J Mol Evol (1984) 21:19–32

Journal of
Molecular Evolution

© Springer-Verlag 1984

The Evolution of a Plant Globin Gene Family

Gregory G. Brown, Jong Seob Lee, Normand Brisson, and Desh Pal S. Verma

Department of Biology, McGill University, 1205 Docteur Penfield Avenue,
Montreal, Quebec H3A 1B1, Canada

Summary. We have analyzed the sequences of soybean leghemoglobin genes as an initial step toward understanding their mode of evolution. Alignment of the sequences of plant globin genes with those of animals reveals that (i) based on the proportion of nucleotide substitutions that have occurred at the first, second, and third codon positions, the time of divergence of plant and animal globin gene families appears to be extremely remote (between 900 million and 1.4 billion years ago, if one assumes constancy of evolutionary rate in both the plant and animal lineages) and (ii) in addition to the normal regulatory sequences on the 5′ end, an approximately 30-base-pair sequence, specific to globin genes, that surrounds the cap site is conserved between the plant and animal globin genes. Comparison of the leghemoglobin sequences with one another shows that (i) the relative amount of sequence divergence in various coding and noncoding regions is roughly similar to that found for animal globin genes and (ii) as in animal globin genes, the positions of insertions and deletions in the intervening sequences often coincide with the locations of direct repeats. Thus, the mode of evolution of the plant globin genes appears to resemble, in many ways, that of their animal counterparts. We contrast the overall intergenic organization of the plant globin genes with that of animal genes, and discuss the possibility of the concerted evolution of the leghemoglobin genes.

Key words: Leghemoglobin — Gene duplication — Gene linkage — Concerted evolution — Nitrogen fixation — Soybean

Introduction

A considerable amount of information has accumulated in recent years on the organization, expression, and evolution of animal globin gene families (Maniatis et al. 1980). Globin genes are present in plants as well (Baulcombe and Verma 1978; Sullivan et al. 1981). These genes encode leghemoglobins (Lbs), the monomeric hemoproteins that are found only in the root nodules of plants participating in symbiotic nitrogen fixation. Soybean, for instance, contains four major Lb proteins, Lb*a*, Lbc_1, Lbc_2, and Lbc_3, which differ slightly from each other in amino acid sequence (Hurrel and Leach 1977; Sievers et al. 1978; Fuchman and Appleby 1979). A number of closely related Lbs are found in other legume nodules as well. Comparison of the amino acid sequences of the plant and animal globins suggests that they have evolved from a common ancestor (Hunt et al. 1978).

Comparisons of animal globin gene sequences have revealed several interesting aspects of their behavior and evolution. These include (i) the identification of putative regulatory sequences (Efstratiadis et al. 1980), (ii) the occurrence of nonfunctional or pseudogenes in the various globin families (Efstratiadis et al. 1980; Maniatis et al. 1980), (iii) possible mechanisms for maintaining sequence homology among certain members of given families within a species (Slightom et al. 1980; Zimmer et al. 1980), (iv) the possibility that sequences within introns are acquired or lost through transposition events (Schon et al. 1981), and (v) the relative degrees of evolutionary constraint on the various nucleotides within the gene (Efstratiadis et al. 1980; Fitch 1980). The globin system is a model for gene evolution in eucaryotes in general.

Offprint requests to: G.G. Brown

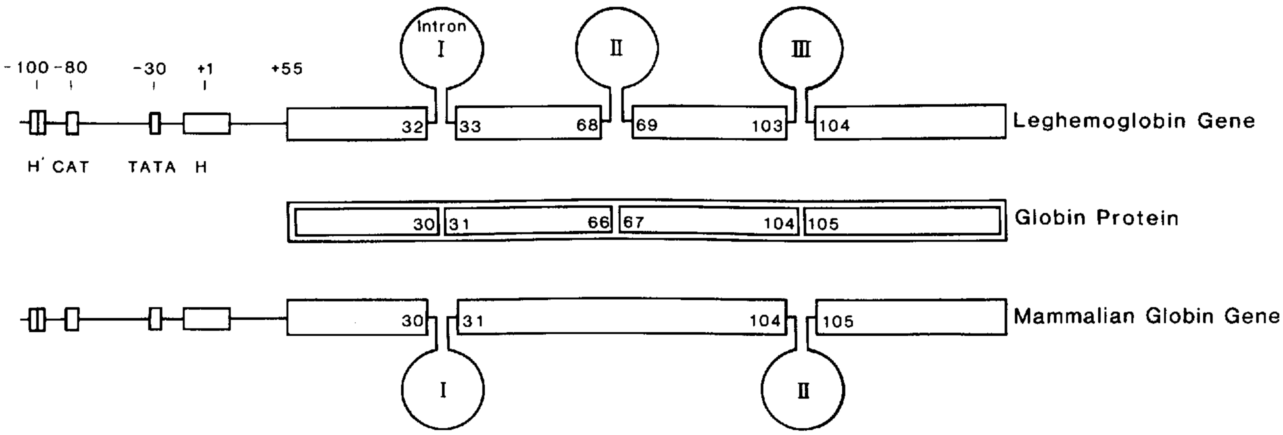

Fig. 1. Positions of the introns in Lbs and globins in relation to the globin structural units as determined by Gō (1981). Also indicated is the presence of two homologous sequences (H and H') in the 5' regions of these genes. For sequence of H, see Fig. 2, and for sequence of H', see Dierks et al. (1983)

The nucleotide sequences of four soybean Lb genes (Brisson and Verma 1982; Hyldig-Nielsen et al. 1982; Wiborg et al. 1982), a Lb pseudogene (Brisson and Verma 1982; Wiborg et al. 1983), and a truncated gene (Brisson and Verma 1982) are now known, and the arrangement of these genes in the soybean genome has been characterized (Lee et al. 1983). The intragenic organization of plant and animal globin genes was found to be very similar, the main difference being that the plant genes possess an extra intervening sequence. The ancestral relatedness of plant and animal globins therefore is reflected not only by their amino acid sequences but also by the overall structures of their genes. The evolutionary implications of the extra intron in the plant genes have already been discussed (Blake 1981; Brisson and Verma 1982; Hyldig-Nielsen et al. 1982).

The Lb system provides an opportunity to contrast the evolution of a gene family in plants with that of the homologous family in animals. With this purpose in mind we compare, in this paper, the sequences of the four soybean Lb genes and a soybean Lb pseudogene with each other and with animal globin genes.

Methods

Nucleotide sequences coding for Lb*a*, $-c_1$, $-c_2$, and $-c_3$ and those coding for the mouse α- and β-hemoglobins (Konkel et al. 1979; Nishioka and Leder 1979) were aligned codon for codon as indicated by the amino acid sequence alignments of Hunt et al. (1978). The number, per nucleotide site, of base substitutions that have occurred since the divergence of the various sequences (K values) and the standard deviations of those values were then estimated using Eqs. [6] and [12] of the "3ST" model of Kimura (1981). Only those nucleotide positions that were represented in all six sequences were considered. K values for the noncoding (5', intron, and 3') regions were calculated after these sequences had been fit to a "best" alignment with the aid of the SEQ homology computer algorithm of the Stanford SUMEX-AIM system. Again, only those nucleotide positions that were represented in all four Lb gene sequences were considered. All sequence alignments are given in the Appendix.

Sequences were searched for the presence of direct and inverted repeats using the homology and symmetry options, respectively, of the SEQ system.

Results

Plant vs. Animal Globin Genes

Leghemoglobin and animal globin nucleotide sequences were aligned to give the maximum homology between the genes. For the noncoding regions this was accomplished with the aid of a computer (see Methods). For the coding regions, we aligned the exon sequences of the animal genes with those of the plant genes according to the amino acid alignments of Hunt et al. (1978). We chose mouse α- and β-globin gene sequences as representatives of animal globins and compared them with the Lb*a*, Lbc_1, Lbc_2, and Lbc_3 gene sequences.

