiefferulus,* sister genus of *Chironomus* (Diptera: Insecta): evolution of the Hb VIIB cluster. J Mol Evol 41:909–919

Schmidt ER, Vistorin G, Keyl H-G (1980) An AT-rich DNA component in the genomes of *Chironomus thummi thummi* and *Chironomus thummi piger.* Chromosoma (Berl.) 76:35–45

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Gilbert W (1987) The exon theory of genes. Cold Spring Harbor Symp Quant Biol 52:901–905

Gō M (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 291:90–92

Goodman M, Pedwaydon J, Czelusniak J, Suzuki T, Gotoh T, Moens L, Shishikura F, Walz D, Vinogradov S (1988) An evolutionary tree for invertebrate globin sequences. J Mol Evol 27:236–249

Hankeln T, Luther C, Rozynek P, Schmidt ER (1991) Complete nucleotide sequence of a hemoglobin gene cluster from the midge *Chironomus thummi piger.* In: Vinogradov S, Kapp OH (eds) Structure and function of invertebrate oxygen carriers. Springer Verlag, New York, pp 287–296

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192

Iwabe N, Kuma K, Miyata T (1996) Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. Mol Biol Evol 13:483–493

Jellie AM, Tate WP, Trotman CNA (1996) Evolutionary history of introns in a multidomain globin gene. J Mol Evol 42:641–647

Jukes TH, Cantor CR (1969) Evolution of protein molecules In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kao W-Y, Bergtrom G (1995) Sequence of globin 6 gene alleles and linkage of globin 6 and 7B genes in the insect *Chironomus thummi thummi.* Gene 153:209–213

Kao W-Y, Trewitt PM, Bergtrom G (1994) Intron-containing globin genes in the insect *Chironomus thummi.* J Mol Evol 38:241–249

Kao W-Y, Hankeln T, Schmidt ER, Bergtrom G (1995) Sequence and evolution of the gene for the monomeric globin I and its linkage to genes coding for dimeric globins in the insect *Chironomus thummi.* J Mol Evol 40:354–361

Kimura M (1991) The theory of evolution: a review of recent evidence. Jpn J Genet 66:367–386

Kleinschmidt T, Braunitzer G (1980) Analyse der Primärstruktur des dimeren Insekten-Hämoglobins CTT VIIA (Erythrocruorin) aus *Chironomus thummi thummi,* Diptera. Hoppe-Seyler's Z Physiol Chem 361:933–942

Kumar S, Tamura K, Nei M (1993) MEGA: Molecular Evolutionary Genetics Analysis version 1.0. Computer program distributed by Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, PA 16802

Kwiatowski J, Skarecky D, Ayala FJ (1992) Structure and sequence of the Cu, Zn SOD gene in the Mediterranean fruit fly, *Ceratitis capitata:* intron insertion/deletion and the evolution of the gene. Mol Phylogen Evol 1:72–82

Laufer H, Bergtrom G, Rogers R (1976) Stage and tissue-specific hemoglobin synthesis in an insect. In: Kurstak E, Marmorosch K (eds) Invertebrate tissue culture: applications in medicine, biology and agriculture. Academic Press, New York, pp 227–240

Lewin R (1984) Surprise finding with insect globin genes. Science 226:328

Li W-H, Grauer D (1991) Fundamentals of molecular evolution. Sinauer Associates Inc., Sunderland, MA

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Oliver DR (1971) Life history of the *Chironomidae.* Ann Rev Enomol 16:211–230

Perlman PS, Peebles CL, Daniels C (1991) Different types of introns and splicing mechanisms. In: Stone ED, Schwartz RZ (eds) Intervening sequences in evolution and development. Oxford University Press, New York, pp 112–161

Rogers JH (1989) How were introns inserted into nuclear genes? Trends Genet 5:213–216

Rozynek P, Hankeln T, Schmidt ER (1989) Structure of a hemoglobin gene cluster and nucleotide sequence of three hemoglobin genes from the midge *Chironomus thummi piger* (Diptera, Insecta). Biol Chem Hoppe-Seyler 370:533–542

Rozynek P, Broecker M, Hankeln T, Schmidt ER (1991) The primary structure of several hemoglobin genes from the genome of *Chironomus tentans.* In: Vinogradov S, Kapp OH (eds) Structure and function of invertebrate oxygen carriers. Springer Verlag, New York, pp 287–296

Saffarini DA, Trewitt PM, Castro M, Wejksnora PJ, Bergtrom G (1985) Deoxynucleotide sequence of an insect cDNA codes for an unreported member of the *Chironomus thummi* globin family. Biochem Biophys Res Comm 133:641–647

Sanger F, Nicklen S, Coulson AR (1987) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Sidow A (1992) Diversification of the Wnt gene family on the ancestral lineage of vertebrates. Proc Natl Acad Sci USA 89:5098–5102

Stoltzfus A, Doolittle WF (1993) Slippery introns and globin gene evolution. Curr Biol 3:215–217

Swofford DL (1993) PAUP: Phylogenetic Analysis Using Parsimony, version 3.1. Computer program distributed by Illinois Natural History Survey, Champaign, IL

Trewitt PM, Saffarini DA, Bergtrom G (1987) Nucleotide sequence of the intronless gene expressing a member of the globin VIIB subfamily from *Chironomus thummi* (Diptera). Nucleic Acids Res 15: 5494

Trewitt PM, Saffarini DA, Bergtrom G (1988) Multiple clustered genes of the haemoglobin VIIB subfamily of *Chironomus thummi thummi* (Diptera). Gene 69:91–100

Trewitt PM, Luhm RA, Samad F, Ramakrishnan S, Kao W-Y, Bergtrom G (1995) Molecular evolutionary analysis of the YWVZ/7B globin gene cluster of the insect *Chironomus thummi.* J Mol Evol 41:313–328

Wobus U (1975) Molecular characterization of an insect genome: *Chironomus thummi.* Eur J Biochem 59:287–293