

# The rhomboids: a nearly ubiquitous family of intramembrane serine proteases that probably evolved by multiple ancient horizontal gene transfers

Eugene V Koonin\*, Kira S Makarova\*, Igor B Rogozin\*, Laetitia Davidovic<sup>†</sup>, Marie-Claude Letellier<sup>†</sup> and Luca Pellegrini<sup>†</sup>

Addresses: \*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. <sup>†</sup>Centre de Recherche Université Laval Robert Giffard, Université Laval, Chemin de la Canardiere, G1J 2G3 Quebec, Canada.

Correspondence: Luca Pellegrini. E-mail: Luca.Pellegrini@crulrg.ulaval.ca

Published: 28 February 2003

*Genome Biology* 2003, 4:R19

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/4/3/R19>

Received: 30 September 2002

Revised: 20 December 2002

Accepted: 3 February 2003

A previous version of this manuscript was made available before peer review at <http://genomebiology.com/2002/3/11/preprint/0010> (*Genome Biology* 2002, 3(11):preprint0010.1-0010.26)

© 2003 Koonin et al.; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

## Abstract

**Background:** The rhomboid family of polytopic membrane proteins shows a level of evolutionary conservation unique among membrane proteins. They are present in nearly all the sequenced genomes of archaea, bacteria and eukaryotes, with the exception of several species with small genomes. On the basis of experimental studies with the developmental regulator rhomboid from *Drosophila* and the AarA protein from the bacterium *Providencia stuartii*, the rhomboids are thought to be intramembrane serine proteases whose signaling function is conserved in eukaryotes and prokaryotes.

**Results:** Phylogenetic tree analysis carried out using several independent methods for tree constructions and the corresponding statistical tests suggests that, despite its broad distribution in all three superkingdoms, the rhomboid family was not present in the last universal common ancestor of extant life forms. Instead, we propose that rhomboids evolved in bacteria and have been acquired by archaea and eukaryotes through several independent horizontal gene transfers. In eukaryotes, two distinct, ancient acquisitions apparently gave rise to the two major subfamilies, typified by rhomboid and PARL (presenilins-associated rhomboid-like protein), respectively. Subsequent evolution of the rhomboid family in eukaryotes proceeded by multiple duplications and functional diversification through the addition of extra transmembrane helices and other domains in different orientations relative to the conserved core that harbors the protease activity.

**Conclusions:** Although the near-universal presence of the rhomboid family in bacteria, archaea and eukaryotes appears to suggest that this protein is part of the heritage of the last universal common ancestor, phylogenetic tree analysis indicates a likely bacterial origin with subsequent dissemination by horizontal gene transfer. This emphasizes the importance of explicit phylogenetic analysis for the reconstruction of ancestral life forms. A hypothetical scenario for the origin of intracellular membrane proteases from membrane transporters is proposed.

## Background

Polytopic transmembrane proteins are, in general, not particularly strongly conserved during evolution. Inspection of the database of Clusters of Orthologous Groups of proteins (COGs) [1] revealed only one family of such proteins that is represented in most of the sequenced bacterial, archaeal and eukaryotic genomes. The prototype of this family is the rhomboid (RHO) protein from *Drosophila melanogaster*, a developmental regulator involved in epidermal growth factor (EGF)-dependent signaling pathways [2-4]. Not only were homologs of rhomboid detected in prokaryotes and eukaryotes, but the pattern of sequence conservation in this family appeared uncharacteristic of nonenzymatic membrane proteins, such as transporters [5,6]. Specifically, several polar amino-acid residues are conserved in nearly all members of the rhomboid family, suggesting the possibility of an enzymatic activity. As three of these conserved residues were histidines, it has been hypothesized that rhomboid-family proteins could function as metal-dependent membrane proteases [5,6]. Recently, however, it has been shown that RHO cleaves a transmembrane helix (TMH) in the membrane-bound precursor of the TGF $\alpha$ -like growth factor Spitz, enabling the released Spitz to activate the EGF receptor, and that a conserved serine and a conserved histidine in RHO are essential for this cleavage [7,8]. Thus, it appears that rhomboid-family proteins are a distinct group of intramembrane serine proteases. Altogether, the genome of *Drosophila* encodes seven RHO paralogs (now designated RHO1-7, with the original rhomboid becoming RHO-1), at least three of which are involved in distinct EGF-dependent pathways, apparently through proteolytic activation of diverse ligands of the EGF receptor [9,10].

