Exon-Intron Structure and Evolution of the Lipocalin Gene Family

Diego Sánchez,*¹ María D. Ganfornina,*¹ Gabriel Gutiérrez,† and Antonio Marín†

*Departamento de Bioquímica y Fisiología y Genética Molecular-IBGM, Universidad de Valladolid-CSIC, Valladolid, Spain; and †Departamento de Genética, Universidad de Sevilla, Sevilla, Spain

The Lipocalins are an ancient protein family whose expression is currently confirmed in bacteria, protoctists, plants, arthropods, and chordates. The evolution of this protein family has been assessed previously using amino acid sequence phylogenies. In this report we use an independent set of characters derived from the gene structure (exon-intron arrangement) to infer a new lipocalin phylogeny. We also present the novel gene structure of three insect lipocalins. The position and phase of introns are well preserved among lipocalin clades when mapped onto a protein sequence alignment, suggesting the homologous nature of these introns. Because of this homology, we use the intron position and phase of 23 lipocalin genes to reconstruct a phylogeny by maximum parsimony and distance methods. These phylogenies are very similar to the phylogenies derived from protein sequence. This result is confirmed by congruence analysis, and a consensus tree shows the commonalities between the two source trees. Interestingly, the intron arrangement phylogeny shows that metazoan lipocalins. We also analyze the relationship of intron arrangement and protein tertiary structure, as well as the relationship of lipocalins with members of the proposed structural superfamily of calycins. Our congruence analysis validates the gene structure data as a source of phylogenetic information and helps to further refine our hypothesis on the evolutionary history of lipocalins.

Introduction

Ever since the rise of Gram-negative bacteria the lipocalins have been functioning and evolving in these organisms and in their eukaryotic symbionts, which possibly acquired the primordial lipocalin gene through a horizontal transmission event (Bishop 2000). A protein family developed through the standard evolutionary mechanisms of gene duplication and divergence (Ohno 1999), giving rise to at least 10 different genes that are currently recognized in the most recently evolved organismal taxa. Previous studies of three-dimensional and sequence similarity grouped lipocalins in kernel or outlier subfamilies based on the presence of three structurally conserved protein regions (SCRs) (Flower, North, and Attwood 1993) in the β barrel-based structural fold of lipocalins (fig. 1*A*).

Several phylogenetic reconstructions have been built upon the alignment of amino acid residues of lipocalin sequences (e.g., Igarashi et al. 1992; Toh et al. 1996). We have performed comprehensive phylogenetic analyses of the lipocalin family using protein sequence alignments with guidance based on protein structure data, and tree-building methods based on maximum likelihood (ML) (Ganfornina et al. 2000; Gutiérrez, Ganfornina, and Sánchez 2000). These studies group lipocalins in well-supported clades. When rooted with the bacterial lipocalins, the topology of the tree and the organismal distribution of lipocalins suggest that these proteins tend to increase the rate of sequence divergence and of gene duplication during evolution. Also, their internal pocket appears to have evolved toward binding smaller hydrophobic ligands with more efficiency.

An extensive literature supports the conservation of exon-intron structure in clades of orthologous genes

¹ Contributed equally to this work.

Key words: lipocalin, calycin, molecular evolution, gene phylogeny, exon-intron, intron evolution.

E-mail: opabinia@ibgm.uva.es.

Mol. Biol. Evol. 20(5):775-783. 2003

DOI: 10.1093/molbev/msg079

© 2003 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

(COGs) (Rokas, Kathirithamby, and Holland 1999; Wada et al. 2002), as well as in families of paralogous genes (Krem and Di Cera 2001) and protein superfamilies (Betts et al. 2001). These findings support the use of gene features as sources for phylogenetic inference (Rokas and Holland 2000; Krem and Di Cera 2001). In a previous report Salier (2000) proposed a scenario for the evolution of the lipocalin gene family by studying the gene structure and chromosomal location of 15 lipocalins. However, this view of lipocalin evolution is very dependent on the concepts of kernel versus outlier subfamilies, and it conflicts with our proposed evolutionary history of lipocalins (Gutiérrez, Ganfornina, and Sánchez 2000). To reassess our hypothesis of lipocalin evolution, we have used gene structure features as characters to build phylogenetic trees through different tree-reconstruction methods.

In this report we present the gene structure data of three insect lipocalins that we have been studying for their role in nervous system development (Ganfornina, Sánchez, and Bastiani 1995; Sánchez, Ganfornina, and Bastiani 1995; Sánchez et al. 2000b). The position and phase of introns in a number of lipocalins are used to reconstruct a phylogeny by maximum parsimony methods and by a distance matrix built with a measure of gene structure similarity (Betts et al. 2001). We also analyze the variability in introns present in the C-termini of lipocalins belonging to different COGs, and compare lipocalin intron arrangement with tertiary structure. We test the conservation of intron arrangement within the calycins, a proposed structural superfamily (reviewed by Flower, North, and Sansom 2000). Finally, we analyze the congruence of phylogenies based on protein sequence and gene structure, and build a consensus tree to refine our hypothesis on lipocalin evolution.

Materials and Methods

Genomic PCR Amplification of the Lazarillo Gene

Genomic DNA was purified from grasshoppers (*Schistocerca americana*). Brain tissue was lysed in 25 mM EDTA, 0.5% SDS, and 0.1 mg/ml proteinase K. The



FIG. 1.—A, Schematic diagram of the topology of the lipocalin structural fold. β strands are represented by white arrows, lettered A–H. α helices are shown as barreled cylinders. The boxes outline the three structurally conserved regions (SCRs). *B*, Schematic representation of the position of introns in representatives of lipocalin clades. White boxes correspond to the gene CDS, and gray boxes represent the untranslated regions. Size (in nucleotides) is shown above each exon. Lines represent intron insertions (not drawn to scale), and the phase of each intron is indicated above the line.

DNA was extracted with phenol and RNAse A-treated. We used the Expand Long Template PCR System (Roche Biochemicals) following manufacturer's specifications to perform polymerase chain reaction (PCR) amplifications in a thermal cycler (GeneAmp 9700 Perkin Elmer) using thin-walled plastic tubes (PE Biosystems). Primers were designed from the Lazarillo cDNA sequence (GenBank Accession Number U15656) with the primer3 program (www-genome.wi.mit.edu/cgi-bin/primer/primer3 WWW.cgi), and synthesized by Amersham Pharmacia Biotech. Primer sequences were: A5' (GTGCTGC-TGTCTGTAAGCTG) and A3' (TGGAGTTGACG-ACTGTGATG) for amplification of intron A; B5' (ACGGCAGAGTACTCCATGTCG) and B3' (AGCTCGCTCGCGAACTCTGC) for testing the presence of intron B; D5' (CGACAACTACTCCATTGT-GTGG) and D3' (GCTGCAGATTCTTCAGCTCATC) for amplification of intron D; and primers EF5' (TCCTA-TTACGATCACGGAAC) and EF3' (TCATGACTCGCT-GACCATAC) for testing the presence of introns E/F. Polymerase chain reaction products were sequenced with an ABI Prism 377 automated DNA sequencer using Taq FS DNA Polymerase.

Sequence Searches and Alignments

We searched for lipocalin genes whose intron-exon structure has been confirmed by the knowledge of their mRNA sequence. No deduced intron-exon arrangement was included in the analysis to avoid "noise" produced by poorly predicted splice sites, and to discard pseudogenes. Using the same seeding process and selection criteria previously described (Ganfornina et al., 2000; Gutiérrez, Ganfornina, and Sánchez 2000), a search for lipocalin cDNA and EST sequences was performed using the Blast program (Altschul et al. 1990) in the GenBank database available December 13, 2001. Thirty-seven of the sequences retrieved contained the complete CDS of the lipocalins and had the corresponding genomic sequence available on the databases. These genes are shown in table 1. We evaluated the presence, location, and phase of introns for these genes, and made a selection (asterisks in table 1) based on two criteria: (1) being the representative of a lipocalin COG (to avoid sampling bias), and (2) showing an intron pattern unique in the family. Thus we are accounting for the overall gene structure variation present in the lipocalin family.

Protein sequences were aligned with ClustalX (1.8) (Thompson et al. 1997) using a Gonnet series scoring matrix and a gap penalty mask based on the aligned secondary structures of the lipocalins with known tertiary structure. Based on our knowledge of lipocalin structure and function, we made minor manual corrections to the alignment. Intron positions and phases were then mapped onto the protein sequence alignment. Intron phase was named 0 when the intron splits two consecutive codons; I if an intron locates between the first and second codon nucleotides; and II if an intron locates between the second and third codon nucleotides. We used only the introns intervening the ORF of lipocalin genes, as those located in the 5'- and 3'-UTR can not be mapped onto the protein sequence alignment.

Phylogenetic Analyses

Phylogenetic analyses based on protein sequences were carried out using the maximum likelihood method with the MOLPHY 2.3 software (Adachi and Hasegawa 1996) as previously reported (Ganfornina et al. 2000). Bootstrap support for tree branches was estimated using the resampling log likelihood method (Hasegawa and Kishino 1994) to calculate local bootstrap proportions (LBP).

We have used intron positions of 23 representative lipocalins as phylogenetic characters. We built three input matrices based on three intron character states: (1) the presence/absence of a given intron, (2) the intron phase, and (3) the intron position in the alignment. Two procedures were carried out: The first was a maximum parsimony analysis using the intron presence matrix as input. Characters were considered as unordered. We made heuristic tree searches by the TBR method of PAUP* (Swofford 1998). A majority rule consensus tree was constructed from the most parsimonious trees found in the analysis. The second procedure started with the construction of a distance matrix based on a measure of gene structure similarity (Betts et al. 2001) that uses the presence, location, and phase of intron matrices described above to estimate the exon-intron similarity between two of the aligned proteins.

$$S_G(a,b) = rac{1}{2N_{ ext{max}}}\sum_{i=1}^{N_{ ext{equiv}}} \left(rac{1}{1+e^{\gamma(d_i-\delta)}}+ arphi(a_i,b_i)
ight)$$

 Table 1

 List of Experimentally Determined Lipocalin Gene Structures

Protein	Species	Abbreviation	Clade	Taxon ^a	Accession Number
Lipocalin	Dictyostelium discoideum	Ddis.Lip*	Ι	Р	JC1b154f03
Lipocalin fly neural-Lazarillo	Arabidopsis taliana	Atha.OML*	Ι	Pl	NC_003076
	Drosophila melanogaster	Dmel.Nlaz*	II	А	L81559
Lipocalin fly glial-Lazarillo	Drosophila melanogaster	Dmel.Glaz*	II	А	DS01087
Lipocalin Karl	Drosophila melanogaster	Dmel.Karl*	II	А	AE003487
Insecticyanin A	Manduca sexta	Msex.IcyA*	II	А	X64714
Insecticyanin B	,,	Msex.IcyB	II	А	X64715
Lazarillo	Schistocerca americana	Same.Laz*	II	А	In process
Apolipoprotein D M16695–6	Homo sapiens	Hsap.ApoD*	Π	PM	M16648-9
"	Mus musculus	Mmus.ApoD	II	PM	NW 000107
Retinol binding protein	Homo sapiens	Hsap.RBP	III	PM	NT 030084
<i>C</i> ,1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	Rattus norvegicus	Rnor.RBP*	Ш	PM	M10610
K03045-6					
Beta-lactoglobulin B	Bos taurus	Btau.BLB*	IV	PM	Z48305
.,	Capra hircus	Chir.BLB	IV	PM	Z33881
Beta-lactoglobulin	Macropus eugenii	Meug.BL*	IV	MM	L14954-60
Beta-lactoglobulin B	Ovis aries	Oari.BLB	IV	PM	X12817
Beta-lactoglobulin A	,,	Oari.BLA	IV	PM	M32232-37
Glycodelin	Homo sapiens	Hsap.Glyc*	IV	PM	M34046
Prostaglandin D synthase	Homo sapiens	Hsap.PGDS	V	PM	M98537-39
"	Rattus norvegicus	Rnor.PGDS*	v	PM	M94134
"	Mus musculus	Mmus.PGDS	v	PM	Y10138
Neutrophil gelatinase lipocalin	Homo sapiens	Hsap.NGAL*	v	PM	X99133
,, I 3, I	Mus musculus	Mmus.NGAL	v	PM	X81627
Quiescence protein-21	,,	Ggal.OS-21*	v	В	AF121346
Alpha-1 microglobulin	Homo sapiens	Hsap.A1mg	VI	PM	M88165
M88243-47	1	1 0			
M88249					
"	Mus musculus	Mmus.A1mg*	VI	PM	AF034692
Complement C8y subunit	Homo sapiens	Hsap.C8GC*	VII	PM	U08198
Major urinary protein 1	Mus musculus	Mmus.MUP1*	VIII	PM	X03208
Aphrodisin	Mus musculus	Mmus.Aphr	Х	PM	NW 042625
Aphrodisin	Mesocricetus auratus	Maur.Aphr*	Х	PM	AJ225170
Alpha-1 acid glycoprotein 2	,,	Hsap.a1G2*	XII	PM	AH007409
von Ebner's gland protein	Homo sapiens	Hsap.VEG	XIII	PM	L14927
,, ¹	Sus scrofa	Sscr.VEG*	XIII	PM	V96150
von Ebner's gland protein 1	Rattus norvegicus	Rnor.VEG1	XIII	PM	X74805
von Ebner's gland protein 2	"	Rnor.VEG2	XIII	PM	X74807
Epididymal RA-binding prot.	Rattus norvegicus	Rnor.ERBP*	XIV	PM	X59831
Epididymal RA-binding prot.	Mus musculus	Mmus.ERBP	XIV	PM	U68381
Odorant binding protein	Mus musculus	Mmus.OBP1	Х	PM	NW_042625
Odorant binding protein IIa	Homo sapiens	Hsap.OBP2a*	Х	PM	AJ251029
Odorant binding protein IIb	Homo sapiens	Hsap.OBP2b	Х	PM	AJ251025