The intragenic organizations of the plant and animal globin genes are contrasted in Fig. 1. In addition to the two intervening sequences common to all animal globin genes, plant globin genes contain a third, central intervening sequence, which occurs at a position that separates the coding sequences for two different globin structural units (Gō 1981). When the sequences of the plant and animal globins are aligned for maximum homology (Hunt et al. 1978), the splicing points of the first and third intron in the Lbs (between codons 32 and 33 and between codons 103 and 104, respectively) coincide precisely with the positions of the two introns in animal globin genes.

Table 1. Base substitutions per nucleotide site (K values) and standard errors as estimated by the "3ST" model of Kimura (1981) for comparisons of the coding regions of plant and animal globin gene sequences[a]

	Mouse α-globin			Mouse β-globin		
	K_1	K_2	K_3	K_1	K_2	K_3
Lb*a*	1.71 ± 0.81	0.86 ± 0.15	—[b]	1.83 ± 0.63	0.80 ± 0.19	2.18 ± 0.77
Lb*c*$_1$	1.82 ± 0.51	0.86 ± 0.26	—[b]	1.73 ± 0.46	0.85 ± 0.17	1.89 ± 0.60
Lb*c*$_2$	1.63 ± 0.58	0.86 ± 0.17	—[b]	1.76 ± 0.52	0.83 ± 0.13	2.62 ± 2.05
Lb*c*$_3$	1.69 ± 0.63	0.81 ± 0.13	—[b]	2.06 ± 0.94	0.80 ± 0.16	1.75 ± 0.85
Mouse	—	—	—	0.67 ± 0.29	0.47 ± 0.08	1.01 ± 0.23

[a] K_1, K_2, and K_3 denote the estimated numbers of substitutions occurring at the first, second, and third codon positions, respectively. At each position 130 nucleotides were compared (n = 130)

[b] Undefined numbers were obtained

Table 1 lists the proportion of bases at the first, second, and third codon positions that have undergone substitution since the divergence of the plant and animal globin gene families and since the divergence of the mouse α- and β-globin genes. The values have been corrected for the occurrence of multiple substitutions and reversions by means of the "3ST" method of Kimura (1981). The sequence alignment employed is given in the Appendix. We estimate that since the plant–animal globin divergence, bases in the first position have undergone from 1.6 to 2.1 substitutions, while those in the second position have undergone from 0.8 to 0.9 substitutions. Values for the third position ranging from 1.7 to 2.6 were obtained in the Lb–mouse β comparisons. The extent of third-position change could not be estimated for the Lb–mouse α comparisons. Our values for each codon position in the mouse α–β comparison are similar to those given by Kimura (1981) for rabbit α- vs. β-globins (K_1 = 0.67 vs. 0.60; K_2 = 0.47 vs. 0.44; K_3 = 1.01 vs. 0.90).

If we assume a relative constancy of evolutionary rate at each codon position, these values can be used to estimate the time elapsed since the plant and animal globin genes diverged. Using the averages of Kimura's and our values for the α vs. β nucleotide divergences and the generally used value of 5×10^8 years for the α- and β-globin divergence time, we estimate evolutionary rates to be 6.35×10^{-10} and 4.55×10^{-10} substitutions/year for the first and second codon positions, respectively. From these values, we can then estimate that the Lbs and animal globins diverged between 900 million and 1.4 billion years ago, the lower figure being that obtained according to the second codon position value and the higher being that obtained according to the first. An intermediate value of 1.1 billion years is obtained using third codon position values.

No significant stretches of homology between the plant and animal globin genes could be identified in either the introns or the 3′ flanking regions, with the exception of the polyadenylation signal (Brisson and Verma 1982; Hyldig-Nielsen et al. 1982). In the 5′ noncoding region, however, several significant stretches of homology were observed. In addition to the presence of the general eucaryotic regulatory sequences (the TATA and CAT boxes), an approximately 30-bp homology (designated "H" in Fig. 1) exists in the region surrounding the cap site. The alignment of the consensus sequences for this region of ten β-like globin genes (Efstratiadis et al. 1980) and four Lb genes is shown in Fig. 2. A tetranucleotide that is complementary to the 3′ end of 18S rRNA and known to affect the rate of mRNA translation (Yamaguchi et al. 1982) is present in this region as well. It should be pointed out that the sequence identified in Fig. 2 appears to be specific to globin genes, since such high homology is not observed around the cap sites of nonglobin animal genes. In addition, a region near position −100, consisting of an imperfect tandemly repeated sequence (designated "H′" in Fig. 1) that is known to be essential for optimum promotor functions in rabbit β-globin genes (Dierks et al. 1983), was also found to be present in soybean Lb genes (see also Lee and Verma 1984).

Intergenic Comparison of Leghemoglobin Genes

The Lb gene sequences were aligned with one another by means of a computer algorithm. The alignments of the coding, 5′ noncoding, and 5′ and 3′ flanking regions were relatively straightforward. No gaps had to be inserted into the coding regions, although additional codons precede the termination codon in the Lb*c*$_2$ and Lb*c*$_3$ genes and in the pseudogene, ψLb$_1$. Only small gaps occurred in the alignments of the 5′ and 3′ noncoding and flanking regions. Large length differences were found in the introns, however. The alignments of homologous stretches in the intervening sequences of the four soybean Lb genes and the pseudogene ψLb$_1$ are schematically depicted in Fig. 3. The plant intervening sequences

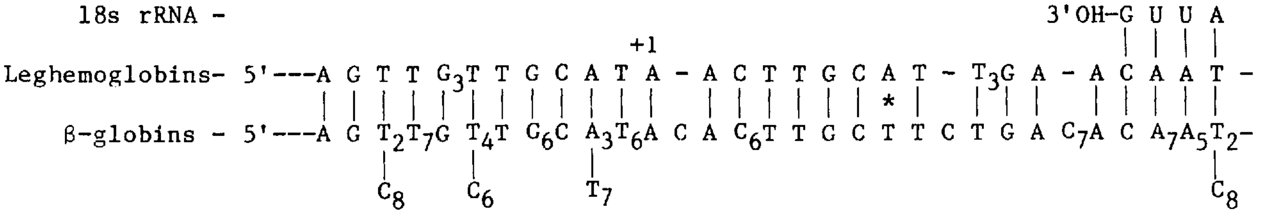
18s rRNA -
3'OH-G U U A
+1
Leghemoglobins- 5'---A G T T G3T T G C A T A - A C T T G C A T - T3G A - A C A A T -
*
β-globins - 5'---A G T2T7G T4T G6C A3T6A C A C6T T G C T T C T G A C7A C A7A5T2-
C8
C6
T7
C8

Fig. 2. Comparison of consensus sequences derived from the regions surrounding the putative cap sites of ten β-like globin genes (Efstratiadis et al. 1980) and four Lb genes. The subscripts indicate that less than nine globin genes and less than four functional Lb genes have the indicated nucleotide at that position. The first nucleotide of Lb mRNA is indicated as "ti".

exhibit more variation in size than do the intervening sequences of animal globin genes. Intervening sequence IVS-1 varies from 95 bp (ψLb$_1$) to 169 bp (Lb*c*$_1$) in length, in contrast to the 116- to 130-bp length variation seen among the mammalian β-globin genes (Efstratiadis et al. 1980), and IVS-3 varies in length from 197 bp (Lb*c*$_2$) to 778 bp (ψLb$_1$), as compared with the 572- to 904-bp mammalian size range.