The newly discovered intramembrane proteolytic activity of RHO places the rhomboid family within the framework of regulated intramembrane proteolysis (RIP), a new paradigm of signal transduction, which appears to be prominent in all forms of life [11,12]. Under RIP, signaling proteins undergo site-specific proteolysis within TMH, resulting in the release of active fragments, which are the actual effectors in signal transduction cascades. Until recently, the only characterized cases of RIP in eukaryotes involved presenilin-1, an aspartyl protease, which cleaves a transmembrane helix in type-1 membrane proteins such as amyloid  $\beta$ -precursor protein (A $\beta$ PP), Notch and Ire1 [13], and the metalloprotease S2P, which cleaves a TMH in a type-2 transmembrane protein, the sterol-dependent transcription factor SREBP [11]. Notably, S2P has highly conserved bacterial homologs, and the protease domain of presenilins also might be homologous to bacterial and archaeal type IV prepilin peptidases, although, in this case, the sequence similarity is low [14,15].

In the case of the rhomboid family, the existence of homologs of RHO in most prokaryotes is particularly remarkable because animal RHO proteins are involved in signaling pathways that are not found outside metazoa,

which seems to make functional conservation in prokaryotes a remote possibility. The only prokaryotic protein of the rhomboid family that has been characterized experimentally in considerable detail is AarA from the bacterium *Providencia stuartii* [16,17]. This protein is involved in the export of a quorum-sensing peptide, a function that, in physiological terms, resembles that of RHO, although the signaling molecules, other than RHO and AarA, are obviously unrelated [18]. In a striking recent development, two independent research groups have shown that several bacterial rhomboid-family proteins, including AarA, can cleave the EGF receptor ligands (Spitz, Keren and Gurken) that are normally cleaved by RHO paralogs [19,20]. The cleavage depended on the conserved serine and histidine residues [19] and, moreover, transgenic flies that expressed AarA developed a phenotype indistinguishable from that induced by overexpression of RHO, whereas RHO could substitute for AarA in *Providencia stuartii* [20]. These unexpected findings demonstrated the conservation of a RIP mechanism producing extracellular signals in eukaryotes and prokaryotes. Eukaryotic rhomboid family proteins seem to show considerable functional variability; in particular, cross-talk might exist between different RIP pathways. A distinct representative of the rhomboid family has been shown to physically interact with presenilins 1 and 2, and was accordingly named presenilins-associated rhomboid-like protein (PARL) [6]. The yeast ortholog of PARL has been suggested to participate in the processing of cytochrome *c* peroxidase precursor during its import into the mitochondrion [21].

The near ubiquity of the rhomboid family among bacteria, archaea and eukaryotes, along with the remarkable functional conservation, suggests that a signaling mechanism mediated by rhomboids might have functioned already in the last common ancestor of all extant life forms, with subsequent loss in several lineages. To address this possibility, we performed a detailed phylogenetic analysis of the rhomboid family.

## Results and discussion

### Sequence and structural features and phyletic distribution of the rhomboid family

Although the sequence similarity between eukaryotic and prokaryotic rhomboid family proteins is relatively low (around 10-15% identity in the conserved region), the entire superfamily could be retrieved from the protein sequence databases within three iterations of the PSI-BLAST program with a high statistical significance and without any false positives. The conserved core of the rhomboid family consists of six conserved TMHs (Figure 1). The predicted catalytic serine is located in TMH5, whereas the predicted catalytic histidine is in TMH7; TMH3 contains two additional histidines and an asparagine, which are conserved in the great majority of the rhomboid-family proteins (Figure 1). The roles of these conserved residues are not known, but, given

the remarkable evolutionary conservation, it seems likely that they also contribute to catalysis; indeed, it has been shown that the conserved asparagine is required for the cleavage of Spitz by RHO [7].