^a Abbreviations: A, arthropod; Am, amphibian; Bi, bird; F, fish; MM, marsupial mammal; PM, placental mammal; P, protoctist; Pl. plant; R, reptile.

* Lipocalins chosen for gene structure phylogenetic analysis (see criteria in Materials and Methods).

 S_G is the similarity measure for the two proteins (a and b); $N_{\rm max}$ is the largest number of introns found in either protein; N_{equiv} is the number of equivalent (homologous) introns; a_i and b_i are the *i*th equivalent intron positions in the two proteins; d_i is the difference in position of the introns within the two proteins (in amino acids); $\varphi(a_i, b_i)$ is 1 if the intron phases are the same and 0 if they are different; γ and δ are constants (0.2 and 30) optimized such that the sigmoid function is insensitive to small changes in intron positions $(\pm 10 \text{ residues})$ (Betts et al. 2001). A matrix of distances calculated by this method was used to reconstruct a tree by the Neighbor-Joining method (Saitou and Nei 1987) implemented in PHYLIP (Felsenstein 1993). The generation of consensus trees and the analysis of congruence were performed with the RadCon program (Thorley and Page 2000). The protein sequence alignments and gene structure data matrices used for the phylogenetic studies are available from the authors upon request.

Results and Discussion

Gene Structure in the Lipocalin Family

Lipocalin genes have been found to contain four to eight exons (Salier 2000). A schematic representation of the position of introns is shown in figure 1*B* for representative lipocalins. Most introns interrupt the ORF, and only a few appear to be located in the 5'- and 3'-UTR of lipocalins. The position and phase of introns intervening the lipocalins ORF appeared to be fairly conserved. These similarities in exon-intron organization provide strong support for a common origin of the lipocalin genes and therefore make gene-structure information suitable for phylogenetic inference.



FIG. 2.—Exon-intron arrangement of the insect Lazarillo genes. *A*, Diagram of the gene structure of a model lipocalin. Introns in the ORF region are named A–F. Arrows and numbers below them show the primer sets designed to amplify specific introns from the genomic DNA. *B*, Photograph of an ethidium bromide gel showing the results of PCR amplifications from grasshopper genomic DNA with the primer sets shown in *A*. *C*–*E*, Diagram of the Lazarillo genes in grasshopper and Drosophila. Introns size are shown by numbers in *C*, and scaled in *D* and *E* as represented by the scale bars. The asterisk in *E* shows a unique intron in the 5' region of the DNLaz ORF.

Before the present analysis, sampling of metazoan lipocalin genes was strongly biased toward the chordate phylum. Only one gene from arthropods was reported (Li and Riddiford 1992). Therefore, we set out to study the gene structure of other known arthropodan lipocalins.

Exon-Intron Arrangement of the Lazarillo Genes in Schistocerca and Drosophila

The gene structure of Lazarillo, a lipocalin found in the grasshopper *Schistocerca americana* (reviewed by Sánchez, Ganfornina, and Bastiani 2000*a*), would be of great value in providing insight into lipocalin evolution because of the ancestral position of orthopteroids within the arthropod lineage (Caterino, Cho, and Sperling 2000).

The ORF of lipocalin genes is interrupted by 6 introns at the most. These introns (named A–F) are represented in a model lipocalin depicted in figure 2A. The predicted location of the six introns in the grasshopper Lazarillo gene was deduced by locating intron positions in a multiple protein sequence alignment of Lazarillo with other lipocalins of known gene structure. We then designed Lazarillo primers that would PCR amplify specific introns from genomic DNA. The primer sets are shown numbered under the lipocalin model in figure 2A. The PCR amplifications using grasshopper DNA appear in the ethidium bromide gel shown in figure 2B. Each numbered lane refers to the set of primers used. These amplifications revealed the presence of three introns in the CDS of the Lazarillo gene (fig. 2C) that corresponded to introns A, C, and D of the model lipocalin gene. Intron size was estimated by band size for introns A and C, and by complete sequencing for the short intron D. Sequencing the PCR products defined the exact location and phase of the Lazarillo introns (see table 2). These intronic sequences are deposited in GenBank (Accession Numbers: AY197702, AY197703, AY197704, and AY197705).

The availability of the Drosophila genome sequence has made possible to locate the introns present in NLaz and GLaz, the two fruit fly lipocalins homologous to Lazarillo (Sánchez et al. 2000*b*). The intron location and size are represented schematically in figure 2D-E, and their sequence boundaries are shown in table 2. Three and four introns are present respectively in the GLaz and NLaz genes that are common to other lipocalins (see below). A unique intron located in the signal peptide (pointed with an asterisk in fig. 2*E*) is present in the N-terminal region of NLaz.

Phylogenetic Analysis of Lipocalins Based on Gene Structure

In addition to the already characterized lipocalin genes (Salier 2000) and the arthropodan Lazarillo genes reported above, we searched for other lipocalin genes whose intron-exon structure was confirmed by the knowledge of their mRNA sequence. All the lipocalin genes found are listed in table 1, with genes selected for the analysis marked with asterisks (23 representatives; see *Materials and Methods*).

We found a protoctist gene (from *Dictyostelium discoideum*, EST #C24642), a plant gene (from *Arabi-dopsis thaliana*, mRNA Acc. Number AY062789), and another Drosophila gene (Karl, EST # NM_132520). The Dictyostelium and Arabidopsis genes are of singular value for our evolutionary analysis because they are the only representatives of lipocalins from unicellular eukaryotes and plants.

Alignment of Lipocalin Gene Structures

The intronic architecture of the selected lipocalin genes was mapped onto a multiple protein sequence alignment in the context of the overall secondary structure of an archetypal lipocalin (fig. 3). Noteworthy, there is a strong conservation of the location and phase of introns, a finding also reported in other gene structure analyses (Igarashi et al. 1992; Holzfeind and Redl 1994; Toh et al. 1996; Lindqvist et al. 1999; Salier 2000). This conservation is evident among COG members, but also among paralogous lipocalins. Some intron positions and phases are very well conserved (e.g., intron A), while others show slight variations (e.g., B and C). Some introns are present in most lipocalins (e.g., introns A and C) while others are present only in a subset of them (e.g., introns D, E, and F).

An important assumption of our analysis is the homology of each intron (A–F) found in the ORF of lipocalins. We accept that some variation in intron position could be due to ambiguities in the alignment of paralogous genes, where nearby insertion/deletions can cause apparent displacement of intron positions (Stoltzfus et al. 1997). A systematic examination of orthologous sequences would be needed to evaluate the presence and relevance of

Table 2
Exon-Intron Boundaries Present in the CDS of Drosophila and Schistocerca Lipocalin Genes

Lipocalin Gene	Intron	Splice Donor		Splic	Codon Phase	
Dmel.DNLaz	α	CAC TCG AG His Ser Se	gtaagcgcca	atccccacag	T TCG CAC r Ser His	2
	А	GCG GAA GCG	gtgagttctg	aatacttcag	TAT ATG GGC	0
	В	AAT CGA TT Asn Arg Le	gtgagtatca	gatgaaaaag	C ACC GGA	2
	С	CCG ACG C Pro Thr G	gtgagtaatg	tacattttag	AG CCA TTG	1
	D	AAT TTC A Asn Phe L	gtgagttaat	ttaattgcag	AA ATT GTT ys Ile Val	1
Dmel.DGLaz	А	ATG AGT CGG Met Ser Arg	gtaagttagt	tatcttgtag	GTC CTT GGA Val Leu Gly	0
	В	AAT CGC AT Asn Arg II	gtatgattaa	tcctttttag	A ACT GGT e Thr Gly	2
	С	GAT TTT AAG Asp Phe Lys	gtatctacaa	tttttcctag	TTT ACC ACC Phe Thr Thr	0
Same.Laz	А	GCC ACG CTG Ala Thr Leu	Unsequenced	gatttcgtag	TAC ATG GGG Tvr Met Glv	0
	С	AGT GTT G Ser Val G	gtgagtttac	aatgttgcag	GT AAC TAC Ly Asn Tyr	1
	D	TCT ACA G Ser Thr G	gtcagtcagt	ctctgtgcag	AA ATC TCA lu Ile Ser	1

intron-sliding as an additional source of variation. Our selection of genes, most of them paralogous to each other, precludes us from answering this question. Nevertheless, independently of its source, the intron position variation is incorporated in our distance measure (see *Materials and Methods*) and is used to assess the evolutionary history of lipocalins.

Taking into account the presence, position and phase of introns, we performed both maximum parsimony and distance-based phylogenetic reconstructions. The resulting trees (fig. 4) are rooted with the Dictyostelium lipocalin for its presence in an ancient organismal lineage (whose origin predates the arrival of metazoans), and because of the ancestral character of this protein sequence as judged by its similarity to bacterial lipocalins.

Maximum Parsimony Analysis

This analysis recovered six equally parsimonious trees (minimum step number 8). The majority rule consensus tree is shown in figure 4A. This tree (that computes the presence or absence of introns A–F as discrete character states) resolves five gene structure-related groups: (1) the Dictyostelium and plant lipocalins, (2) two arthropodan lipocalins (Laz and IcyA), (3) two Drosophila lipocalins plus ApoD and RBP, (4) a numerous group of lipocalins that belong to the clades IV-XIII (defined in our protein phylogeny, Gutiérrez, Ganfornina, and Sánchez, 2000; see table 1 for details), and (5) the three lipocalins bearing six introns (C8GC, a1mg and ERBP). The fruit fly Nlaz gene sets apart, although grouped with the remainder arthropodan lipocalins, due to its unusual set of introns.

Distances Phylogeny

A distance-based phylogenetic reconstruction was carried out by computing a distance matrix with gene

structure data (intron presence, location, and phase). These data were combined to produce a quantitative measure of gene structure similarity (Betts et al., 2001; see Materials and Methods). The Neighbor-Joining (NJ) tree (Saitou and Nei 1987) rooted with the Dictyostelium lipocalin is shown in figure 4B. Similar to the parsimony tree, the NJ tree relates monophyletically most arthropodan lipocalins with ApoD, and segregates the Drosophila NLaz and RBP as genes with unique exon-intron structures. The Arabidopsis and Dictyostelium lipocalins remain at the base of the tree, and the set of lipocalins belonging to clades IV-XIII are forming a monophyletic group, also related to the 6-intron C8GC, a1mg and ERBP. Despite displaying short branch lengths, this tree also establishes relationships among different lipocalin COGs, as can be seen in the cladogram shown in figure 4B.