The large length differences among the introns owe to the presence of relatively long insertions or deletions. For example, an approximately 1250-bp insertion found in IVS-2 of ψLb$_1$ is not found in any of the other three genes, and the IVS-3 sequences of the Lb*c*$_1$, Lb*c*$_2$, and Lb*c*$_3$ genes lack an approximately 400-bp stretch present in the Lb*a* and ψLb$_1$ genes. As for IVS-1, the only major length difference is due to the insertion of the 46-bp-long simple repeat $(AT)_n$ in the Lb*c*$_1$ gene. Other deletions and insertions ranging in size from 1 or 2 up to 130 bp are found throughout the introns.

Because direct repeats have been implicated in the generation of both deletions and insertions in animal globin genes (Efstratiadis et al. 1980; Schon et al. 1981; Spritz 1981), we searched the Lb intron sequences for the presence of such sequences. Their locations in the various Lb introns are shown in Fig. 3. Four classes of direct or near direct repeats are found in IVS-2. Classes *a* and *c* each have three members, *a, a'*, *a''* and *c, c'*, and *c''*, and *b* and *d* each have two, *b* and *b'* and *d* and *d'*, respectively. The *a* repeats are 8 bp long, the *b* repeats are 7 bp in length, the *d* repeats are 9 bp long, and the *c* repeats vary from 14 (*c''*) to 21 bp in length. Within classes, the repeats in the *a, b,* and *d* classes differ in only one base pair, with the exception of *a''*, which shows two differences from *a* and three differences from *a'*. The *c* and *c'* repeats differ in length only. An interesting feature of the *a* and *b* sequences is that they themselves can be aligned to form an inverted repeat 6 bp long. The *a* and *b* sequences are found in close proximity to one another at several places in the IVS-2 sequences of the Lb*a* and Lb*c*$_1$ genes, and by forming hairpin loops may produce important structural features in the pre-mRNAs.

The ends of nearly all insertions and deletions within the introns coincide with the locations of direct repeats, as has been observed in some animal globin genes (Efstratiadis et al. 1980; Schon et al. 1981). In IVS-2, for example, the large deletion in the Lb*c*$_3$ gene is flanked by the *a* sequence on its 5' side and a partial *a* sequence on its 3' side. Similarly, the 36-bp deletion in the Lb*c*$_1$ gene is flanked by a *b'* sequence and the insertion in IVS-2 of Lb*c*$_1$ is flanked by *d* and *d'* sequences.

The occurrence of deletions and insertions in IVS-3 is also associated with the presence of repeated sequences. The first additional stretch of sequence present in the Lb*a* but not the ψLb$_1$ sequence consists primarily of a direct repeat of the 11-bp sequence *e*. Similarly, additional stretches in ψLb$_1$, Lb*a* and Lb*c*$_3$ are flanked by the direct or near-direct repeats *f, g–g'*, and *h–h'*, repsectively. The remainder of the *f*-flanked insertion in ψLb$_1$ consists primarily of the simple sequence $(A)_n$.

Direct repeated sequences are found at the borders of both procaryotic (Calos and Miller 1980) and eucaryotic (Roeder and Fink 1980) transposons. The generation of insertions such as that found in IVS-2 of Lb*c*$_1$ may be accomplished by a mechanism analogous to the one proposed for insertion of the Ty elements in yeast (Roeder and Fink 1980). Similarly, the direct repeats may be involved in the generation of deletions by facilitating excision of integrated transposons (Roeder and Fink 1980) or by promoting strand slippage during DNA replication (Efstratiadis et al. 1980). Inverted repeat sequences such as are found in the Lb introns are a feature common to both procaryotic and eucaryotic transposons. This suggests that mobile genetic elements may play a role in effecting heterogeneity in introns. Schon et al. (1981) have postulated that similar events in goat globin evolution may have been mediated by transposons.

Sequence Divergence in the Leghemoglobin Genes

K values for different classes of nucleotide substitutions observed in the comparisons of the legitimate Lb genes are shown in Table 2. Within the