When examining the multiple alignment of the rhomboid superfamily proteins, we noticed that several eukaryotic members appear to be inactivated proteases, as indicated by the loss of the predicted catalytic serine or histidine (Figure 1, and data not shown); these inactivated forms could be regulators of active rhomboid proteases. Several other proteins lack one or more of the conserved residues in TMH3; it remains unclear whether or not these are active proteases.

Bacterial and archaeal members of the rhomboid superfamily contain six TMH, whereas the eukaryotic members typically have an additional seventh TMH, which may be attached to the core either from the amino terminus or from the carboxy terminus as discussed below.

The phyletic distribution pattern of the rhomboid family shows that this intramembrane protease is extremely common in all three kingdoms of life, but is not necessarily essential for cell function. Rhomboids are missing in the microsporidian *Encephalitozoon cuniculi*, a eukaryotic intracellular parasite with a highly degraded genome, the archaea *Methanothermobacter thermoautotrophicus* and *Thermoplasma volcanium*, and several bacterial species, primarily parasites with small genomes but also species with moderately sized genomes, such as *Xylella fastidiosum* (see COG0705 at [22]). In two instances, a representative of the rhomboid family is present in only one of a pair of relatively close genomes (present in *T. acidophilum* but missing in *T. volcanium*; present in the spirochete *Treponema pallidum* but missing in the related bacterium *Borrelia burgdorferi*), which suggests relatively recent, repeated losses of this gene. Most of the prokaryotic species have a single gene coding for a rhomboid-family protein, although

some have two or three paralogs (see COG0705 [22]); in contrast, eukaryotes show expansion of the rhomboid family, with seven members in *Drosophila*, and as many as 13 in *Arabidopsis*.

### Phylogeny and evolutionary history of the rhomboid family

The multiple alignment of the 6-TMH core of the rhomboid family (Figure 1) was employed to construct a phylogenetic tree using the least-squares algorithm with subsequent optimization using the maximum likelihood (ML) method (see Materials and methods). Only the conserved regions including the TMH and short adjacent stretches shown in Figure 1 were used as the input for tree building, whereas the poorly conserved intervening regions were omitted to avoid noise from potentially misaligned residues (except for the Bayesian analysis, which used the complete alignment; see Materials and methods). The alignment used for phylogenetic reconstructions included 87 sequences and 149 aligned sites. The phylogenetic tree of the rhomboid family presents a complex and unexpected picture (Figure 2). Neither the eukaryotic nor the archaeal subsets of the family appear to form monophyletic clades. Instead, the eukaryotic rhomboids are split between two major subfamilies, which are positioned in the midst of different prokaryotic branches (Figure 2). The first subfamily, which includes six of the seven *Drosophila* rhomboids, clusters with a distinct prokaryotic assemblage, consisting primarily of Gram-positive bacteria as well as a subset of archaea; this clade is strongly supported by bootstrap analysis (Figure 2). The proteins in this group of eukaryotic rhomboids, which we designated the RHO subfamily, typically have an extra TMH added carboxy-terminally to the 6-TMH core; some of these proteins also contain EF-hand calcium-binding domains amino-terminally of the core (Figure 2).

The second eukaryotic subfamily, which we designated the PARL subfamily, after PARL, the human ortholog of