Gene Structure versus Protein Sequence Phylogenies

The gene structure-inferred view of lipocalin evolution shares basic topological features with the protein sequence-based phylogeny (see Gutiérrez, Ganfornina, and Sánchez 2000). Although in principle there are no reasons to expect congruence between these two trees, it is clear that in both phylogenetic reconstructions the arthropodan lipocalins are related to ApoDs, and they appear related to protoctist and plant lipocalins; RBPs form a separate group, related to some insect lipocalins; and the rest of lipocalins form a well supported monophyletic group. To further test this, we built a ML tree using the protein sequence alignment from which the gene structure matrices were derived, and rooted this tree with the Dictyostelium lipocalin (fig. 5A). We used the program RadCon (Thorley and Page 2000) to evaluate the congruence of the protein sequence ML and the gene structure NJ trees. Both source trees are well resolved:

	intron A			Intron B		
Ddis Lin	VIY BYN SFKRYT PREWIE VKP FYP RKY	UCKWYETARI YTYEKU		TTUNSCYNYKKKRE		ICSII
Atha OM	MTRKKRMEVVALINVERY	CRUYETASEPSROPKN		THUT.NETWEN-CKRCF	EGS AYKANP	KSDEAK
Hsao. Apol	ORFHLGKCPNPPVOENEDVNKY	LGRWYEIKKIPTTENG	RCIOANYSLMENICK	INVINDELEADGTVNO	EGEATP VNL	TEP
Dmel.NLaz	RO VPF PCKCPD VKLLDT FDAKAY	GVWYEYAAYPFAKIG		VSVVNAAINIR TOOPS	VIGOAKVLG	PG0
Dmel.GLaz	AGTD AYGFGRCPNYPS) PKFNMSRV	LGHWYE VER SFYLEI ASGCTTI	FOF EPYNKGEOSKESNEK	LAVALKNINE T GNP N	VNIGYAPENSESSI	100FK
Dmel.Karl	YOT RRR SEP SNRCPKVEALKNFDLERO	GCVHVVUYYASTELP	EYACME SHE SESKED	DHITTONESYLFARDPL	RKLVGNITVMI	PKFOEP
Same. Laz	AGET AGCAD RSAIND FNATLY	CKWYEYAKIGSIYEE	GEV CVT AEY SMS SNN	- ETVVNSMKDNITHEVN	T – – – TTUWAEF ASE	L-HT-D
Msex. IcyA	GD LFYPGYCPE VKP VDD FDL SAF	AGAVHE TAKLPLEENE	GKC TVAEYKYDG KKA:	SVYNSF VINGVKEYME	GDLE TAP DAK	LTKQ
Rnor. RBP		SELWATAKKDPELFL	QDNILLAEFS VDE KLED	ISATAKGRVRLLSWE	VCADMVGTFT	DTE -DP
Btau. BLB	LIVTQTMAGLDIQKV	AGTWYSLAMAASDISL	LDAQSAPLRVYVEELI	KPT PEGDLE ILL OK DE	NGEC AQKKII	AEKTKI
Hsap.Glyc	VPAMDI PQT KQDLELPKL	AGTVHSMAMATNNESL	MATLKAPLR VHITSL	LPT PEDNLE IVL HRVE	NNSC VEKKVL	GEKTGN
Meug. BL	VENIRS KNDLGVEKF	VGSVYLREAAKT	MF SIPLFDMDI KEVI	NLTPEGNLELVLLEN	DRC VEKKLL	LKKTQK
Rnor. PGD S	OCHDTVQPNFQQDKF	LGRWYS AGLASN SSW	FREKKELLFMCQTVV7	APSTEGGLNLTSTFLE	KNQCETKVM	LOPAGV
Hisap.NGAL	QDSTSDLIP APPLSKVPLQQNFQDNQF	QGKWYVVGLAGNAI	LREDKDPQK247ATTY1	ELKEDK SYNVT SVLF	KKKC DYWIRT	FVPGCQ
Minus, Alting	DPASTLPDIQVQENFSESRI	CKWYNLAVGSTCWL S	RIKDKMS VQTLVLQEGI	ATETEL SMT STRWERG	VCEEITGAYQ	ktd I
Hsapp.C8GC	QKP QRP RRP ASP I ST I QP KANFDAQQF	AGTVILL VAVGSACFLQ	EQGHRAEATTL HVAJ	POGTALOAVSTFRILDC	ICWQVRQLYG	DTGV
Rnor. ERBP	AVVKDFDI SKF	LGFWYE IAF ASKM	TGLAHKEEKAGAMV	VEL KENLLALTT TYY S	DHCVLEKVTAT	EGDG-P
Ggal.QS-21	ARTVPDS SEV	AGKWYI VALASNTS	FLRENGKOKOVOART	SFLCEDELEVSYAAPS	KGCRKVET TF-	KKT S
Minus, MOP 1	HAEEA SST GRNFNVEKI	NGEVHTIILASDKEKI	EDNGNFRLFLEQIH	VLENSL VLKFHT <mark>V</mark> RDE	ECSELSMVAD	KTEK
Sscr. VEG	AQEFPAVGQPLQDL	LGRWYL KAMT SDPIP	GKKPE SVT PLI LKAI	LECCOLEMOITE	QCQD VIIL VLK	KIN
Maur. Aphr	QDF AEL	QGKWYTIVI AADNEKI	KEGGPLRFYFRHID(CYKNESE@ETTFYV I T	NNQC SKTTVI	GYLK-G
Hsap.0BP2	SFTLEEEDI	IGTWYVKAMVVDKFPE	DRRPRKVSPVKVTAJ	LOG GNL EAT FTF	RCIQKKILMR	KTEE
Hsap.a.1G2	QIPLCANLVPVPITNATLDRI	ICKWFYIIASAFRNEYN	KSVQELQATEFYETPNI	KTEDTI FLREYOT ON	QCFYNSSYLN	VORE
						·>
						~~~
	intron C	ir	ntron D		intron E	「intron F
Ndis Lin	intron C		ntron D	T.BLAKSVEYDTTK-L	intron E	intron F
Ddis.Lip Atha OM	intron C -EGMLKVSEFCPEYSGYNV. LKVKEYVPPEELPTTPVTCHYMV	ir IAI DPDYKYAIAGQ SEDY	ntron D Manlskeptiperki knsv Manlsetaomeretykni	YLELAKSVGYDI TK-L	intron E IWSKUNENEN	intron F
Ddis. Lip Atha. OM Hsap. Apoll	intron C -BGMLKVSFF	ir IAI DPDYKYAIAGQ SFDY LYI DPDYQHAIGQP SRSY YENYAVYSCTCI IOLFHVDJ	ntron D navilskeptiperteriktiknsv rlvalsreaqueretykoj favillarnenlepetvosj	YLELAK SVGYDI TK-L LVE KAVERGYDI SK-L LKNILT SNNIDVKK-M	intron E IWSKONENEN HKTPQ SDT PPE SNT IVTDO VNC PKL S	intron F
Ddis.Lip Atha.OM Hsap.ApoD Dmel.NLaz	intron C -ECMLKVSFFCPFYSGYNU. LKVKFYVPPFLP	ir IAI DPDYKYAIAGQSFDY LYI DPDYQHAIGQPSRSY YENYAVYSCTCI IQLFHVDJ YESYAVYSCTSVTPLANED	ntron D Amilskeptipekiknsy Alwilsetaoneketykdi Fruilarnpnippetvosi Cwillteorepsaeavdad	YLELAK SVGYDI TK-L LVE KAVERGYDI SK-L LKNILT SNNIDVKK-M ARKILEDND V SO AF-L	intron E IVISKONENEN HKTPQ SDT PPE SNT IVIDQ VNC PKL S IDTVO KNC PRL DGN	APEDSKCW
Ddis.Lip Atha.OML Ksap.ApoD Dmel.NLaz Dmel.GLaz	intron C -ECMLKVSEF	ir LAI DPDYKYAIAGQSED LYI DPDYQHAIGQPSRSY YRNYAVYSCTCI IQLEHVDU YESYAVYSCTSVTPLANER YRNFALWSCCSIGSLEHSDU	Itron D Amiliskeptipeki knsy Almilsrtaquæketykdi Franlarnpnipetydsi Cwilltrorepsaeavdaj Invilgedrife vid rsky	YLELAK SVGYDI TK-L LVEKAV KEGYDI SK-L LKNILT SNNIDVKK-M ARKILEDND VSQAF-L VYD VLKRL SLDP ER-L	INTERNET INSKONENEN HKTPQ SDTPPE SNT INTDQ VNCPKL S IDTVQ KNCPKL DEN II SKNADCPKAL	APEDSKOW
Ddis.Lip Rtha.OML Ksap.ApoD Dmel.NLaz Dmel.GLaz Dmel.Karl	intron C -EGALKVSEF	ir IAI DPDYKYAIAGQSED LYI DPDYQHAIGQPSES YENYAVYSCTCI IQLEHUD YENYAVYSCTSVTPLANEB YENFALWSCGSI GSLEHSDQ YENFALWSCGSI GSQPRYL 5	ITON D MATLES KEPTEPEKT KNS KLWILLS RTAQUE KETYKE FRATLARNP NLP PET VDS CWILLT RORRES SARA VDA JUVILLGRORDFE VDI RSKT SALLES RNT SLADRE I SF	YLE LAK SVGYDI TK-L LVE KAVEEGYDI SK-L LKNILT SNNIDVKK-M ARKILE DND VSQAF-L VYD VLKRL SLDPER-L USCKLPQD-IDT SF-M	intron E IV/SKONENEN HKTPQ SDT PPE SNT TVTDQ VNC PRL SNT DT VQ KNC PRL LGRN II SKNKQC PERL ENIQO ESC DML MES	APEDSKGW GTGLAGEDG
Ddis.Lip Atha.OM Ksap.ApoD Dmel.NLaz Dmel.GLaz Dmel.Karl Same.Laz	intron C -ECMLKVSFFCPFYSGYNU LKVKFYVPPFLDCPFYSGYNU AKLEVKFSVBPDSAPYVLL ATD 	IRI DPDYKYAIAGQ SED LYI DPDYHAIGQ SED YENYAVYSCTCI LQ LEHVDJ YENYAVYSCTSVTP LANE YENFALWSCGSIGS LGHSDU YDTWFAWGA BEKKK QD RYL A DWYSVWSC VRRP DS RAS TED	NTON D MANILSKEPTIPEKI KNSY MANILSKTAQUEETYKQI FAMILARNPNLPETYKQI UWILFROREPSAEAVDAI ILWILGKORDFEVDIRSKY SALLSSRATSLADUEISFI SALLSSRATSLADUEISFI	YLE LAK SVGYDI TK-L LUE KAVERGYDI SK-L LUE KAVERGYDI SK-L LKNILT SNNIDVKK-M ARKLLE DND VSQ AF-L VYD VLKRL SLDP ER-L LRGKLP QD-LDT SF-M VED ELKNLDLD, NSYK	intron E IWSKQNENEN HKTPQ SDT PPE SNT TVTDQ VNC PKL S ID TVQ KNC PKL DGN III SKNKQC PKR L FNIGQ ESC DNL MES TEOSAKYC A	intron F APEDSKGW GTGLAGEDG SRDDPLAYV
Ddis.Lip Atha.OM Hsap.ApoD Dmel.NLaz Dmel.GLaz Dmel.Karl Same.Laz Msex.LcyA	intron C -EGMLKVSFF	IRI DPDYKYAIAGQ SFD LYI DPDYQHAIGQP SFD YENYAVY SCTCI IQ LEHVDI YENYAVSCTSVTD LEHVDI YENYALVSCGSI GS LGHSDI YDTYGCV/BC ARKKK OP RYI YDNY SVV/SC VARP -D SRA STD YENYAVYNC NYHP -D SKAHSDD	ITTON D MAILLSKEPTIPEKTENST MAILLSETAQUEEETYKD GAULLARNPNLPETVKD INTLERORDFESARAVDA DIVILLSENT SLADNE ISP ILSWILLSENT SLADNE ISP ISWILLESENT SLADNE ISP ISWILLESENTVLEGYTKEV	YLE LAK SVGYDI TK-L LVE KAVEEGYDI SK-L LUNILT SNNDVKK-M RKLLE DNO VSQ RF-L VYD VLKRL SLDP ER-L LRCKLP QD-DDT SF-M VED ELKNLQ LDL NKYK VFD VLKNLQ LDL NKYK	intron E IWSKQNENEN HKTPQ SDT PPE SNT TVTDQ VNC PKL S IDT VQ KNC PKL S IDT VQ KNC PKL L ENIGQ ESC DNL MES TE QSA KYC A MS NEE SEAACOYST	intron F APEDSKGW GTGLAGEDG SRDDPLAYV TYSLTGPDR