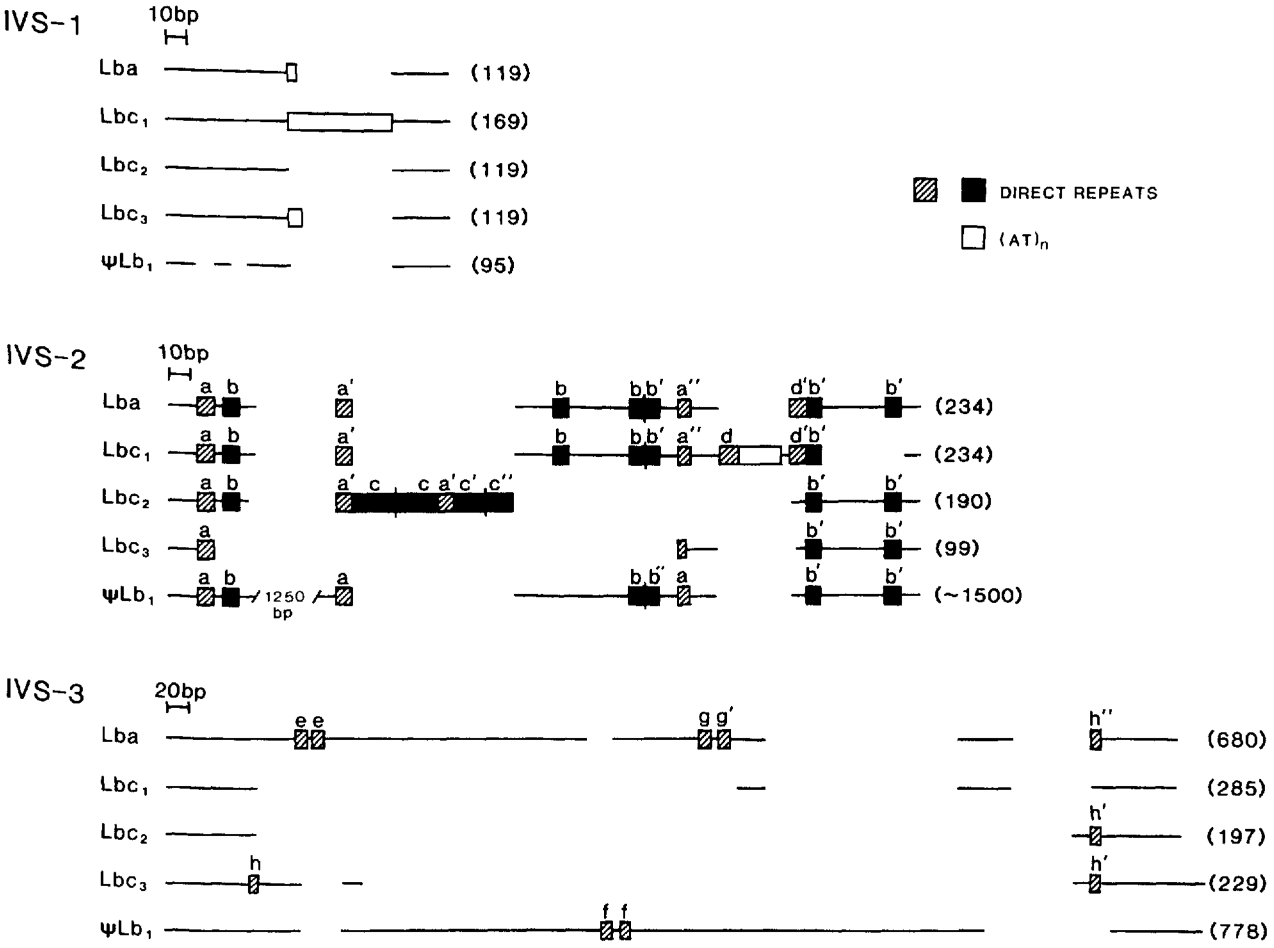

Fig. 3. Relationships among introns of various soybean Lbs. The lengths of the introns are indicated in parentheses. Homologous regions are indicated by lines, and boxes indicate the locations of various repeated sequences. Repeat *a* = TAAAATTA, *a'* = TAAGATTA, *a''* = TAAATTTC, *b* = TATTTTA, *b'* = TATTTTT, *c* = TTAAACATGTATTTAACACTC, *c'* = TTAAACATGTATTTAAC, *c''* = TAAAACATGTATTT, *d* = TGGTAATTA, *d'* = TGATAATTA, *e* = CAATCTTAAAA, *f* = TTGATTA, *g* AGTTCAATATATATTCATTT, *g'* = AGTACAATATATTTTCATTT, *h* = TTTCGTACT, *h'* = TTATGTACT, and *h''* = TTACGTACT. Vertical bars drawn through direct repeat regions indicate the boundaries of juxtaposed repeat units. Sequence homology is found between the *c*–*c* repeat in Lbc_2 and the regions surrounding and including the second b repeat of the Lb*a* and Lbc_1 genes, as well as the homologous stretch in the ψLb$_1$ gene (see Appendix). For the sake of clarity, the *c* repeat region of Lbc_2 was not aligned with the homologous regions of the other genes in the figure

coding regions, the values obtained at a given codon position do not vary greatly among the six comparisons. The first codon position appears to be slightly less variant than the second, although this difference is probably not statistically significant. Approximately 55% of the substitutions are silent, i.e., they do not lead to amino acid replacement (Table 3). This proportion of silent changes is very similar to that found in comparisons of animal globin genes (Jukes 1980). Since most silent changes occur at the third codon position, the third position is considerably more variable than the other two.

In the noncoding and flanking region and in the intervening sequences, the K values within a given category vary more among the six comparisons than in the coding region. This is probably due, at least in part, to the lower number of nucleotides compared in these cases. Because of this intracategory variation, it is difficult to assess the relative amounts of variation occurring in these regions. In general, however, the variation occurring in all of the noncoding categories appears to be comparable to that of the third position in the coding region, with the 5' noncoding/flanking region possibly being slightly more conserved. The relative frequencies of substitutions among synonymous (silent), nonsynonymous (replacement), and intron sites observed for the goat and sheep β-globin genes by Li and Gojobori (1983) are consistent with our findings. This suggests that the modes of base substitution in plant and animal globin genes are similar.

Recently, considerable attention has been drawn to the fact that in animal mitochondrial DNA (mtDNA), a very high proportion of the total num-

Table 2. Base substitution values, estimated as in Table 1, for various regions of the plant globin genes[a]

Com-	Coding regions (n = 130)			Noncoding/flanking		Intervening sequences		
parison	K_1	K_2	K_3	5′ (n = 103)	3′ (n = 123)	IVS-1 (n = 111)	IVS-2 (n = 41)	IVS-3 (n = 141)
Lba/c_1	0.023 ± 0.013	0.048 ± 0.020	0.11 ± 0.03	0.13 ± 0.04	0.15 ± 0.04	0.05 ± 0.02	0.025 ± 0.012	0.12 ± 0.03
Lba/c_2	0.032 ± 0.016	0.031 ± 0.016	0.10 ± 0.03	0.10 ± 0.03	0.12 ± 0.03	0.11 ± 0.03	0.052 ± 0.019	0.11 ± 0.03
Lba/c_3	0.048 ± 0.020	0.040 ± 0.018	0.11 ± 0.03	0.08 ± 0.03	0.12 ± 0.03	0.12 ± 0.03	0.110 ± 0.050	0.10 ± 0.03
Lbc_1/c_2	0.023 ± 0.013	0.048 ± 0.019	0.14 ± 0.03	0.04 ± 0.02	0.11 ± 0.03	0.12 ± 0.03	0.025 ± 0.012	0.10 ± 0.03
Lbc_1/c_3	0.032 ± 0.016	0.073 ± 0.025	0.15 ± 0.04	0.07 ± 0.03	0.08 ± 0.03	0.13 ± 0.04	0.078 ± 0.040	0.17 ± 0.04
Lbc_2/c_3	0.023 ± 0.013	0.040 ± 0.018	0.15 ± 0.04	0.07 ± 0.03	0.05 ± 0.02	0.21 ± 0.05	0.110 ± 0.050	0.14 ± 0.03

[a] The number of bases used for each comparison is given in parentheses

Table 3. Nucleotide changes in the coding regions of a Lb pseudogene and true Lb genes

Gene pair	Silent changes	Replacement changes	Silent changes/ replacement changes
Lba/Lbc_1	10	11	0.91
Lba/Lbc_2	11	9	1.22
Lba/Lbc_3	11	12	0.92
Lbc_1/Lbc_2	15	9	1.67
Lbc_1/Lbc_3	14	13	1.08
Lbc_2/Lbc_3	16	9	1.78
Lba/ψLb$_1$	21	33	0.64
Lbc_1/ψLb$_1$	27	29	0.93
Lbc_2/ψLb$_1$	25	31	0.81
Lbc_3/ψLb$_1$	26	33	0.79

ber of base substitutions are transitions (Brown and Simpson 1982; Brown et al. 1982; Aquadro and Greenberg 1983). This increased frequency of transitions can have a significant effect on methods of calculating sequence divergence (Holmquist 1983; Gojobori 1983). We find a significant bias toward transitions in the base substitutions occurring in the Lb genes, which therefore behave in this respect also like animal hemoglobin genes (Derancourt et al. 1967; Li and Gojobori 1983). In the coding regions, for example, transitions outnumber transversions by a ratio of 1.6:1. This bias is significant, since the opportunity for transversions to occur is twice as great as that for transitions. It is not as extreme as that for animal mtDNA, however, in which the transition/transversion ratio ranges from 8:1 to 32:1 (Brown and Simpson 1982; Aquadro and Greenberg 1983). A similar observation regarding the relative frequencies of transition substitutions in animal globin vs. animal mitochondrial genes has recently been made by Li and Gojobori (1983).

A soybean genomic sequence that possesses a high degree of homology with Lb cDNA but does not code for any of the known Lb proteins has previously been identified (Brisson and Verma 1982; Wiborg et al. 1983). Because this sequence does not appear to be expressed, it has been tentatively designated a pseudogene, ψLb$_1$. However, no structural features that would prevent its expression (i.e., in-frame termination codons, splice junctions lacking the consensus sequence, etc.) have been identified. This gene is linked to normal Lb genes by spacers of about 2.5 kb.

When the sequences of the pseudogenes found in animal gene families are compared with those of their functional counterparts, the fraction of replacement substitutions is generally found to be higher than that observed in comparisons among the functional genes (Miyata and Hayashida 1981). The numbers of silent and replacement substitutions found in comparisons of the last two exons of various Lb genes are given in Table 3. The gene–pseudogene comparisons show a highly significant elevation in the proportion of replacement substitutions. In comparisons between legitimate Lb genes, silent substitutions exceed replacement substitutions by a factor of 1.26 on the average. However, the silent/replacement substitution ratio drops in the gene–pseudogene comparisons to an average value of 0.79. This provides some additional evidence that the ψLb$_1$ sequence is, in fact, that of a pseudogene.

Discussion

In general, the evolution of Lb genes appears to be quite similar to that of their counterparts in animals. The relative frequencies of occurrence of base substitutions in different coding and noncoding regions and at the various codon positions within the coding regions are comparable to those found for animal globins. This suggests that the relative degrees of functional constraint to which these various regions are subjected are the same in plants and animals.