**Figure 1** (see figure on the next two pages)

Multiple alignment of the conserved core of the rhomboid family proteins. The alignment includes the majority of the detected rhomboid family proteins; some closely related sequences were omitted. Only the six conserved (predicted) transmembrane helices (TMH) and short surrounding regions are shown. The boundaries of the predicted TMH are indicated by gray shading and overline and they are numbered 1-6. The number of amino-acid residues in the omitted terminal and internal regions are indicated. The consensus shows amino-acid residues present in at least 90% of the aligned sequences; h stands for hydrophobic residues (A, C, I, L, V, M, F, Y, W in the single-letter amino-acid code) and s for small residues (G, A, S, D, N, V). The proposed catalytic serine (TMH4) and histidine (TMH6) as well as conserved residues in TMH2 with possible ancillary roles in catalysis are highlighted in color. The proteins are identified with the gene identification (GI) number from the nonredundant database and an abbreviated species name. Bacterial species are color-coded green, eukaryotic species blue and archaeal species yellow. Species name abbreviations: Aerpe, *Aeropyrum pernix*; Agrtu, *Agrobacterium tumefaciens*; Anoga, *Anopheles gambiae*; Arath, *Arabidopsis thaliana*; Arcfu, *Archaeoglobus fulgidus*; Bacsu, *Bacillus subtilis*; Brume, *Brucella melitensis*; Caeel, *Caenorhabditis elegans*; Caucr, *Caulobacter crescentus*; Chlte, *Chlorobium tepidum*; Cloac, *Clostridium acetobutylicum*; Corgl, *Corynebacterium glutamicum*; Deira, *Deinococcus radiodurans*; Dicdi, *Dictyostelium discoideum*; Drome, *Drosophila melanogaster*; Escco, *Escherichia coli*; Haein, *Haemophilus influenzae*; Halsp, *Halobacterium* sp.; Homsa, *Homo sapiens*; Lacla, *Lactococcus lactis*; Lisin, *Listeria innocua*; Metja, *Methanococcus jannaschii*; Metka, *Methanopyrus kandleri*; Metma, *Methanosarcina mazei*; Meslo, *Mesorhizobium loti*; Mycle, *Mycobacterium leprae*; Myctu, *Mycobacterium tuberculosis*; Neucr, *Neurospora crassa*; Noss, *Nostoc* sp.; Prost, *Providencia stuartii*; Pyrab, *Pyrococcus abyssi*; Pyrae, *Pyrobaculum aerophilum*; Ralso, *Ralstonia solanaceum*; Sacce, *Saccharomyces cerevisiae*; Schpo, *Schizosaccharomyces pombe*; Sinme, *Sinorhizobium meliloti*; Strco, *Streptomyces coelicolor*; Strpn, *Streptococcus pneumoniae*; Sulso, *Sulfolobus solfataricus*; Sulto, *Sulfolobus tokodaii*; Synsp, *Synechocystis* sp.; Theac, *Thermoplasma acidophilum*; Thema, *Thermotoga maritima*; Thete, *Thermus thermophilus*; Vibch, *Vibrio cholerae*; Xanca, *Xanthomonas campestris*; Xylfa, *Xylella fastidiosum*.