Ddis. Lip Rtha. OML Ksap. ApoD Dmel. NLaz Dmel. GLaz Dmel. Karl Same. Laz Msex. LcyA Rhor. RBP	$\label{eq:constraint} \begin{array}{c} & \text{intron C} \\ \hline \\ - & \text{EGMLKVSFF} - & - & \text{CPFYSGYNU:} \\ & \text{LKVKFYVPP ELPI : IPVTGDYNU: \\ & \text{RKUKKFSM} & \text{RKUKFSM} \\ & \text{RKUKFSM} & \text{RKUKTSM} \\ & \text{RKUKFSM} & \text{RKUKTSM} \\ & \text{RKUKFSM} \\ & \text{RKUMFKM} \\ \end{array} $	ir IAI DPDYKYAIAGQSED LYI DPDYQHAIGQPSRSY YENYAVYSCTCI IQLEHVDJ YESYAVYSCTSVTPLANE YENYAKSCCSIGSLEHSDO YDTVGV/OFCAEKKKQP RYL YENYAWYNCWRP-D SKASTE YKNYAWYNCWRP-D SKASTE YKNYAWYNCWRP-D SKASTE YKNYAWYNCWRP-D SKASTE	ITON D MATLISKEP TIP KKI KNST VIMILSRIAQUE EETYKDI FRAILLARND NLP PET VOSI ITWILGEDRIPE VOI RSK' SALLISRKT SLADNEI SP ISWILLSRKT SLADNEI SP ISWILLSRNKVLE GATKEV HERVLSRNKVLE GATKEV	VLE LAK SVGYDI TK-L LVE KAVEEGYDI SK-L LUNILT SINIDUKK-M ARKILEDND VSQ AF-L WID VLIKIL SLDP ER-L LRGKLP QD-IDT SF-M VED KLINIQ LDL INSYK VDNVLKTE SHLI DAS FI	intron E INSEQNEMENT INSEQNEMENT INTOQUE SUT PPE SNT INTOQUE SUT DEVELOCAT INTOQUE SUT DEVELOCAT INTOQUE SUT ALCONST INSERT SERACOVST INTENTACOSTOS	intron F APEDSKCW/ GTGLAGEOG SRDDPLRYV TYSLTGPDR NSL
Ddis. Lip Rtha. O.L Hsap. ApoD Dmel. NLaz Dmel. Karl Same. Laz Msex. LcyA Rnor. RBP Btau. BLB	intron C -BCMLKVSEF	IRI DPDYKYRIAGQSED LYI DPDYQHAIGQPSEY YENYAVYSCTCI IQLEHVDJ YESYAVYSCTSVTPLANE YINYAVYSCGSIGSLEHSDU YDTWSVNSCGSIGSQPRYL YDNYSVNSCVKRP-D SARSTE YNNYRNYNC NYHO-D KKAHS <u>D</u> YDTFAQYSCRLQNLDGTCADSY YDKYLFCURINSAEPEQSLJ	ITON D MATLES KEPTIP KAI KNSY VIMILS RTAQUE KETYND FRATLARNP NLP PET VDS I WILL ROREP SA BAVDAS I WILL ROREP SA BAVDAS SALLLSRAT SLADNE I SP I SWILL RSRNSS MATLER WEAVLS RANKVLE GYT KEV YSF VES RDP MAL TPE TRR)	YLELAK SWGYDI TK-L LUE LAK SWGYDI TK-L LUN TLT SINTDUKK-M ARKTLEDND VSQAF-L WTD VLKRL SLDP ER-L LIGKLPQD - IDT SF-M VED ELKNLQ LDL NKYK VDN VLLKTF SHLI DAS FI LVRQRQ FEELCLE RQ-R FUKALKALP JOH ELSN	intron E IV/SQNENEN HKTPQ SDT PPE SNT IV/TOQ VAC PKL SC ID TVQ KAC PKL DCAN II SKNRCP FAL FNIGQES CONLMES ITEQSAKYCA MSNEE SEAACQYST MI KHNG/CQSRPSR PULEBJCCH	APEDSKGW/ GTGLAGEOG SRDDPLAYV TYSLTGPDR NSL
Ddis. Lip Atha. O.L Hsap. Apol Dmel. N.Jaz Dmel. Glaz Dmel. Karl Same. Laz Msex. LcyA Rhor. RBP Btau. BLB Hsap. Clyc	$\label{eq:constraint} \begin{array}{c} \mbox{introd C} \\ - \mbox{EcmlKvSff} & - \mbox{CoppySgynv.} \\ \mbox{LkvKfyVppFL} & - \mbox{Lfirstraint} & - \mbox{CopySgynv.} \\ \mbox{RkfyVpfD} & - \mbox{LirknvLvLgyp} \\ - \mbox{LirknvFyffD} & - \mbox{LirknvLvLgyp} \\ \mbox{Charge for the firstraint} & \mbox{LirknvLvLgyp} \\ \mbox{Charge for the firstraint} & \mbox{CopySgynv.} \\ \mbox{Charge for the firstraint} & \mbox{CopySgynv.} \\ \mbox{Charge for the firstraint} & \mbox{CopySgynv.} \\ \mbox{CopySgynv.} \mbox{CopySgynv.} $	IRI DPDYKYAIAGQSRD LYI DPDYCHAIGQPSRD YENYAVYSCTSUTPLANES YENYAVYSCTSUTPLANES YENFALWSCGSIGSLGHSD YDTYSUSCSUTRSRASTE YNYANYNCNYHP-DKKAHSD YNYANYNCNYHP-DKKAHSD YKKYLFCMENSREPEQSLJ YDNFKLCLQUTTTPLQSND	ITTON D MANTLISKEP TIP EKI KNST MANTLISKEP TIP EKI KNST MANTLISKEP TIP EKI KNST MANTLISKEP SAER VOA JUVILGRORDFE VDI RSKT SALLISSKT SLADDNE ISE MANVLISKNKVLEGNTKEV YSE VESROP NGL TPETRRI ACQ CLÜRTP EVD DEALEKS MOYLINGEN VLOODELLOGI	YLE LAK SVGYDI TK-L LUE KAVERGYDI SK-L LKNILT SNNIDVKK-M RKLILE DND VSQ RF-L UYD VLKRL SLDP ER-L LKSKLPQD - DJT SF-M VED ELAVLQ LDL NSYK VNVLKTF SHLL DAST LURQRQ EKL CLE RQ-R FDKRLKALP MH RL SN ETDRAFELP MH.WF.D	intron E INSKNENEN HITPO SITPPE SIT ITTDO VINC PKL S III SKNNOC PKAL FRIGO ESCINLINES TEOSANYCA MISNEF SEAACOYST MI ENGFICOSRPS PTOLESICHI LOWEFC FE	APEDSKCW/ APEDSKCW/ GTGLAGEDG SRDDPLRYV TYSLTGPDR NSL
Ddis. Lip Atha. O.M. Hsap. Apol Dmel. NLaz Dmel. Claz Dmel. Karl Same. Laz Msex. LcyA Rnor. RBP Btau. BLB Hsap. Clyc Meug. BL	intron C -EGMLKVSFFGPFYSGYNU: LKVKFYVPPPLPIGPFYSGYNU: RKLEKVRESMENDSAPYULLATD -LAVAPYPTDITXNTVLLTD GHUHT KUTYECTYNTVLLTD GKUSVHENNSPSVGNYNUL STD GKYNATKEEGBNVVQUPVALATD RKFKINFKEGBNENKU/ULTD PAVFKINFLNENKU/ULTD PKEFKINFTVANENKU/ULTD PKEFKINFTVANENKU/ULTD PTEFECTYIS SESNENKU/ULTD	IRI DPDYKYRIAGQSED IXI DPDYDHAIGQPSED YRNYRVYSCTCI IQLEHVDJ YRSYRAVYSCTCI IQLEHVDJ YRSYRAVSCTSVTPLEHVDJ YRNFALWSCGSI GSLGHSD YDTFGQVSCTSVTPLGHSD YDTFAQYSCRLQNLDGTCADS YDTFAQYSCRLQNLDGTCADS YDNFLLCLQDTTTPLGMD YDNFLLCLQDTTTPRKD	ITTON D MANILSKEPTIPEKTENST MANILSETAOMEETYKD TRAILARNPNLPETVOJ TRAILARNPNLPETVOJ DIVILGRORDFEVDIRSK SALLSENT SLADNEISF ISMILLSENT SLADNEISF ISMILLSENT SLADNEISF MANULSENKVLEGNTKEV YSF VESROPNLTPETRI NOQULARVLVEDBELMG MOYLARVLVEDBELMG MANMARELE-ENKISME	VLE LAK SVGYDI TK-L LVE KAVEEGYDI SK-L LUNILT SNNDUKK-M ARKLLE DNO USQAF-L UYD VLKRL SLDP ER-L LREKLP QD - DDT SF-M VED ELKNLQ LDL NKYK WINVLKTF SHLI DAS E LVRQRQ EEL CLE RQ-R FDKALKALP MHI KLS N FTR AFRPLP RHLW7-D FTR AFRPLP RHLW7-D	intron E INSKINEMEN HKTPOSDTPPESNT TYTDOVOCKLS IIISKNNCCPKAL FNIGOSSCDMLMES TEOSAKYCA SILKINGCOSSCDMLMES FTOLESICHI LKOMERCRF LKOMERCRF	GTGLAGEDG SRD DPL RYV TYSLTGPDR NSL
Ddis, Lip Rtha. OML Hsap, ApoD Dmel. (Laz Dmel. (Laz Imel. Karl Same, Laz Msex, LcyA Rnor, RB Btau, BLB Hsap, Clyc Meug, BL Knor, PGD S	intron C -EGMLKVSFF	IRI DPDYKYAIAGQSED LYI DPDYQHAIGQPSED YENYAVYSCTCI IQLEHVDI YESYAVYSCTSVTPLAHVDI YENYAWSCCSIGSLEHSDI YDTWCV/MCAEKKKQPRYL YNYANYNCWYHP-DSKAHSDI YNNYANYNCWYHP-DSKAHSDI YNTFAQYSC KLQNLDGTCADS YKKYLECMENSAEPEGSLI YDNYHCLCLQDYTTPTQSKO YDSYFFCLYNISDRRODFRI YDEYALFSKCTKCPQDFRI	IT ON D MATLISKEP TIP KKI KNST MIMILSRTAQUE EETYKD FRAILARNP NLP PET VOSI ITWILF ROREPS AEA VDAI ITWILF ROREPS AEA VDAI ITWILF ROREPS AEA VDAI SALLLSRNT SLADNE I SPI ISWILL BRRNT SLADNE I SPI ISWILL BRRNT SLADNE I SPI ISWILL BRRNT SLADNE I SPI ISWILL BRRNT SLADNE I SPI ISWIL SRNKVLE GAT KEV NO YLARVL VED DEALEN MCANTYRRIE - ENNEMMEN MATLISF GAQ LLKEKL KEN	YLE LAK SVGYDI TK-L LUE KAVEEGYDI SK-L LUNILT SNNIDVKK-M ARKILE DND VSQ AF-L WID VLKRL SLDP ER-L URG KLP QD-IDT SF-M VED ELKNLQ LDL INSY K VDNVLKTF SMLI DAS F FUKALKALP MHI RLSN FIR AFRPL P RHLWY-D FRKILR TLAMPYTV-I FIK FIR KLAMPYTV-I FIT F SKOQGLTE KD-V	intron E IV/SULNENEN HKTPQ SDT PPE SNT IV/TOQ VAC PKL SC ID TVQ NC PKL DCN II SKNQC PKAL FNIGQ ESC DNL MES ITE QSANYCA MSNEF SKARACYST UKENGC QSR PSR PTQLE SJC HL EVRETRACK-HV EVRETRACK-HV EVRETRACK-UP	APEDSKGW/ GTGLAGEDG SRDDPLAYV TYSLTGPDR NSL
Ddis, Lip Rtha. OM Hsap, ApoD Dmel. Klaz Dmel. Glaz Dmel. Karl Same, Laz Msex, LcyA Msex, LcyA Btau, BLB Hsap, GlyC Meug, BL Rhor, PGD S Hsap, NGAL	intron C -BCMLKVSFF	ITAI DPDYKYAIAGQ SFD LXI DPDYKAIAGQ SFD YENYAVYSCTSUTP LANES YENYAVYSCTSUTP LANES YENYAVSCTSUTP LANES YENYAWSCTSUTF LANES YENYAWSC KIRC CASA YENYAWYAC NYHD -DKKAHSU YINYSUKCU KIRC CHSON YDSYALCLQDTTP IQSM YDSYALFSKGTKGP GUFRJ YDBYALFSKGTKGP GUFRJ YDBYALFSKKUSQM EKYFJ	NTON D MANILSKEPTIPEKIKNSY MANILSRTAQMEETYNQI FRAILARNPNLPETVDSI FRAILARNPNLPETVDSI FRAILARNPRS PARA VDAI ILWILTERDRIPES ARA VDAI ILWILLERDRIPES ANDRE ISE SALLLSRNT SLADNE ISE HERVULSRNKVLEGNTKEV TSE VES SOD MALTDETRRN MOQULARVL VED DELINGSI MCAHVURRIE – ENNSMEI MATINSRAQLIKKELKEN INTINGERKELTSELKEN	VLE LAK SWGYDI TK- L LUE KAVEEGYDI SK- L LUNILT SINIDUKK- M ARKILE DND VSQ AF- L WYD VLRKI. SLDP RR- L LRGKLPQD - IDT SF- M VED ELKNIQLDL NSYK VDIVLKTF SHLI DAS FI LURQEQEEK LLE RQ. F FDKALKALP MIT ELS N FTR AFR.PLP RH. WY- D FTK FLKTLAMPYTV- I FTF FSKDQGLTFE ED- V FTF FSKDQGLTFE D- V	intron E INSENSIONENT HITPO SUTPPESNT TVTDQ SUTPPESNT IT SKINIC PERLOR IT SKINIC PERLOR IT SKINIC PERLOR IT SKINIC PERLOR IS SER SER ACOYST IN ENDER SER ACOYST I	APEDSKGW/ GTGLAGEDG SRDDPLAYV TYSLTGPDR NSL