We have used the divergence values between the mouse α- and β-globins and the Lb to estimate the plant–animal globin gene divergence time as 900 million to 1.4 billion years ago. This estimate is based on the assumptions that the α- and β-globin gene families diverged approximately 500 million years ago and that the globin genes have evolved at a constant rate in both the plant and animal lineages. Since the rate of evolution of animal globin genes appears to be subject to some fluctuation (Czelusniak et al. 1982; Li and Gojobori 1983) and since errors are involved in both the estimation of the α–β divergence time from the fossil record and the estimation of the number of base substitutions, it is possible that the actual divergence time differs significantly from the one we have calculated.

It has recently been suggested that the Lbs arose as a result of a horizontal transfer of a globin gene from an animal to an ancestral legume plant (see Lewin 1981). The very ancient divergence time for the plant and animal globin genes obtained from our analysis does not support this hypothesis. Cytochrome *c* amino acid analyses (Brown et al. 1972) as well as the fossil record (Valentine 1973) indicate that the major metazoan radiations took place between 700 and 800 million years ago. Hunt et al. (1978) suggest that the most ancient animal globin gene duplication occurred also at about this time. It is therefore unlikely that the globin genes of the putative animal donor would have diverged prior to this time from those of the lineage that gave rise to the vertebrates. Our divergence figure is more consistent with the alternative hypothesis that globin genes were present in the common ancestor of present-day plants and animals. However, the qualifications placed on the divergence calculations as well as the possibility that globin genes may be evolving more rapidly in plants (Lee et al. 1983) make the elucidation of evolutionary relationships between plant and animal globin genes difficult.

The relatively larger fluctuations in the lengths of the introns of the plant globin genes constitute one major difference between their mode of evolution and that of animal globins. The reason for this difference is unclear. It may be that insertion or deletion events, possibly due to transpositions, occur more frequently in the plant genome, or that the constraints on intron length are lesser in plants. Since the sites of deletions and insertions often coincide with the locations of repeated sequences in both plants and animals, it seems likely that the mechanisms that give rise to this type of variation are similar in both kingdoms.

The preservation of a globin-specific sequence in the 5′ flanking region is perhaps the most striking feature of the comparison of plant and animal globin genes. The significance of the conservation of this particular sequence is unclear. If the region serve a regulatory role, then it is possible that some of the regulatory mechanisms that govern the expression of animal globin genes operate in plants as well. Alternatively, it is possible that the conservation of this sequence is simply fortuitous, although we deem this unlikely, particularly in light of the fact that another sequence, at position −100, that is known to be involved in β-globin transcription (Dierks et al. 1983) is also found in Lb genes.

A question of primary importance regarding the evolution of the Lb genes of soybean concerns their mode of duplication. As dicussed by Jeffreys and Harris (1982), gene families appear to be organized into two basic patterns—as sequences in linked clusters and as sequences dispersed over widely separated chromosomal locations. As with animal globin genes, both types of organization are observed in the Lb genes. One chromosomal locus contains three Lb genes, *a*, c_1, and c_3, and a complete pseudogene, ψLb$_1$ (Fig. 4; see also Lee et al. 1983). Lbc_2 is found at a different chromosomal location, where it is linked to another pseudogene, ψLb$_2$ (J. Lee and D.P.S. Verma, unpublished observations). Truncated Lb pseudogenes, LbT$_1$ and LbT$_2$, consisting of only the fourth exon and related flanking sequences, are found at at least two other chromosomal locations. All four chromosomal sites are bounded at their 3′ ends by the dispersed repeat element labeled *s* in Fig. 4. The Lb*a* and Lbc_2 loci contain two types of this repeat, *s* and *s*′. Furthermore, the main locus is flanked by two sequences that are expressed more abundantly in root and leaf tissues.

The means by which this particular organization was achieved is uncertain. The main locus was undoubtedly produced by a series of duplications, possibly resulting from unequal crossovers occurring during meiosis. We were unsuccessful, however, in constructing a tree based on Lb sequence comparisons that gave a good indication of the order in which these duplications took place. Trees constructed using different methods gave different branching arrangements, and in no case was any one arrangement significantly better than any of the others.

One possible reason for this may be that the Lb genes are subject to concerted evolution. This view is suggested by the fact that the Lb proteins occurring within the soybean species are all more closely related to one another than to the Lb proteins of different species. Thus homogenization of the sequences at the main locus, again possibly through unequal crossovers (Zimmer et al. 1980) would obscure the ancestral relatedness of the different genes. Zimmer et al. (1980) cite the occurrence within the same population of variants of the human α-globin locus that possess greater and fewer than the normal number of genes as evidence that concerted evolution of globin genes takes place by this mechanism.

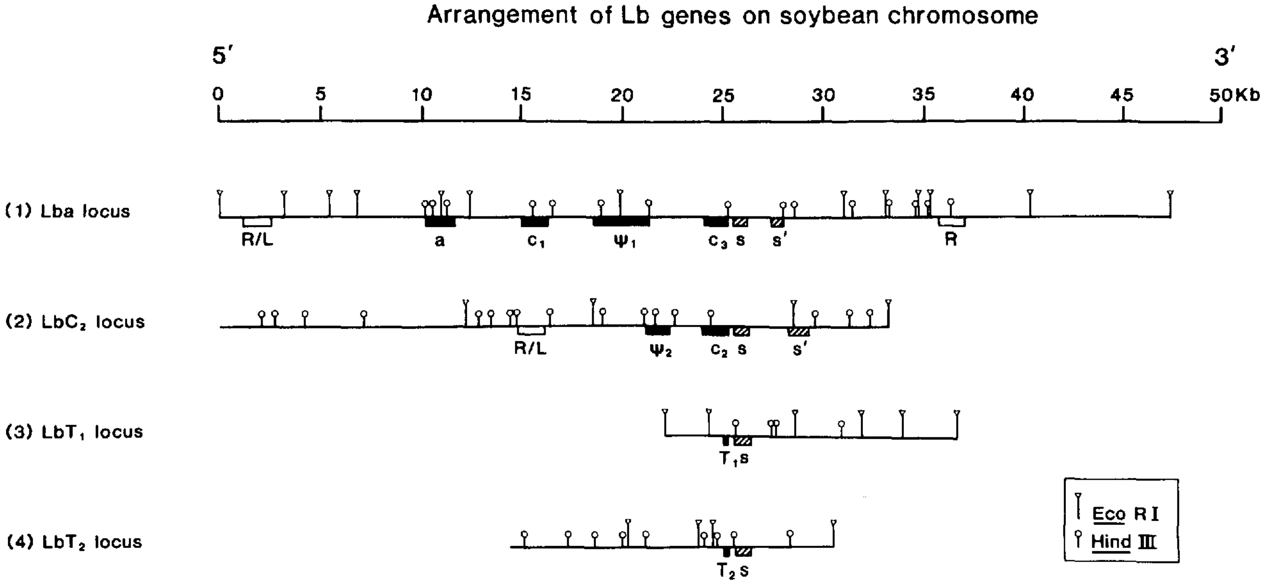

Fig. 4. Chromosomal arrangement of Lb genes in soybean. Note the presence of at least four loci, two of which contain truncated Lb sequences. *R* and *R/L* are sequences expressed in root and root/leaf, respectively; *s* and *s′* are two repeat elements (see Lee et al. 1983 for further details)

This may also be true for Lb genes. Alternatively, it is possible that a concerted evolution could proceed through gene conversion events, as suggested by Slightom et al. (1980).