		TMH1	TMH2	TMH3
6325010	Sacce	17 LTTGLVVFLLTAYLLSFPFA	14 LQMSRLSLSPLIHLSPHLLFNVLAIWAPLNLFEET	4 YTGVFNLNSALPAGLILYLLGKLLY
19075999	Schpo	10 ILKLPWITQIITYYIAILVYA	21 RQLYEIIITVYLLSMLHIVFNFVSLPAMSQPEKK	5 CILVTVIPYLPFGIMHLLVYHFFLN
21593075	Arath	25 LFSVVVVCVGYLICLLTG	17 FQVYRFYTAIIFHGSLHLVFNMMALVPMGSELERI	6 LYLTVLLATNNAVLHLLIASLAGYN
19570079	Didci	39 ATKVISIICSILFALSIVAP	19 LDNRLIILSNFAHLSIHYIVNMIPTFLDLAK-LERL	1 FGTLKYPYLLFLFGIITNLCILRTY
18676811	Homsa	28 PVTATLALNIIWFFLNPKQ	15 KDWRLLLSPLHLADWHLYFNMAASMLWKGINLER	0 LGSRFVAVITAFSVLTCGVYLLLO
18401578	Arath	33 PVTASLLAANTLVVLRPAF	21 KDLKRLFLSAFYHVNEPFLVYNNMSLLWKGKLETS	0 MGSSEFASVMVFTLIGMVYLLLLA
11498616	Arcfu	133 ANNTVLIICITLFFISIVAP	17 AMPWQLITSMFLVHEFFVFFVNFVLLFPFGTELE	0 LGDRKYLEIFPVSGLAGNVGYIAYS
6321538	Sacce	143 KNLVYALLGINVAVFGLWQL	18 TSKIISIGSFAHQEFVHFGNMLLALWSFGTSLATM	0 LGSNFFSLYMNASITAGSLFSLWYP
11066250	Homsa	166 QRTVTGIIAANVLFVCLWRV	18 VLCSMPLSTFSHFSLFHMAANMYVLSFSSSIVNI	0 LGQEQFMVAVLSAGVINSFVSYLKG
17647867	Drome	145 KMPFAPILLCNLVAFAFMRV	18 VVCWPMFLSTFSHSYAMHLFANMYVMHSFANAASV	0 LGKEQFLAVLSAGVAGSLSMSVLYK
18394631	Arath	133 RDVVLGLVIANAGVFMVVRV	19 GRHLTIITSAFSDIIGHIVSNMIGLYPFGTSTARN	0 FGPQPLKLYLAGALGGSVYFLIHH
19112976	Schpo	117 IMVAVIVCLVNGVVFVHWDL	30 GRWTVLVVSIFSHQNLALHLLVNCVAIYSFSLIVVYK	0 FGVWKALSIVLGGVFGVNGVALQRM
21295914	Anoga	163 ERIFAPICALNVIVYGLWRI	18 AVCWPMFLSTFSHSYSLFHILANMYVLHSHSAAVA	0 LGREQFLGVVLSAGVIASFASHVFK
22327066	Arath	81 ANGIWFIIILINLIYGLADHF	15 PAWYQVATATFCHANWNHLSNLFVFLYIFGKLVEEE	0 EGNFGLWLSYLPFGVANLVSWLVL
7509358	Caeel	392 PWFYTWITTIQIFVCLLSLL	257 NQFYRLFTSLFVHAGVQHLALSLFFQYVVMKDLN	0 IASKRMALVYFASGITGNLASAIFV
13375799	Homsa	165 PFFTYWLTFFVHVIIITLVVIC	230 DQFYRLWLSLFLHAGVHVCLVSVVFMPTILRDLK	0 AGWHRIAIIFLPSGITGNLASAIFL
17647863	Drome	1246 PFFTYWINTQVVVLLISII	236 DQLYRLTSLFCMHAGILHLAITLIFQHLFLADLERL	0 IGTVRIAIIVYIMSGFAGNLTASAILV
15240744	Arath	55 SMLVPMFVANVAVFVVFVAMP	57 KEGWRLLTCIWLHAGVILHGANMLSLVFIGIRLEQQ	0 FGFVRIIVYVLSGIGGSVLSLFI