Ddis. Lip Atha. O.M. Hsap. Apol Dmel. N.Laz Dmel. Karl Same. Laz Msex. Lcy:A Rnor. RBP Btau. BLB Hsap. Glyc Msag. Glyc Msag. N.GAL Mmus. Almg	$\label{eq:constraint} \begin{array}{c} \mbox{introd C} \\ - \mbox{EcmlKvSff} \mbox{CppySgynv.} \\ \mbox{LkvKfyWppFLp} \mbox{CppySgynv.} \\ \mbox{RkfyWfp} \mbox{LkvKfyWfp} \mbox{LkvKfyWfp} \\ - \mbox{LkvKfyWfp} \mbox{LkvKfyWfp} \\ \mbox{Cpp} - \mbox{LkvKfyWfp} \\ \mbox{CkvKfyWfp} \mbox{LymyvLhfp} \\ \mbox{CkvKfyWfp} \\ \mbox{CkvKfyWfp} \\ \mbox{RkfyWfp} \\ \mbox{RkfyWffp} \\ \mbox{RkfyWfp} \\ \mbox{RkfyWffp} \\ $	IRI DPDYKYAIAGQSRD LYI DPDYCHAIGQPSRD YENYAVYSCTSUTPLANE YENYAVYSCTSUTPLANE YENFALWSCGSIGSLGHSD YDNYSUWSCUKRP-DSAASTE YRNYANYNCNYD-DSAASTE YRNYANYNCNYD-DSAASTE YNYFLCLEUTTPTQSD YDSYFFCLYNISDREKM YDFYLLESKETKCPQDFRI YDFYLLESKETKCPQDFRI YDFYLLESKETKCPQDFRI YDFYALLESKETKCPQDFRI YDFYALLESKETKCPQDFRI YDFYALLESKETKCPQDFRI YDFYALLTKKSSHHHELTT	ITTON D AMULS KEP TIP EKI KNST AMULS SETAQUE EETYKDI FRAJLARNP NLP PET VDS I FRAJLARNP NLP PET VDS I INTLGROR DFE VDI RSKT SALLS SRKT SLADNE I SE I ISWILL SSRKT SLADNE I SE I ISWILL SSRKT SLADNE I SE I SALLS SRKT LE STAT SALLS SRKT SLADNE I SE I SALLS SRKT SLADNE I SALLS SRKT SLADNE I SE I SALLS SRKT SLADNE I SALLS SRKT SLADNE I SALLS SRKT SLADNE I SALLS SRKT SLADNE I SALLS SL	YLE LÄK SVGYDI TK-L LUE KAVEEGYDI SK-L LKNILT SNNIDVKK-M ARKLIE BIN UVSQ RF-L UVD VLKRLS LDP ER-L LIGKLP QD - DJT SF-M VED ELKNLQ LDL NXYK UVD VLKTE SHLI DAS F CURARPLE KLUE RQ-R FDKALKALP MHI RLSN FTRAFEPLE PRH.WF-D FKKILRTLAMPYTV-J FTR FSKDQGJTE RD-V FTR FSKDQGJTE RD-V FTR FSKSLGLPE NH-V FTR FSK SLGLPE NH-V	intron E INSEQUENTEN HETPO SITPPE SIT ITVIDQ VIC PKL S IT SKNNQC PKAL S IT SKNNQC PKAL S HATGO ESCINULIES SKNCE SEARCOYST SITENESCH SCORESE PTOLESCH SCORESE PTOLESCH SCORESE EVERTRUCH V FP VPI DC IDG S FND VPI DC IDG S FND PKI CV FGRE	APEDSKGW/ GTGLAGEOG SRDDPLAYV TYSLTGPDR NSL
Ddis. Lip Atha. OM. Hsap. Apol Imel. NLaz Imel. Karl Same. Laz Msex. LcyA Rnor. RBP Btau. BLB Hsap. Clyc Msug. BL Rnor. PCD S Hsap. C&CC	intron C -EGMLKVSFF	IRI DPDYKYAIAGQSFD IXI DPDYQHAIGQPSFD YENYAVYSCTCI IQLEHVDJ YENYAVYSCTCI IQLEHVDJ YENYAVYSCTSVTPLANKE YENYAVYSCYSCE YDTFGQYSCRIQNIDGTCADST YDTFAQYSCRIQNIDGTCADST YDTFAQYSCRIQNIDGTCADST YDNFLICLQDTTTPTQSM YDNFLICLQDTTTPTQSM YDNFLICLQDTTTPRKM YDFYALFXSSHHREXFI YDYBYALFXSSHHREXFI YDSFFECINSSSHGREXFI YDFYALFXSSHHBLTTT YQSFALYLERAGQLS	ITTON D MANILIS KEPTIPEKI KNST VIMILS ETAOME EETYKDI FRAILARND NLP PET VOSI DIWILGRORDFE VDI RSK SALLS EKT SLADNE ISP ISWILLS RKT SLADNE ISP ISWILLS RKVLE GAVKEV VSF VES RDD NLI TPETRI MQYLARVI. VED DELLEKI MQYLARVI. VED DELLEKI MATINS RQ LLKEEL KEKI KTI INS RQ LLKEEL KEKI	VLE LAK SVGYDI TK-L LVE KAVEEGYDI SK-L LUN LLT SNNIDVKK-M ARKLILE DNO VSQ AF-L UYD VLKRL SLDP ER-L LREKLP QD - DIT SF-M VED ELKNLQ LDL NNYK WINVLKTF SHLI DAS E LVRQRQ EEL CLE RQ-R FINAFRPLP FIL DAS E FINAFRPLP FIL DAS T FTR AFRPLP FIL WA'D FTR FSK DQG LTE ED - V FTR FSK SLG LPE NH- V FTR FSK SLG LPE NH-V FTR FSK SLG LPE NH-V FTR FSK SLG LPE NH-V FTR DY AL NUK G SENS-1 FED RUQ EANLTE QQ-F	intron E INSEQUENTED SUT PPE SWI TUTDO VIC PKL SS- III SKINDC PKAL FNIGO SSC DNL MES III SKINDC PKAL FNIGO SSC DNL MES III SKINDC PKAL FNIGO SSC DNL MES III SKINDC PKAL FNIGSC SSC DNL MES III SKINDC	intron F APEDSKGW/ GTGLAGEOG SRDDPLRYV TYSL/GPDR NSL VEPTSL/R NVLDEVRR
Ddis, Lip Rtha. O.K. Hsap, ApoD Imel. NLaz Imel. Karl Same. Laz Isex. LcyA Rnor. RBP Btau. BLB Hsap, Clyc Meug, BL Rnor. PCD S Hsap. NGAL Imus. Almg Hsap. C&CC Rnor. KBP	intron C -EGMLKVSFF	IRI DPDYKYAIAGQSED LYI DPDYQHAIGQPSED YENYAVYSCTCI IQLEHVDI YESYAVYSCTSVTQLEHVDI YESYAVYSCTSVTQLEHVDI YDTWSV/SCTSVTGLEHVDI YDTWSV/SCTGIGSLEHSDI YDTWSV/SCTGIGSLEHSDI YDTWSV/SCTURESEA YNNYANYNCNYHP-DKAHSU YINYAVYNCNYHP-DKAHSU YNTFAQYSCTLQNLTGTDGGL YDSYFECLYNISDREKM YDSYFECLYNISDREKM YDSYFECLYNISDRREYE YDSYFELTKKSSMREJTT YDSFALJERRG0.	IT ON D MATLISKEP TIP EKI KNST MIMILSRTAQUE EETYKD FRAILARNP NUP PET VOSI IWILT FOREPS AEA VOAI IWILT FOREPS AEA VOAI IWILT FOREPS AEA VOAI SALLLSRNT SLADNE I SPI ISVILL RSRNSSING LERY ISVILL RSRNSSING IERY NOQCLART PEVD BEALEXI MOQCLART PEVD BEALEXI MOQCLART PEVD BEALEXI MOQCLART FUD BEALEXI MOQCLART FUD BEALEXI MOQTLARV I VED BEALEXI MOQLARY IS BOD MILTOPETRRI MOQLARY IS BOD MILTOPETRRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPETRI MOLTOPET	VLE LAK SVGYDI TK-L LUE KAVEEGYDI SK-L LUNILT SNNLDVKK-M ARKILE DND VSQ AF-L WID VLKRL SLDP ER-L URGKLP QD-IDT SF-M VED ELKNLQ LDL INSY K VDNVLKTF SMLI DAS F FUKALKALP MHI RLSN FTR AFRPL P HLWY-D FKKLLR LAMPYTV-L FTR FSK SLGLPENH-V FTR FSK SLGLPENH-V FRO RVGEMLTEDO-F FRKTLS DHGF SE TD-Y.	intron E IVISU NENEN HKTPQ SDT PPE SNT IVITOQ VAC PKL S IDIVVQ NC PKL DCN III SKNQC PKAL FNIGQ ESC DNL MES ITE QSANYC A MSNEF SKRAACOYST UKENGC QSR DSR PTQLE SDC HI LKQME SDC RF LKQME SDC RF FL POP DC LOE FL POP DC LOE FL POP DC LOE FL POP INC LOE FN PRICE CYPG DRE IL KHULTC VKVLQS	APEDSKGW/ GTGLAGEDG SRDDPLAYV TYSLTGPDR NSL VEPTSLAR- VEPTSLAR- NUDEVR- AAESRP-
Ddis. Lip Atha. OM Hsap. ApoD Jmel. Naz Imel. GLaz Imel. GLaz Imel. GLaz Imel. Karl Same. Laz Msex. LcyA Rnor. RBP Btau. BL Btau. BL Hsap. GlyC Msap. C&CC Rnor. ERP Ggal. QS-21	intron C -ECMLKVSFF	IT IT IT IT IT IT IT IT IT IT	NTON D MANILSKEPTIPEKI KNST MANILSKEAQUEETYKDI FRAILARNPALPETVDJ FRAILARNPALPETVDJ FRAILARNPALPETVDJ IVILLGRORDFEVDIRSKT SALLSSRT SLADDREISEJ IVILLGRORDFEVDIRSKT KOVILGRORDFEVDIR KANVLSRNKVLEGATKEV YSEVESROPNLTPETRRI AQQLIVERRUVKODELLAGI HANNVRTE-ENLIJMEI KATINGRAQLLKEELKEN FAKINGREPOLRDSLLOEJ SVKINGRSLDVSDSULSGI VARLSSRSTUNGRELYN	YLE LAK SVGYDI TK-L LUE KAVERGYDI SK-L LKNILT SNNIDVKK-M RKLILE DND VSQ RF-L UVD VLKRL SLDP ER-L LKGKLPQD - DJT SF-M VD VLKRL SLDP ER-L LKGKLQ DLD. NSYK VDN VLKTF SHLI DAS F LURQRQ EKL CLE RQ-R FDKALKALP MHI RLSN FTK SKSLGLPENH-V FTR FSKSLGLPENH-V FKD VALNVGISENS-D FFQLTS DKES FTD-Y FFKLTS SHGFS FTD-Y FKKLARERNYTD EM-A	intron E INSEQUENCESSION INTEGESCONSECTION INTEGESCONSECTION INTEGESCONSECTION INSECTIONS INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECTION INSECT	APEDSKGWI GTGLAGEOG SRDDPLAYV TYSLTGPDR NSL VEPTSLAGEOD VEPTSLAR HVLDEVER ARE SRP
Ddis. Lip Atha. O.M. Hsap. Apol Dmel. NLaz Dmel. CLaz Dmel. Karl Same. Laz Msex. LyA Rnor. RBP Btau. BLB Hsap. Clyc Msug. BL Rnor. PCD S Hsap. C&CC Rnor. KBBP Ggal. QS-21 Mnus. MUP1	intron C -ECMLKVSFFCPFYSGYNU LKVKFYVPPFLPICPFYSGYNU RKLEVKFSMEDPSAPYWLLKID LAVAFYPTDPLTXRNTULGTD FTRFPFPDVIARLLPCSGXYQVLYTD GHUMT KUTYSCTYNTVLITD GKUSMFFNSFSVENWAULSTD RKFNKTWGVASELQRGDDHWILITD PAVFKLDLNENKVLVLDTD PKKFKLDLNENKVLVLDTD PKKFKLDLNENKVLVLDTD PGYTYNSPBGSFHSLSVVFTD PGEFTLGNLKSFP-GLTSYLVKWSTD GCKLYJNSSSNTILESYVVHIN LGRELLOBRGARENTYFUPKUDTD AKEQVTRLSE	IT IN DPDYKYAIAGQSRD VENYAVYSCTSUTQLEHVDJ YENYAVYSCTSUTQLEHVDJ YENYAVYSCTSUTQLANK YENYAVYSCTSUTQLANK YENFALWSCGSIGSLGHSDJ YDTYSVSCVACRAKKQPRYL YDNYSVSCVACRAKKQPRYL YDNYSVSCVACRAKKDISAA YDNYSVSCVACRAKKCODFRY YDNFLLCLONTTPTQSAD YDSYFFCLYNISDREKM YDSYFFCLYNISDREKM YDYNFLCLONTTP CODFRY YDYNFLCLONTTP YDSYFFCLYNISDR	It on D AMULS KEP TIP EKI KNST AMULS SETAQUE EETYKDI FRAULARNP NLP PET VOS J FRAULARNP NLP PET VOS J ILVILLGRORDFE VDI RSKT SALLS SRKT SLADNE I SF ILSVILL RSRNSTNATIERT MAVLS SRKVLE GOT KEV YSF VFS RDP NLL TPE TRRI MCQ TLARVL VKD DEI LOGI MCA HY RRI E-E NNSMNEI MGT INSEAQ LLKKEL KEN KIT INSEAQ LLKKEL KEN KIT INSEA LOBI SVKLYARSL PVS DSVL SGI INKLYS SSL DUNGEALYNI MGRINS SSL DDINGEALYNI MGRINS SREVS PTNAAT I LMGLNSREP DLS SDI KER	VLE LAK SVGYDI TK-L LUE KAVEEGYDI SK-L LKNILT SNNIDVKK-M ARKLIE END VSQ AF-L UYD VLKRL SLDP ER-L LURGKLP QD - DJT SF-M VED ELKNIQ LDL NKYK VID VLLKRL SLDP ER-L LURQ RQ EEL CLE RQ - R FDKALKALP MHI RLSN FTRAFREPLE MH. WY FTR FSKDQG JTE EN-V FTR FSKDQG JTE EN-V FTR FSKDQG JTE EN-V FTR VAL NVG TSENS-J FEQ VAL NVG TSENS-J FEQ VAL NVG TSENS-J FEQ VAL NGC SENS-J FEQL RAGENTTD EM-A FRALARERNTTD EM-A FRALARERNTTD EM-A FRALARERNTTD EM-A	intron E INSEQUENCESSION INTERNET STORES INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNET INTERNE	Intron F APEDSKGW7 GTGLAGEOG SRDDDLAYV TYSLTGPDR NSL VEPTSLAR WLDEVRR ARE SRP