The mechanism by which the additional Lb loci arose is unclear. One intriguing possibility is that these loci appeared initially as a result of tetraploidization events (Hadley and Hymowitz 1973) that took place during the recent evolutionary history of the soybean. This view is supported by the finding that the overall arrangement is very similar at the 3′ ends of all three loci (Fig. 4). The deletion of Lb genes subsequent to such chromosomal duplications would thus give rise to the structure of this gene family as it is observed in present-day legumes. Gene duplication through tetraploidization has been invoked by Jeffreys et al. (1980) to explain the organization of α- and β-globin genes in *Xenopus.*

The temporal sequence of induction at animal globin loci is in the 5′ to 3′ direction. In the case of Lbs, the Lbc_3 gene, which is located on the 3′ end of the main locus, appears to be induced before Lb*a* (Verma et al. 1979), which is located on the 5′ end. The reason for this difference between plants and animals is not apparent. Furthermore, although the animal and plant globins have regions of significant homology at their 5′ ends, they are induced under very different sets of conditions, suggesting that many other sequences are involved in regulation of the expression of these genes in vivo.

Acknowledgments. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and FCAC, Quebec. N.B. acknowledges the receipt of a postgraduate fellowship from NSERC. We wish to thank Dr. Forrest Fuller for his assistance in computer analysis and Miss Yvette Mark and Mrs. Sylvie Gauthier for their assistance in preparing the manuscript.

References

Aquadro CF, Greenberg BD (1983) Human mitochrondrial DNA variation and evolution: analysis of nucleotide sequences from several individuals. Genetics 103:287–312

Baulcombe D, Verma DPS (1978) Preparation of a complementary DNA for leghaemoglobin and direct demonstration that leghaemoglobin is encoded by the soybean genome. Nucleic Acids Res 5:4141–4153

Blake CCF (1981) Exons and the structure, function and evolution of haemoglobin. Nature 291:616

Brisson N, Verma DPS (1982) Soybean leghemoglobin gene family: normal, pseudo and truncated genes. Proc Natl Acad Sci USA 79:4055–4059

Brown GG, Simpson MV (1982) Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. Proc Natl Acad Sci USA 79:3246–3250

Brown RH, Richardson M, Boulter D, Ramshaw JAM, Jeffries RPS (1972) The amino acid sequences of cytochrome *c* from *Helix aspersa* Müller (garden snail). Biochem J 128:971–974

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Calos MP, Miller JH (1980) Transposable elements. Cell 20: 579–595

Czelusniak J, Goodman M, Hewett-Emmett ML, Weiss ML, Venta PJ, Tashian RE (1982) Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. Nature 298:297–300

Derancourt J, Lebor A, Zuckerkandl E (1967) Séquence des acides aminés, séquence des nucleotides et évolution. Bull Soc Chim Biol 49:557–591

Dierks P, van Doyen A, Cochran MD, Dobkin D, Reiser J, Weissmann C (1983) Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit β-globin genes in mouse 3T6 cells. Cell 32:695–706

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JD, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ (1980) The structure and evolution of the human β-globin gene family. Cell 21:653–668

Fitch WM (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three β hemoglobin messenger RNAs. J Mol Evol 16:153–209

Fuchman WH, Appleby CA (1979) Separation and determination of the relative concentrations of the homogeneous components of soybean leghemoglobin by isoelectric focusing. Biochim Biophys Acta 579:314–324

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

Gojobori T (1983) Codon substitution and the "saturation" of synonymous changes. Genetics 105:1011–1027

Hadley HH, Hymowitz T (1973) In: Caldwell BE (ed) Soybeans: improvement production and uses. American Society of Agronomy. Madison, Wisconsin, pp 97–114

Holmquist R (1983) Transitions and transversions in evolutionary descent: an approach to understanding. J Mol Evol 19:134–144

Hunt LT, Hurst-Calderone S, Dayhoff MO (1978) Globins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Silver Spring, Maryland, pp 229–251

Hurrel JGR, Leach SJ (1977) The amino acid sequence of soybean leghemoglobin C_2. FEBS Lett 80:23–26

Hyldig-Nielsen J, Jensen EO, Paludan K, Wiborg O, Garret R, Jorgensen P, and Marker KA (1982) The primary structure of two leghemoglobin genes from soybean. Nucleic Acids Res 10:689–701

Jeffreys AJ, Harris S (1982) Processes of gene duplication. Nature 296:9–10

Jeffreys AJ, Wilson V, Wood D, Simons JP, Kay RM, Williams JG (1980) Linkage of adult α and β-globin genes in *X. laevis* and gene duplication by tetraploidization. Cell 21:555–564

Jukes TH (1980) Silent nucleotide substitutions and the molecular evolutionary clock. Science 210:973–978

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Konkel DA, Maizel JV, Leder P (1979) The evolution and sequence comparison of two recently diverged mouse chromosomal β-globin genes. Cell 18:865–873

Lee JS, Brown GG, Verma DPS (1983) Chromosomal arrangement of leghemoglobin genes in soybean. Nucleic Acids Res 11:5541–5553

Lewin R (1981) Evolutionary history written in globin genes. Science 214:426

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

Maniatis T, Fritsh EF, Lauer J, Lawn RM (1980) The molecular genetics of human hemoglobins. Annu Rev Genet 14:145–178

Miyata T, Hayashida H (1981) Extraordinarily high evolutionary rate of pseudogenes. Evidence for the presence of selective pressure against changes between synonymous codons. Proc Natl Acad Sci USA 78:5739–5743

Nishioka Y, Leder P (1979) The complete sequence of a chromosomal mouse α globin gene reveals elements conserved throughout vertebrate evolution. Cell 18:875–882

Roeder GS, Fink GR (1980) DNA rearrangements associated with a transposable element in yeast. Cell 21:239–249

Schon EA, Cleary ML, Haynes JR, Lingrel JB (1981) Structure and evolution of goat γ-, β^C- and β^A-globin genes: three developmentally regulated genes contain inserted elements. Cell 27:354–369

Sievers G, Huhtala ML, Ellfolk N (1978) The primary structure of soybean (*Glycine max*) leghemoglobin *C*. Acta Chem Scand [B] 32:380–386

Slightom JL, Blechl AE, Smithies O (1980) Human fetal G and A-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638

Spritz RA (1981) Duplication deletion polymorphism 5′ to the human β globin gene. Nucleic Acids Res 9:5037–5047

Sullivan D, Brisson N, Goodchild B, Verma DPS, Thomas DY (1981) Molecular cloning and organization of two leghemoglobin genomic sequences of soybean. Nature 289:516–518

Valentine JW (1973) Coelomate superphyla. Syst Zool 22:97–102

Verma DPS, Ball S, Guérin C, Wanamaker L (1979) Leghemoglobin biosynthesis in soybean root nodules. Characterization of the nascent and released peptides and the relative rate of synthesis of the major leghemoglobins. Biochemistry 18:476–483

Wiborg O, Hyldig-Nielsen JJ, Jensen EO, Paludan K, Marker KA (1982) The nucleotide sequences of two leghemoglobin genes from soybean. Nucleic Acids Res 10:3487–3494

Wiborg O, Hyldig-Nielsen JJ, Jensen EO, Paludan K, Marker KA (1983) The structure of an unusual leghemoglobin gene from soybean. EMBO 2:449–452

Yamaguchi K, Hidaka S, Miura KI (1982) Relationship between structure of the 5′ non-coding region of viral mRNA and efficiency in the initiation of protein synthesis in an eukaryotic system. Proc Natl Acad Sci USA 79:1012–1016

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC (1980) Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc Natl Acad Sci USA 77:2158–2162

Received August 17, 1983/Revised May 24, 1984

Appendix: Alignment of Globin Gene Sequences (Figs A-1-A-5)