16944591	Neucr	161 PFFVYFPTVQIAVFAELV	56 NQWRFLITPMLHAGVVIHIGFNMLLQMTIGKEMERS	0 IGSIRFFIVYVYVAGVIFGVFMGGNFA
8923490	Homsa	61 PFFVVISISLAELAVFIYAV	26 EEARVRFISYMLVHAGVQHLNGLCMQVLVGIPLM	0 HKGLRVGLVYLAGVITAGSLASIFD
17647865	Drome	72 PWFILLMSFVQISLHWIASE	13 VEYWRLLTYMLLSDYHWSLNLICPQCFIGICLEVE	0 QGHWRLLAVVYVGGVAGSLANAWLQ
17647869	Drome	102 PWFILVLSIIIEIAIFAYDRY	26 LQVWRFFSYMFLHAGVHFLGNFVIVIQVFFGIPLEM	0 HGTRIGVIMYAGVAGSLGTSVDD
17864410	Homsa	98 PFFIILATLLEVLVFLVWGA	15 LQLWRFSLYALLHAGVHLGYNVLTQLLFGVPLEV	0 HGSLRTGVIYVAGVLAGSLGTSVDD
21264326	Drome	163 PFMFITVTLLEVAFFLYNGV	26 AQVWRLYTYIFMHAEGHGLHLNVVQLVGVPLEM	0 HGATRIGLVVYVAGVLAGSLVAVD
17933592	Drome	179 PLTMVLSIIEIIMPLVDVI	31 YEGWRVFSYMFVHVGIMHLMNLIQIFLIGIALELV	0 HHWRVGLVYLAGVLAGSMGTSVDD
17977674	Drome	168 PFFIILVTLVLELGFVYVHSV	24 HEIWRFLPYMVLHAGVHLGFLVAVQVFLGPLEM	0 HGSTRICACIYFSGVLAGSLGTSVDD
17553192	Caeel	174 PFMFLITIIQVGFPPFYWE	33 GEAWRFITSPMLHAGLNLHGLNVIQLLVGIPLEVA	0 HKIWRIGIYLVAVTSGVLSLQYAD
21297308	Anoga	157 PLFVILVTFVLEGFVYVHSL	24 QEVWRFLPYMVLHAGVHFLGNLIQLLVGIPLEM	0 HGSTRIGCVYVAGVLAGSLGTSVDD
3219925	Schpo	77 RSLVLSIIGINGVFLWLRWA	20 INMPSMIYASAFHQSQWHLFNMFVAFYFAPAIVDV	0 FGNQVYVAFYISLSLFSNVASLHSH
15218144	Arath	48 TWLVSFVLLQIVLFAVTMG	52 HEIWRLLTSPWLSGLFHLFNLGLSILFVGIYMEQQ	0 FGPLRIAVIYVLSGIMGKFAVFLV
15222545	Arath	153 RRTWNVLLAINVIMYIAQIA	18 QQLWRLATASVLANPMLMNCYSLNSIGPTAESL	0 GGPKRFLAVVLTSAVAKPIRLVLSG
15231701	Arath	14 ATSCIVTLCVSIWVFIQKKS	15 GHYWRMITSALSHISVHLVFNMSALWSLGV-VEQL	8 YYLHYTLVVLVPSGLVIGIYHLLI
18312405	Pyrae	15 PFTVKALVFINVAVFIYELL	16 SEPYRWVTHMFLHGGGLHIVGNMILVWVFGDNVED	0 YGHFRFLALYLMWGLAAAFVHYVAV
15789622	Halasp	94 AFLPLGVMVWTFVIQGIAP	22 EYVWTVVTSVFAHGGFSHIVLSIVLVYFFGPIVEDR	0 IGSKRFVALLPAGLILAGLQAVGAS
20093492	Metka	1 MSLTMLMFLNVLAVYLSVG	21 VHEPCLITVYFHLANLHLLFNMLGLLTFVQVLERV	0 LSTSEFVLLVYLSGLMGLAQATL
21226784	Metma	24 APSMIAIIFLCIVSFFLEMV	19 TRPWTLVYTFIHLHAGLHGLFFNMIVLVYFFGTALER	0 VGNKQLLGFPTAGLISLAIYFTFLS
14520881	Pyrae	28 TFSLMIITAVPIYEVIVGF	16 QQWRLTALIFLHMGVHFALNFAVWFLVYGLTDLGI	0 VGTKRFLIVFASALAGNVLSPFTL