Ddis. Lip Atha. OM. Hsap. ApoD Imel. NLaz Imel. NLaz Imel. Karl Same. Laz Msex. LyA Rnor. RBP Btau. BLB Hsap. Clyc Msap. CGJC Msap. NGAL Msap. CGC Rnor. ERBP Ggal. QS-21 Sscr. VEG	intron C -EGMLKVSFF	IRI DPDYKYAIAGQ SFD IXI DPDYQHAIGQP SFD IXI DPDYQHAIGQP SFD YENYAYSCTCI IQ LEHVD YENYAYSCTSVTP LEHVD YENYAKYCNACAEKKX QPRYL YDNYSVWSCVARP -D SAA STE YDNYSVWSCVARP -D SAA STE YDNYSVWSCVARP -D SKAHS YDTFAQYSCRLQNLDGTCADS YDNYSVHCLAVHD -D SKAHS YDNYSVHCLAVHD D SKAHS YDNYSLALLCLQDTTTP TQSM YDNFLALLCLQDTTTP TQSM YDNFLALLSSHH EKM YDFYAELKXSSHH ELMITI YQSFALYLERAG RURR YLTYAI DIT SLVAG RURR YNNFLAALI NEKKG FTFQ XDNFLALLI NEKKG FTFQ	ITTON D MANTLISKEP TIP EKI KNST MANTLISKEP TIP EKI KNST MANTLISKEP TIP EKI KNST MANTLISKEP SREA VDA DIVILLGRORDFE VDI RSKT SALLISRKT SLADNE ISF ISWILL SRKT SLADNE ISF ISWILL SRKT SLADNE ISF MANTLISKKVILE GATKEV SF VESROP NUL TPETRI MANTLISKKVILE GATKEV KATHVREI E – ENKJANEI MATINS RAQ LLKEEL KEKI KATHVREI E – ENKJANEI MATINS RAQ LLKEEL KEKI KATHVREI E – ENKJANEI MATINS RAD LLKEEL KEKI KATHVREI E – ENKJANEI KATHVREI E – ENKJANEI MARINS RAD LLKEEL KEKI KAKLVGED ENNPERLEEI	VLE LAK SVGYDI TK-L LUE KAVEEGYDI SK-L LUN LLT SNNIDVKK-M ARKLIE BNO USQ RF-L LWYD VLKRL SLDP ER-L LWR KLP QD - DIT SF-M VED BLKNLQ LDL NNYK WIN VLKTF SHLI DAS F LUR QRQ EEL CLE RQ-R FINAERPLP MILWY-D FUTF SKDQGLTE ED-V FITF SKDQGLTE ED-V FITF SKLQELTE ED-V FITF SKLQELTE ED-V FITF SKLQELTE ED-V FITF SKLQELTE ED-V FRO VAL NWG JSENS-L FED RUDEAHLTE QQ-F FRKUTS DHGF SETD-Y FRKLARERNYTD EM-A FRQ LCE KMG JLREN-U FKQ LCE KMG JLREN-U	intron E INSEQUENTED SUT PPE SWI TUTDO VAC PKL S III SKNNQC PKL S III SKNNQC PKL S III SKNNQC PKAL FNIGO SSC DNL MES ITE OSAKYC A ENDERSCH EL SOME SCAROYST MI KHNGTC OSRPSR PTOLEBICHI FL POP DACIOS FL POP DACIOS	intron F APEDSKGW/ GTGLAGEOG SRDDPLRYV TYSL/GPDR NSL 
Ddis, Lip Rtha. OM Hsap, ApoD Imel. NLaz Imel. Karl Same. Laz Isex. LcyA Rnor. RBP Btau. BLB Hsap. CLyC Meug. BL Rnor. PCD S Hsap. NCAL Imus. Almg Ggal. QS-21 Imus. MUP1 Scr. VEG Kaur. Aphr	intron C -EGMLKVSFFGPFYSGYNU: LKVKFYVPPELPIGPFYSGYNU: LKVKFYVPPELPISAPYVLLACDD -LAVAFYPTD:SAPYVLLACDD FTTREPFDVIARLPGGXYU/LTD GKJVMFYBDYBYGTYNTYVLDTD GKJVMFYBGYGY	IT I DPDYKYAIAGQ SFD LXI DPDYKAIAGQ SFD YENYAVYSCTSUTP LANES YENYAVYSCTSUTP LANES YENYAVYSCTSUTP LANES YENYAWYGCTSUTP LANES YENYAWYGCTSUTP LGNSQ YINYSUKSCURP DSKANSU YINYFULCLQDTTP IQSQ YDNFLLCLQDTTP IQSQ YDSYALFSKSIGH EXFEI YDSYALFSKSIGH EXFEI YDSYALFSKSG QUFR YNNFLAMLI NEKGC ETHQ SUKLFTNKNQDRGC QETN	NTON D NAVILLSKEPTIPEKIENSY NAVILLSKEAQUEETYNQI FRVILLARNDNLPPETVDSI FRVILLARNDNLPPETVDSI ILVILLSRKTSLADNEISEN ILVILLSRKTSLADNEISEN ILVILLSRKTSLADNEISEN FRVVLSRNKVLEQNTKEV TSF VESKOPNLLTPETREN MOQULVETPEVDBELLEG MQYLARVLVEDBELLEG MQYLARVLVEDBELLEG MQYLARVLVEDBELLEG MQYLARVLVEDBELLEG MQYLARVLVEDBELLEG MQYLARVLVEDBELLEG MQYLARVLVEDBELLEG MQYLARVLVEDBELLEG MQRLYSRSLDDNERALYN MQRLYSRSREVSPTAMAI LUGLYCHEPDLSSDIKER MQKLVEGDRANPENDERLEG MUVACUGNALTPEENELI	VLE LAK SVGYDI TK- L LVE KAVEEGYDI SK- L LUN LIT SNNLDVKK- M ARKILE DND VSQ AF- L WYD VLKRL SLDP RR- L LRGKLP QD - IDT SF- M VED ULKNLQ LDL INSY K VDN VLKTF SHLI DAS F FURALKALP MAI RLSN FTR AFRPLP RHLWY- D FTR FSK SLGLPENH- W FTR FSK SLGLPENH- W FRO VALNVGISENS- L FFOR VALNVGISENS- L FFOR VALNUGISENS- L FFR VARALTE DD-F FRKLTSDHGF SETD-Y FRKLARERNYTD EAL- A FAQ LCE KHGILREN- L FAQ VALNAGISL-NDD- V LVQ FANEKKIP VEN- L	intron E INSERVENTION STOPES ST TVTDQ STOPES ST IT STOD STOPES ST IT SCALL S IT SCALL S IT SCALL S IT SCALL S IT SCALL ST IT SCALL	intron F APEDSKGW/ GTGLAGEDG SRDDPLAYV TYSLTGPDR NSL VEPTSLAR- VEPTSLAR- NVLDEVRR- AAE SRP
Ddis. Lip Atha. O.M. Hsap. ApoD Jmel. N.Jaz Imel. GLaz Imel. GLaz Imel. Karl Same. Laz Msex. LoyA Rhor. RBP Btau. BLB Hsap. ClyC Msap. NGAL Mnus. Almg Hsap. C&CC Rnor. ERBP Ggal. Q5-21 Mnus. MUP 1 Sscr. VEG Maur. Aphr Hsap. OBD 2	intron C -ECMLKVSFF	IT DPDYKYAIAGQ SED LYI DPDYHAIGQP SED YENYAVYSCTSUTP LANES YENYAVYSCTSUTP LANES YENYAVYSCTSUTP LANES YENFALWSCGSIGS LGHSDG YDTWSVSCURP DSAASTE YNNYAWYSCTUUDGTCADS YNNYAWYSCNUDGTCADS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNNYALFSKELS YNYSYNYALFSKELS YNYSYNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYALFSKELS YNYAL	IT ON D AMULS KEP TIP EKI KNST AMULS SKTAQUE EETYKDI FRUILARNP NLP PETVDSI DIVILLGRORDFE VDI RSKT SALLSSRAT SLADDNE ISE I SALLSSRAT SLADDNE ISE I SALLSSRAT SLADDNE ISE I AMULS RNKVLE GAT KEV YSE VES ROP NGL TPETRRI ACQ CLURTP EVD DEALEKSI WCY LIA KVL VKO DETL MGG ACA HTWRRIE – E NNG MNE MAT INS GAQ LLKEKL KEN TAKLYS REP QLR DSL LOEI SVKLYARSLEVS DSVL SGI TAKLYS RST DONGEALENI TAKLYS RST DONGEALINGEN MARLYS RST DONGEALINGEN MARLYS RST DONGEALINGEN MARLYS RST DONGEALINGEN MARLYS RST DONGEALINGEN MARLYS RST DONGEALINGEN MAKLYS REP DLS SOLKER MAKLYS RNP NINNLE ALEI	YLE LAK SVGYDI TK-L LUE KAVERGYDI SK-L LKNILT SNNIDVKK-M ARKLIE BIN US SR = L UYD VLKRL SLDP ER-L LKGKLP QD - DJT SF - M VED ELKNLQ LDL NXYK UVR QR EKL CLE RQ - R FDK RLKALP MHI RL SN FTRAFFPL FRH.WF - D FTR STS SLG JEN H-V FTR SK SLG JEN H-V FTR SK SLG JEN H-V FKD VAL NVG I SEN S- D FFQ RVUEAHLTE QQ - F FRK I RS BHG FSE TD -Y FRK I RS BHG FSE TD -Y FRK I RS BHG FSE TD -Y FRK I RS RK G LLE RN-D FAKLARERNYTD EM - A FAQ LCE KHG JL REN-D EFK KLAKETD VR-L EFK KLVQHKGL SEED F	INTON E INSERVENEN HETPO SITPPE SIT ITVTDQ VINC PKL S IT SKINIC PEAL FNICO ESCINLINES TEOSANYC A MSNEF SEAACOYST MI ENDERCER EVERTRUCHV EVERTRUCHV EVERTRUCHV FNDDR: EC VPGDRE IL KHOLTCVKVLOS VI PSOESC SUBEV- DL SNANSCLOARE- RP QOS FC SPGCH- IL AND COS CVLEX TP LOT CSC VLEX	APEDSKGW/ GTGLAGEDG SRDDPLAYV TYSLTGPDR NSL VEPTSLAR- HVLDEVER- ARE SRP
Ddis, Lip Atha. O.M. Hsap, Apol Imel. NLaz Imel. NLaz Imel. Karl Same. Laz Msex. LyA Rnor. RBP Btau. BLB Hsap. ClyC Msug. BL Rnor. PCD S Hsap. NCAL Msus. Almg Hsap. C&CC Rnor. KBBP Ggal. QS-21 Sscr. VEG Maus. Aphr Sscr. VEG Maus. Aphr Ssap. OB 2 Hsap. alG2	intron C -ECMLKVSFFCPFYSGYNU LKVKFYVPPFLPSAPYNLLKID AKLEVKESMENPSAPYNLLKID LAVAFYPTDLTXANT.U.GTD FTRFPFPDVLARLLPGSGXYQVLYTD GHJMTKUTYSE	IT IAI DPDYKYAIAGQSRD IXI DPDYOHAIGQPSRD YENYAVYSCTCI IQLEHVDJ YENYAVYSCTCI IQLEHVDJ YENYAVYSCTSVTPLANKE YENYAVYSCVAC AEKKKQP RYL YDNYSVWSCVKRP-D SAA STE YNYAVYNCNYHD-D IXAHSID YDT FAQYSCRLQNLDGTCADS' YNYAYINCNYHD-D IXAHSID YDT FAQYSCRLQNLDGTCADS' YNKYLECMINSAEPEQSLJ YNFLLCLYNIS DREKM YDEYALFSKGTKCPGDFRN YDEYALFXKSSHHHELTT YQSFALYLERAGQLSY XINYALDIT SLVACEYTQI KDHYTCEGRLDGQKWR SDKLFTNKNBURGCETTQI KDHYTCEGRLDGQFNN WVFYKDQRBGCL	NTON D MANILS KEP TIP EKI KNST MMILS KEP TIP EKI KNST MMILS BETAQUE EETYKDJ FRAJLARNP NLP PET VOS J DIVILGRORDFE VDI RSKT SALLS EKT SLADNE I SF ISMILL SERT SLADNE I SF ISMILL SERT SLADNE I SF ISMILL SERT SLADNE I SF MAVUS SRKVLE GYN KEV YSF VFS KOP NL TPETREN MQ VILAWU VED DEL MQG MANY REI E-ENNGMEN MAT INS FAQ LLKERL KEN KIT INS FAQ LLKERL KEN SVKINARSL PVS DSVL SGJ TAKINS SSREVS PTAMATI MARINS SSREVS PTAMATI MARINS SSREVS PTAMATI MARINS KIT PENDIN KEN MARING SON MINNE ALEBI SKI VEND PINNER ALEBI MAN DEND PINNER ALEBI MSKI VEND PINNER ALEBI LSFYND KPETTKEQL GEFT	VLE LAK SVGYDI TK-L LVE KAVEEGYDI SK-L LKNLLT SNNIDVKK-M ARXLLE DNO VSQ AF -L WYD VLKRL SLDP ER-L UYD VLKRL SLDP ER-L LKR KLP QD - DIT SF-M VED ELKNLQ LDL NKYK WINVLLTF SMLL DAS F EVKQ RQ EELCLE RQ-R FUKALKALP MHU KUSN FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLP BHU.WY-D FTRAFRPLF BHU.WY-D FTRAFRPLF BHU.WY-D FTRAFRPLF BHU.WY-D FTRAFRPLF BHU.WY-D FTRAFRPLF BHU.WY-D FTRAFRLF BHU.FT FTRAFRLF BHU.FT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFTT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFT FTRAFTFTT	INTON E INSEQUENCESSION INTON SUPPESSION INTON SUPPESSION INTON SUPERIDES INTON SUPERIDES INTON SUPERIDES INTON SUPERIS INSUES SEARAOYST SUPERIS INTON SUPPERIS INTON SUPPERIS INTO SUP	APEDSKCW APEDSKCW GTGLAGEOG SRDDPLAY TYSLAGEOD NSL VEPTSLAGEOD VEPTSLAR NUDEVER AAESER HEKERKQEE