Coding sequences

```
mouse α   GTG   CTCTCTGGGGAAGACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGCCATGGTGCTGAATATGGAGCTGAAGCCCTGGAAAGGATGT
mouse β   ---CAC--GA---AT-CT--G--GGCTGCTG--TCTTGCCTG-----A---G-G      A-CTCC-A----GT---T-G---G-------GC---C--C
Lba       GTTGCTTTCACTGAGAAGCAAGATGCTTTGGTGAGTAGCTCATTCGAAGCATTCAAGGCAAACATTCCTCAATACAGCGTTGTGTTCTACACTTCGATAC
Lbc1      -G---------------------G-------------------------------------------------------------------A------T-
Lbc2      -G---------------------G----------------------------------------------------------------------------
Lbc3      -G------------T--------G--------------------T-----------A--------------------T--------------C-------
ΨLb1      -G------T--A----------------------A------G--T------------------C--------CC---------A----T--A-------T

mouse α   TTGCTAGCTTCCCCACCACCAAGACCTACTTTCCTCACTTT   GATGTAAGCCAC               GGCTCTGCCCAGGTCAAGGGTCACGGCAA
mouse β   -G-T-GT--A---TTGG--GC--CGG------GA-AG----GGA--CC--TC-TCTGCCTCTGCTTCTATG--TAA----A-A--G----CC--T-----
Lba       TGGAGAAAGCACCTGCAGCAAAGGACTTGTTCTCATTTCTA         GCAAATGGAGTAGACCCC   ACTAATCCTAAGCTCACGGGCCATGCTGA
Lbc1      -----------------------------------------         ------------------   -----------------------------
Lbc2      -------------C--------------------G------         T-T-----------T--T   -G---------------------------
Lbc3      ---------------T-------------------------         ------------------   -----------------------------
ΨLb1      -------------A---------A--A-----------T--         --TG---C------T--G   -A-------------G-------------

mouse α   GAAGGTCGCCGATGCGCTGGCCAGTGCTGCAGGCCACCTCGATGAC               CTGCCCGGTGCCTTGTCTGCTCTGAGCGACCTGCATGCC
mouse β   ------GATAAC---CT-TAA-GA--GCCTGA-T---T-G--CAG-               --CAAG--CA----TG-CAGC--C--T--G--C--CTGT
Lba       AAAGCTTTTTGCATTGGTGCGTGACTCAGCTGGTCAACTTAAAGCAAGTGGAACAGTGGTGGCTGATGCCGCACTT         GGTTCTGTTCATGCC
Lbc1      -------------------------------------------A---A---------------------T------         -T----A-C------
Lbc2      -----------G-----------------------------------A---------A------------------         ------A-C------
Lbc3      ---A-------G------A-----T------------------------------------AT-------------         ------A-C------
ΨLb1      -----------G--------------------T-------C------AG---TT----------------A-----         ---C-------CA--

mouse α   CACAAG      CTGCGTGTGGATCCCGTCAACTTCAAGCTCCTGAGCCACTGCCTGCTGGTGACCTTGGCTAGCCACCACCCTGCCGATTTCACCCCCG
mouse β   G-----      ----A---------T-AG-------G-------G--A TATGA-CG--A-TGTGC---GCCA-----TTGGCAAG-------------
Lba       CAAAAAGCAGTCACT      GATCCT   CAGTTCCTGGTGGTTAAAGAAGCACTGCTGAAAACAATAAAGGCAGCAGTTGGGGACAAATGGAGTGACG
Lbc1      ---------------      ------   --A----------------------------------------AC--T-----C-G---T----------
Lbc2      ---------A-----      ------   --------------------------------------------G-----------------------T-
Lbc3      ---------A-----      ------   --A--T-------------------------------------AG-------------------------
ΨLb1      -------G---T--C      ------   -----T-C---------------------T-------------A---------------------C--A-

mouse α   CGGTACATGCCTCTCTGGACAAATTCCTTGCCTCTGTGAGCACCGTGCTGACCTCCAAGTACCGT
mouse β   -T-C---G--TG-CT-CC-G--GG-GG-G--TGGA---GC---T-CCT--G-TCA--------AC
Lba       AGTTGAGCCGTGCTTGGGAAGTAGCCTACGATGAATTGGCAGCAGCTATTAAGAAGGCA
Lbc1      -A------A-------------------T-----------------A-----A------
Lbc2      -A------A-------------------T------------------------------TTT
Lbc3      --------A-------------------T------------------------------TTT
ΨLb1      -A------AA-C-----------------------A-T-----C---------------ATGGCTATAGGATCATTAGTA
```

The coding sequence of the mouse α-globin and the Lb*a* genes are given in full and aligned over their complete lengths according to the amino acid alignments of Hunt et al. (1978). Nucleotides in the mouse β-globin sequence that differ from those in the mouse α-globin sequence and nucleotides in the Lb*c*$_1$, *c*$_2$, or c$_3$ sequences that differ from those in Lb*a* are listed below the corresponding bases in the mouse α-globin and the Lb*a* sequences, respectively. Dashes (–) indicate that the corresponding nucleotide is the same as that in the mouse α-globin or Lb*a* gene. For the flanking/noncoding and intervening sequences, the Lb*a* sequence is given in full and base differences in the other Lb genes are listed below it. The gaps represent the sites of deletions or insertions. In the intron sequences, the stretches of $(AT)_n$ are indicated by bars both above and below the sequence. The direct or near-direct repeats indicated in Fig. 3 are shown by bars above the corresponding sequence. The repeat sequences are given in full for each gene.

5' non-coding / flanking

```
Lba    AAGCT TT   GGTT TTCT   CACTCTCCAAGCCCTCTATATAAACAAATATTGGA GTGAAGTTGTTGCATAACTTGGCATCGAACAAT      TAA

Lbc1   ---T- G-AAA  C- ----   --  ---- ----T---------  --TG-------T---------A---------------T-------AGAAAA---

Lbc2   ---T- G-AAT  -- ----AA -    --------T-------------CG-------T-------------------------T-------AGAAA ---

Lbc3   ---T- --ATTAG--A----GAT------T------T-----------T--G-------T-------------------------T-------      ---

ΨLb1   --CAAA--        --T-   --G-------A--A--------------G-------T-----CT--     ----- -G--T---
```

```
Lba    TA GAAATAA     CAGAAAATTAAAAAA    GAAAT

Lbc1   C-A ---A--    GT-A----G--G----    -----

Lbc2   C--C---G--      - -T--G- G----AA  -----

Lbc3   -- -------     -------G--G----    -----

ΨLb1   -- -----T-AATAA----------C----GATC-----
```

3' non-coding

```
Lba    TAATTAGTATCTA     TTGCAGTAAAGT      GTAATAAATAAATCTTGTTTCA CTATAAAACTTGTTACTATTAGACAAGGGCCTGATACAAAA

Lbc1   -------G-----CTGCA----C-------      ------------- --------A--  ----------   ----A-----TT--CT---T- -

Lbc2       ---G-----CTA  ----C--C----      -------------T--------  --  ----------   ----A-----TC--C----T-T--

Lbc3       ---G-----C  AA----CT------      -------------A--A-----  --  ----------   ----A-----TT-TC----T- -

ΨLb1                         ------CTAGTA-------------T--------  --  ----T-----   --- A--   TT-------T- -
```

```
Lba    TGTTGGTTAAAATAA TGGAATTA  TA TAGT ATTGGATAAA AATCTTA

Lbc1   -----T--------- GTA----T  C- -T-- ----------C-C-T---

Lbc2  A--------------- GTA-----  -- CG-- --- ------C-------

Lbc3   --------------- GTA-----  -- -G-- ----------C-------

ΨLb1   ---------C- -T-GGTA-----CAG--C-T--C----A----C---A---
```