14601690	Aerpe	19 PIVNMSIIALNFAAFIVGLT	29 ERLYTVFTSMFLHGSWAHILGNMILYVIFGDNIESI	0 LGRARYIILYISGSLGAVVPHIASI
15669882	Metja	1 -MINILIVGICIAMPIISVF	16 NMPQVITSIIFMHAGLTHLLVNLVLFIFGTYLENI	0 VGSKKYLIIFLPSGIGNLAYIAYAA
15790000	Halasp	96 GFWPGTLLVAGIVAGFYTLV	18 AYPGLVLTSPIAHANLGHVTVNLIIGTLALAPVAEYR	7 RGTAAFGSWRTPNYPYRAGVVPVAG
15897391	Sulso	35 TFFMLFLVTLGPMVGLLATF	18 GYSELFTSIFITNSFVDFIFNFISLVYIYLVIFGSR	0 ACKGHE-GIFLAGLGNLTLVFTV
15920355	Sulto	28 TVLTLITLITIGYIIGQLLSL	18 GFYQVLTSTFVTPNFDWAFNTIAMFYIYVLYKGE	0 ACKGLE-IIFLAGLIGVINSLSLYLY
16081803	Theac	2 FLELAFFFLLGLYISSYPGA	7 RTFVGLYTSIFIVDGSNVEYFLIFAILFSAANISH	6 KRTAVALLASVLSGSIANLDDLALF
15598282	Pseae	85 SPMTAAVLLTFVVAAVTYL	33 QQWRLTAPMLIFHGWLHLAMNMFVWELGRRRIEFR	0 QGRPMLGLLTLFGLVSNVQYAVS
17549219	Raliso	1 --MISSLILANVIVFAELF	24 FSPWQLLYTAFHLSVPLHVFNMFGMFMFRGDRVERA	0 LGRVTRGVLLVLSASLAAFTQMAV
17549744	Raliso	205 PHLTHALIANLVLAWLATLV	26 GEWRLLSATFLHAGVHLAVNMIGLYAAGVTVERI	0 YGPVATLTYLGLGALLSALSLSL
17987022	Brume	17 VIALIGLCVAVVYVQNYLS	27 AVIFTFIISYFPMHGSFAHIAVNMILWAAFGSPLAGR	0 IGAVRMLLFWVPTVSVAGLTHYALH
19553712	Corgl	45 VRTGLTIAIGYVYVIAVHL	23 SALWGHFTSPLHSGFSHILIGNTVPGFIFSPGLGMS	3 VFWEVTTIAGLIGVLTGTFVGGITG
20806909	Thete	14 PVITLSLIIINSLIFFTLSS	32 SNLYPFIITSMFLHGNTFHLISNMWLLWVFGDNVEDR	0 MGHIRFLIYVLLSAGVLAGVHLVFN
21220616	Stroco	39 LCLLFLISPAAGLNPVYGT	27 GSALTPTAPALFVHSGVWHLHGNMFLYVFGAMTEER	0 MGRQLFALFVYLCGVYLAGVYAGAN
21222264	Stroco	84 HLVTKILIGINAVVPIAVQA	28 GEWYRLVTMTFHEEIIHIGFNMSISLWFLGGPLEAA	0 LGRARYALVYVSLGAGSLVAYLLA
21224370	Stroco	135 ANVLVFLFTPGMAGSAGDG	54 SPELSVLTAMFLHGGWHLHGNMFLWVIFGNNVEDR	0 MGHVFPFLVYVCGYAATYFALLD
21229496	Xanca	13 PRWAVPLLFAAVMLAYLSWI	33 GSVLRLTALFLHADWSHLHGNMFLVLLIFGLPAERI	0 LGPWRLLLFLPGLGGAANLAAIFP
21230863	Xanca	1 -MITLILIAITIGVSWMAFN	18 KQYDRLITVYGFHADLGHVFNMIITLFFFGRIEDV	0 MTRLLTGSVLTYPFLYFALVLSILP
21233650	Xanca	140 SRVLRAPNLSLAVALLVAV	19 DGLIGLITLAPLHAGVHLHGNMFLVLLIFGLTAGSV	3 ATAMALPLLWLSGSLGAGVYVLS