Fig. 3.—Alignment of the mature proteins of lipocalin representatives with known gene structure. The position and phase of the introns are mapped onto the alignment in the context of the overall secondary structure of an archetypal lipocalin ( $\beta$  strands are represented by white arrows, and  $\alpha$  helices by cylinders). Intron phase 0 is shown as a line between the split codons; phase 1 or 2 introns as open or shaded boxes around the amino acids presenting the split codon.

their cladistic information content, a normalized measure of how much a tree reduces uncertainty regarding phylogenetic relationships (Thorley, Wilkinson, and Charleston 1998), is 0.98 for the gene structure NJ tree, and 1.00 for the protein sequence ML tree.

We also analyzed the positional congruence of each lipocalin COG in the two source trees. The normalized congruence measure, called "explicitly agree" (EA) similarity (Estabrook 1992), is shown in figure 5A for each lipocalin, and the average EA similarity (the EA similarity of the trees) is 0.789. This measure reflects the high congruence of the topology of both trees, and suggests that both phylogenetic reconstructions are good estimates of the evolutionary history experienced by lipocalins. Following the strict nesting method (Adams 1972), we built a consensus tree (fig. 5B) that shows the commonalities between the two source trees. The consensus tree further corroborates the orthology of ApoDs to the arthropodan lipocalins, and the monophyletic relationship of other chordate lipocalins.

Thus, two sets of independent characters have produced the same phylogenetic relationships between extant lipocalins.

Phylogenetic Distribution of Intron Numbers Within the Lipocalin Family

Another finding revealed by the intron arrangement phylogeny is that lipocalins that have originated more recently contain more introns in their CDS. In figure 5*C* we mapped the number of exons onto an updated version of the ML-based lipocalin protein phylogeny (see Gutiérrez, Ganfornina, and Sánchez [2000] for clade ascription). The ancient unicellular eukaryotic and plant lipocalins are encoded by 2 exons; the arthropodan lipocalins by 4–5 exons; and the chordate lipocalins by 4–7 exons. Introns E and F are absent in nonchordate lipocalins, whereas introns A–D show much wider phylogenetic distributions.



FIG. 4.—Phylogenetic trees derived from lipocalin gene structure information. *A*, Maximum parsimony analysis based on the presenceabsence of introns in the gene ORF. *B*, Neighbor-Joining (NJ) phylogenetic reconstruction based on a distance matrix with gene structure data (intron presence, location, and phase) combined in a measure of gene structure similarity (see *Materials and Methods*). The NJ tree is shown both as a phylogram (left) and as a cladogram (right). All trees are rooted with the Dictyostelium lipocalin. The scale bar in the phylogram represents branch length (number of amino acid substitutions/site).

A first look at these data might suggest an evolutionary trend to gain introns. In this hypothesis, the origin of introns A and D could be placed early in eukaryotic evolution, introns B and C originated at the base of the metazoan lineage, and introns E and F appeared later during early chordate radiation. The acquisition of introns was accompanied by diverse intron losses in different branches, giving rise to the pattern observed today.

However, we have to be critical when interpreting these observations in the context of lipocalin introns origin and evolution. First, the current insufficient sampling of lipocalins outside the metazoan kingdom generates uncertainty about the very assumption of homology of introns A and D in Dictyostelium and Arabidopsis, respectively. Second, the set of metazoan lipocalins available encompasses only two phyla within the kingdom; any proposal about which set of introns was present in the common ancestor of all metazoans awaits confirmation coming from other phyla. A scenario with a set of four ancient introns and subsequent losses in different lineages (Fedorov et al. 2001; Roy et al. 2002) would be as probable as a scenario with fewer or no ancient introns and a prevalence of intron gain at preferred "hot spots" (or proto-splice sites; Dibb and Newman 1989; Logsdon 1998).

Nevertheless, the extensive sampling of lipocalins in the chordate phylum allows us to make a stronger case for the acquisition of introns E and F during early chordate radiation. Both introns are absent in all arthropod lipocalins and in ApoD, the lipocalin COG that branches off at the base of the chordate lipocalin subtree in our two independent phylogenetic reconstructions (figs. 4B and 5C). Therefore, intron gain within the chordate lineage is the most parsimonious explanation for the current distribution of introns E and F.



FIG. 5.—*A–B*, Comparisons of gene structure versus protein sequence phylogenies. *A*, Phylogenetic ML tree built upon the protein sequence alignment used to derive the gene structure matrices (see fig. 3), and rooted with the dictyostelid lipocalin. Explicitly agreed similarity values are shown for each lipocalin, according to the RadCon program (see *Materials and Methods*). *B*, Consensus tree obtained by a strict nesting method (Adams 1972), representing the commonality between the gene structure and the protein sequence trees. *C*, Updated (March 2002) maximum-likelihood tree based on the protein sequence of 148 lipocalins (see Ganfornina et al. [2000] for details on tree building) showing the number of introns present in the ORF of each lipocalin clade. LBP values are indicated in each node (see *Materials and Methods*). The tree was rooted with bacterial lipocalins. The scale bars represent branch length (number of amino acid substitutions/100 residues).