IVS-1

```
Lba     TAAGTTTTCTCTCTAA GCATGTGTCTTCCATTCTATGTTTTTC TTTTGGA AATTTGTTGTGTTTGAAAAAAGATATA

Lbc1    ------------A--- -----------T--------------- --C---- -----T-----------------TATATATATATATATATATATATA

Lbc2    ------------TA-- ------A----T------C--------C--- C--C- ---T-------------G---TA

Lbc3    -----A----A-----ATT---------  ---G-------AA- -- -C-T GG---------------------TATATA

ΨLb1    ---T----------C-              C--     ------ -C-C--                 -A- -------G--GTA


Lba                                              TTGT   TAATGTGAGTGG T TTTGGTTTGATTAAAAA       TGAATAGG

Lbc1    TATATATATATATATATATATATATATATATTT----   ------------ - -----------------       -A------

Lbc2                                             G---   C-----------G-A---TT---T--------       -T--C---

Lbc3                                             ----   ------------ - ----------C------       --------

ΨLb1                                             G---ACA----   ----T C ----------------ACAAAA- ------
```

IVS-2

```
                           a           b                                    a'                 b
Lba     GTAAGTATCACCCAACTAAAATTATAACTATTTTATGTGATT                              AATTTTAAGATTAAGCAT-CATGTATTTTAACACTCTTAAAACA
                           a           b                                    a'                 b
Lbc1    ----------G-----TAAAATTA----TATTTTA-------                              -----TAAGATTA-A--- ----TATTTTA------------T-
                           a           b
Lbc2    ----------T-----TAAAATTA--G-TATTTTA-------                                  a'               c
                           a                                                    -----TAAGATTAAACAT    GTA TTTAACACTCTTAAA CA
Lbc3    -------CT-G--T--TAAAATTA
                           a           b                (approx. 1250               a
ΨLb1    ------G-T-GT----TAAAATTA--GTTATTTTA-------TTAGGG  nucleotides)          TTAAAATTA-A---G--CT-- -A---T----------T-


                                                                                             b      b'
Lba     T                                                    CAATGAACATTAATTGTTTGAATTGTATTTTATATTTTTGCCATATCTTGAACTA
                                                                                             b      b'
Lbc1    -                                                    ---------------T---------TATTTTATATTTTTA---------------
              c            a'       c'              c"
Lbc2    TGTATTTAACACTCTTAAGATTAAACATGTATTTAACTAAAACATGTATTTGC

Lbc3
                                                                                             b      b"
ΨLb1    -                                                    -----------  ----A -----TATTTTATATCTTTAA--------C-T----


                           a"                                                       d      b'
Lba     GG    AATAGTATATAAATTTCTATTAGTATTTGT                                        TGATAATTATTTTTCTTTCATAACTATCTTGTCACA
                           a"                d                                      d'     b'
Lbc1    --    ----A----TAAATTTC------------TGGTAATTACATATATATATATATATATATAATCCTTGTGATAATTATTTTTC
                                                                                           b'
Lbc2                                                                                --    --TATTTTTT---T----T------------
                                                                                           b'
Lbc3                     AAATCC-----------T-T                                            --TATTTTT---C---G-T-G----------
                          a
ΨLb1    --GATT----A-G--TAAAATTA------------T-G                                        --T-GT TTT   ---TA--G-TC------A---T-


                    b'
Lba     TATTAT ATA TTTTTTC AATT GTAG

Lbc1                       G----T----
                   b'
Lbc2    ------ -TATTTTT-  G---- ----
                   b'
Lbc3    ------ -TA TTTTT- G----A ---
                   b'
ΨLb1    ------G-TA  TTTTT- --A-T----
```

IVS-3

```
Lba     GTATGATAAATAA      TGAAATGTTATAATAAATTATGCA    TACTTCAATTTT     TCATGGAGCAGT A TA ATGA  TCAA CACACACTT
Lbc1    -------------TACTAG-A---------C-------- ---    A----A-G----ACGTA---A-T-      - -C ----CT---TG--
Lbc2    -------------      -A--------C-------   ---CATA-----A------     A-----T-----GT--TG--C-  ---
Lbc3    -------------      -----A-C--C-------   ---CAAA----- ------     A---A-T-----GC --T----  ---T ---
ΨLb1    -------------         -T-------------   ---    -----A---C-A     A-----T-T-T-CT GTA----  ---T     -----

                                       _____e______   ______e_______
Lba     CTTTTGT  TT   CATGCATTTGATAACTACAATCTTAAAATGTTGCAATCTTAAAAATAGTATTAAAAATATA  ACATTTAATTAGCTCATCAATATTT
Lbc1
Lbc2    T------  --
           ___h___
Lbc3    ---CGTACTAAGTA---A----                                          -C-T-TTT-T-TT---G
ΨLb1    -------  --AGTA---A----                                         -C----C----- ---------   G--C--------A

Lba     TTCTGTTGCAATTTTTTATGAAAAAATT ATAATTATGAATTCTTTGAGCAATGTTTAATTAAAAAATTGATTTAATAATGAAATAACTAAGCTACCTCTG
Lbc1
Lbc2
Lbc3
ΨLb1    -AT-A---T-------CA----   TT--TT------AT--------------------  AG---  ---------C-C--C-------- ---T-----T

Lba     TCTC GTTTTTCATTTAAACTATGACATAAACAATG AA TAAAGTAAACTAAACCATGACAT GTTTATTTTTGAATGAGGTTATTAATA   ATTTTTTT
Lbc1
Lbc2
Lbc3
ΨLb1    ----AA----- --A----T---C-----T--T---G--A----A---G------T---TG--TA----------G--T------------GC----- A-

Lba     TCACTAT   CTATTGCAATG                            TTCATTGATTTATC AATTATCTT G GTT   GCATTGATTCTC     TC
Lbc1
Lbc2
Lbc3
                              ___f___                          ___f____
ΨLb1    AAT----TGG----------C-TTGATTAGATTCTCAAAAAAAAAAAAACT--G-TTGATTAA-T -------A- TTC--TTT------------GCCTG--

                                   __________g__________                __________g__________
Lba     GATTTTTTTCTT GAGGTTAA GCTTCAGTTCAATATATATTCATTTTTTGATAAAAAAAAATAGTACAATATATTTTCATTTAGCTGATCATATTTATTT
Lbc1                                                                               -GG -TA---TT-TT-----------G
Lbc2
Lbc3
ΨLb1    --CA--------T--A-----  -A----A----- -                                    ------------TTATT   ----------A
```

IVS-3

```
Lba  AA GTTCAACTTAAAATTTT                          ATAGAT   GTT AATTGATA TAATTTGTTGAGATGATGAGAAGACCAATACC

Lbc1 -- -- --------------                          G--A--ACA-A-CG-- -C--G--------------C-------A--G  ----

Lbc2                                                                                          TAGTAATGAAT

Lbc3                                                                                          AAGTAATGGAT

ΨLb1 --A-- --------  ----GTT (132 nucleotides) ATT


                      h"             h"
                    _____          ______
Lba                 ATTAC          GTACTC TTTTGAAA GTGTT  ATA  TG GA TTTTAATTATAAGGAAAA    ATGTAAGAGCTA

Lbc1                -C---TCCAATAGCAT -----A---------AT---  ---AC--T--  -C---------------AGTGT--A---------
                      h'             h'
                    _____          _____
Lbc2 TTACTTAAAATCTTAA-TTAT          GTACT  -------- -- --  ---  -- --A-----------G------    ------------
                      h'             h'
                    _____          ______
Lbc3 TTACTTAAAATCTTAA-TTAT          GTACTT C---  ---A-A---TTG--  -- --A------------A-----    ------------

ΨLb1                        AATAGCAT ----GC    ---AC- TTA ---  --AA-A -----------------    ------------


Lba  AACCAT      TGCTG       ATTTTGAAG

Lbc1 -T----      -AT-ATTTTTTAT------T--

Lbc2 -T----      -AG--TTTTT   G-C--T--

Lbc3 ------      -----ATG    ----C----

ΨLb1 --A---CATTGT--          --- C-T--
```