In summary, although many questions about the origin of lipocalin introns and their subsequent evolution remain unanswered, a combination of ancient and recent introns is the most plausible scenario. Our results show that, independent of their origin, the variations in gene structure can be used to reconstruct the history of descent of lipocalin genes.

#### N-termini Conservation versus C-Termini Variability?

It is remarkable that the introns specific to chordate lipocalins are located in the C-termini of the proteins, whereas introns in their N-terminal region are the most conserved in the family (see fig. 3). This polarity, also noticed by other researchers (Salier 2000), is related neither to a particular distribution of lipocalins length nor to a C-terminal–specific protein sequence variability. Rather, we propose it might be related to a propensity for intron gain/loss in this gene region. The analysis of the 3' region



FIG. 6.—A, Alignment of the nucleotide sequences coding for the Cterminal region of the ERBP genes from mouse (Mmus.ERBP) and rat (Rnor.ERBP). Stop codons are marked by gray boxes. Intron insertions are shown by vertical lines. A conserved TGA codon is present (underlined) in the 3'-UTR of the rat gene. B, Alignment of the protein sequences of two representative FABP and CRBP proteins with the lipocalin alignment of figure 3. Intron insertions of FABP and CRBP (intron phase shown as in fig. 3) are shown in the context of lipocalin are represented by a white box and the intron positions (not considering their phase) are highlighted with a thick black line.

of lipocalins bearing 5-6 introns reveals that several lipocalins of particular chordate lineages (e.g., PGDS, VEG, NGAL; data not shown) show introns in the 3'-UTR that are located 0–7 nucleotides away from the stop codon. These introns would be equivalent to intron F if they happened to be in the CDS. Any form of intron sliding (Stoltzfus et al. 1997), or any frameshifting mutation that moves the stop codon in the 3' direction, could include/ exclude a given intron in the gene CDS, generating an apparent intron gain/loss. Moreover, a puzzling case of Cterminus variability comes from the comparison of mouse and rat ERBP (fig. 6A). Intron F of mouse ERBP locates 9 nucleotides away from the stop codon. The rat ERBP gene has an insertion that accommodates a short exon and another intron (alternatively, the mouse ERBP could have experienced an equivalent deletion). Were it not for the existence of an in-frame stop codon in the short exon present in the rat ERBP gene, we would have a unique lipocalin with 7 introns.

In summary, the C-termini of chordate lipocalins show genomic plasticity, accommodating introns and mutations that modify the protein length. It is not known whether this genomic plasticity is causally related to a higher probability of intron gain, loss, or sliding in the 3' end of lipocalins, but this possibility is worth investigating.

Gene Structure and the Three-Dimensional Structure of Lipocalins and Calycins

We mapped the location of exon boundaries in the tertiary structure of lipocalins that belong to different phylogenetic clades (IcyA, RBP, BLB, NGAL, MUP, and ERBP). Most lipocalin introns are located in the boundaries of  $\beta$  strands (see fig. 3). In spite of a certain variability in number and position, introns A–D seem to demarcate the lipocalin  $\beta$  barrel, whereas introns E and F are present in the C-terminal flexible region.

A way of testing a relationship between lipocalin intron-exon boundaries and tertiary structure would be to analyze the gene structure of proteins with a tertiary structure like that of lipocalins. A similar structure and a marginal sequence similarity have been used to propose a structural superfamily, the calycins, that relates lipocalins to proteins such as FABP, CRBP, avidin, and a group of protease inhibitors (Flower, North, and Sansom 2000). We find no gene structure similarity after aligning representatives of these proteins with the lipocalins and comparing intron positions (fig. 6B). This finding suggests that (1) we do not have compelling evidence for a relationship between intron-exon arrangement and the tertiary structure of these  $\beta$  barrel-based proteins; (2) the evolutionary relationship of lipocalins with the other proposed calycins, already questioned after analyzing their protein sequence (Ganfornina et al. 2000), remains to be demonstrated; and (3) the homology of introns A-F in lipocalins, the foundation for the phylogenetic inferences that we present in this work, is a reasonable assumption: the pattern and properties of intron-exon boundaries are good markers of the lipocalins history of descent.

Concluding Remarks: Evolutionary Hypothesis for the Lipocalin Gene Family

In conclusion, gene structure is well preserved among lipocalins, and our results validate its use for the reconstruction of lipocalin evolution. The congruence of phylogenetic trees built from two independent sets of data (protein sequence and gene structure) increases the verisimilitude of both reconstructions of the lipocalins history. Furthermore, in the future we can use gene structure data to assay the lipocalin nature of novel proteins whose amino acid sequence and/or protein structure show similarity to lipocalins.

Our results give support to the following hypothesis about the evolutionary history of lipocalins: Bacterial lipocalins were inherited by unicellular eukaryotes and passed on to both plants and metazoans. The primitive metazoans spread a low number of ancient lipocalins into some of their successors, the arthropods and chordates, although these proteins might have been unexploited and subsequently lost in other phyla. The primordial arthropod and chordate lipocalins were likely similar to the Lazarillo and ApoD lipocalins now present in these phyla. Alongside the chordate radiation, the ApoD-like ancestral lipocalin suffered duplications. On the one hand, it gave rise to the ancestor of RBPs, and on the other hand, to one or more ancestors of all other paralogous groups of lipocalins that diverged into the current diverse catalog of chordate lipocalins.

# Acknowledgments

We thank C. González for allowing us to use his lab resources to collect the Lazarillo gene structure data needed for this work, and M. Bastiani for the grasshopper genomic DNA. We thank J. P. Salier for helpful discussion and comments on the manuscript. This work was supported by a grant (DGICYT BIO-1999-065-C02-02) to A. M., and a visiting professor grant to D.S. We are grateful to R. Escalante for alerting us to the sequence of the Dictyostelium gene, and to B. Akerstrom for alerting us to the Drosophila gene Karl. We thank C. Santiago at the Centro de Investigaciones Biológicas sequencing facility.

#### Literature Cited

- Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: programs in molecular phylogenetics based on maximum likelihood. The Institute of Statistical Mathematics, Tokyo.
- Adams, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool. 21:390–397.
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.
- Betts, M. J., R. Guigo, P. Agarwal, and R. B. Russell. 2001. Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J. 20:5354–5360.
- Bishop, R. E. 2000. The bacterial lipocalins. Biochim. Biophys. Acta 1482:73–83.
- Caterino, M. S., S. Cho, and F. A. Sperling. 2000. The current state of insect molecular systematics: a thriving tower of Babel. Annu. Rev. Entomol. 45:1–54.
- Dibb N. J., and A. J. Newman. 1989. Evidence that introns arose at proto-splice sites. EMBO J. 8:2015–2021.
- Estabrook, G. F. 1992. Evaluating undirected positional congruence of individual taxa between two estimates of the phylogenetic tree for a group of taxa. Syst. Biol. **41**:172–177.
- Fedorov, A., X. Cao, S. Saxonov, S. J. de Souza, S. W. Roy, and W. Gilbert. 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. 98:13177–13182.
- Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.
- Flower, D. R., A. C. T. North, and T. K. Attwood. 1993. Structure and sequence relationships in the lipocalins and related proteins. Protein Sci. 2:753–761.
- Flower D. R., A. C. North, and C. E. Sansom. 2000. The lipocalin protein family: structural and sequence overview. Biochim. Biophys. Acta 1482:9–24.
- Ganfornina, M. D., G. Gutiérrez, M. J. Bastiani, and D. Sánchez. 2000. A phylogenetic analysis of the lipocalin protein family. Mol. Biol. Evol. 17:114–126.
- Ganfornina, M. D., D. Sánchez, and M. J. Bastiani. 1995. Lazarillo, a new GPI-linked surface lipocalin, is restricted to a subset of neurons in the grasshopper embryo. Development **121**:123–134.
- Gutiérrez, G., M. D. Ganfornina, and D. Sánchez. 2000. Evolution of the lipocalin family as inferred from a protein sequence phylogeny. Biochim. Biophys. Acta 1482:35–45.
- Hasegawa, M., and H. Kishino. 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. Mol. Biol. Evol. 11:142–145.
- Holzfeind, P., and B. Redl. 1994. Structural organization of the gene encoding the human lipocalin tear prealbumin and synthesis of the recombinant protein in Escherichia coli. Gene 139:177–183.
- Igarashi, M., A. Nagata, H. Toh, Y. Urade, and O. Hayaishi. 1992. Structural organization of the gene for prostaglandin D synthase in the rat brain. Proc. Natl. Acad. Sci. USA 89:5376–5380.
- Krem, M. M., and E. Di Cera. 2001. Molecular markers of serine protease evolution. EMBO J. 20:3036–3045.

- Li, W., and L. M. Riddiford. 1992. Two distinct genes encode two major isoelectric forms of insecticyanin in the tobacco hornworm, *Manduca sexta*. Eur. J. Biochem. 205:491–499.
- Lindqvist, A., P. Rouet, J.-P. Salier, and B. Akerstrom. 1999. The alpha1-microglobulin/bikunin gene: characterization in mouse and evolution. Gene 234:329–336.
- Logsdon, J. M. Jr. 1998. The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. 8:637–648.
- Ohno, S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin. Cell Dev. Biol. **10**:517–522.
- Rokas, A., and P. W. Holland. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. **15**:454–459.
- Rokas, A., J. Kathirithamby, and P. W. Holland. 1999. Intron insertion as a phylogenetic character: the engrailed homeobox of Strepsiptera does not indicate affinity with Diptera. Insect Mol. Biol. 8:527–530.
- Roy, S. W., A. Fedorov, and W. Gilbert. 2002. The signal of ancient introns is obscured by intron density and homolog number. Proc. Natl. Acad. Sci. USA 99:15513–15517
- Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
- Salier, J.-P. 2000. Chromosomal location, exon/intron organization and evolution of lipocalin genes. Biochim. Biophys. Acta 1482:25–34.
- Sánchez, D., M. D. Ganfornina, and M. J. Bastiani. 1995. Developmental expression of the lipocalin Lazarillo and its role in axonal pathfinding in the grasshopper embryo. Development 121:135–147.
- 2000a. Lazarillo, a neuronal lipocalin in grasshoppers with a role in axon guidance. Biochim. Biophys. Acta 1482:102–109.
- Sánchez, D., M. D. Ganfornina, S. Torres-Schumann, S. D. Speese, J. M. Lora, and M. J. Bastiani. 2000b. Characterization of two novel lipocalins expressed in the Drosophila embryonic nervous system. Int. J. Dev. Biol. 44:349–359.
- Stoltzfus, A., J. M. Logsdon, Jr., J. D. Palmer, and W. F. Doolittle. 1997. Intron "sliding" and the diversity of intron positions. Proc. Natl. Acad. Sci. USA 94:10739–10744.
- Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Suderland, Mass.
- Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tool. Nucleic Acids Res. 24:4876–4882.
- Thorley J. L., and R. D. M. Page. 2000. RadCon: phylogenetic tree comparison and consensus. Bioinformatics **16**:486–487.
- Thorley, J. L., M. Wilkinson, and M. A. Charleston. 1998. The information content of consensus trees. Pp. 91–98 in A. Rizzi, M. Vichi, and H.-H. Bock, eds. Advances in data science and classification. Springer-Verlag, Berlin.
- Toh, H., H. Kubodera, N. Nakajima, T. Sekiya, N. Eguchi, T. Tanaka, Y. Urade, and O. Hayaishi. 1996. Glutathione-independent prostaglandin D synthase as a lead molecule for designing new functional proteins. Protein Eng. 9:1067–1082.
- Wada, H., M. Kobayashi, R. Sato, N. Satoh, H. Miyasaka, and Y. Shirayama. 2002. Dynamic insertion-deletion of introns in deuterostome EF-1 alpha genes. J. Mol. Evol. 54:118–128.

Peer Bork, Associate Editor

Accepted December 23, 2002