21675030	Chlita	21 PAALKAIIITNVIIVFLPQNS	24 FHLWQPIITLFLHGSFAHIFNMFALWVMEGVEIENY	0 WGTNRVFSYFVICIGIAGLNLINLAT
15608477	Prost	21 IALTLTLVLLNIAVYFYQIV	25 GDWWRPISIMMLHNSGTHLAFNCLALFVIGIGCERA	0 YGKFKLLAIYIISGIIAALFSAVYQ
13470470	Meslo	16 VLAIVIGICAAVFLQYVVLN	26 FLTRPPTVYAFMHGGFAHIAINMVLWAAFGSPLANR	0 LGGLRFALFVAVTGLASVALFWAMH
13473011	Meslo	17 QVVTIGLIVVNAVLYVCATAL	33 PELSLSYTLVSHLADIFHLGNNMFLWVFGDNVEDA	0 LGHIRYLIYFYLCAAGAAFPQGLVA
15606530	Aquae	14 PIVNLSIIVACSLIWLVEYS	31 QKPYTLTLTSPFLHSGVSGWGHIGNMFLWVFGDNVEDK	0 LKGFYRIIFYLCCAGLALQTFIS
15607252	Myctu	37 PVVYTYLISLNLAVFVQMVT	17 QTYRLVTSFALFHYGAMHLLNMMWALYVVGPPLEMW	0 LGRLRFALYVAVLGGVSLVYLLIA
15608477	Myctu	37 VGGTITLTFVALLYLVELI	18 DGLWGVIFAPLLHANWHLMANTIPLLVGLFMTLA	3 RFVWATAIWILGGLGTWLVGNVVS
15639966	Trepa	13 TNVTLSLVLANGAVFVITSL	18 RMYWQIFTYQFVHSGVWHLHGNMFLVGLVFFGQTEKK	0 MGSSEMLLFLVYLVGLTCGACAAAY
15640131	Vibch	97 GVFTLIFMALCIIIFTLQTF	19 WQIWRVWVSHALLHFSVMHIAFNLLWVWQFGDLEQR	0 LGSVRLIKLFPVVAISISGAGQYVE
15641983	Vibch	32 LGTITGHDVNLVYLLLLAISL	33 GQWWRILTNFARHTNFAHWAMNLAALWIIISFVKPT	0 ARQLLIPLLLSLAVGVMLIASDMQ
15643350	Thema	3 KRAVYFILLFNAPFIVMMFT	29 GDWFRILTALFVHGGILHIFLNSYALYVYGLIVEDI	0 YGTEKFLVGYFTPTVGNLATHVFT
15643845	Thema	14 PVVTIALILINNVVVFVYELM	30 FSLLPFITSMFLHGGFVHILGNMFLWVIFGDNTEDE	0 MGHVGYTLFVLSAGIFAALQTFVFT
15672152	Lacla	15 ATYILSIIITLVLVWVQFFTY	25 SQMWRFLTALFHIGIWAHVLNVAITLFFPIGRQIENV	0 FGWLRFTLIVLSSGIFGNAMVFLLT
15803931	Escco	94 GPVTVWVMIAACVVVFIAMQI	19 FEFWRVFTHALMHFSLMHLIFNLLWWWVYGGAVEKR	0 LSGSKLIVITLISALLSGYVQKFS
15806990	Deira	50 VKAAAGVTAAGLIALWQGEV	20 GTFWHVHTAFPLHAGVPHLIANTVPLAVLAPMTAVR	3 RFLVATPLIALIGGGLVWLIGRSGS
15827590	Mycele	36 MVGGVTILTFMALLYLVELI	18 DVLWGISFAPVLANHQQHLVANTIPLLVGLFIALA	3 RFIWVTAMWVIFGGSATWLVGNMMS
15837251	Xylfa	10 PTVTKGLLLTNVVPFLPQMM	27 FMPWQLLYTGLFHEGFGHLFFNMLAVFMFGAALEHT	0 WGEKRFLTYLVVAVAGVQVQLLVS
15837656	Xylfa	9 LMAVAPLFFAVLIAFLPWSI	33 GSALRLTALFLHADWHLHGNMFLVLLIFGLPAERI	0 LGSWRLLLFLPGLGALANLAAVLT
15838777	Xylfa	4 MLITLILIAMNAVVSWSLSPN	18 RQYDRLITVYGFVHANISHLHFNMTVLYFFGSMIEAV	0 MGEITLGSLLTYPLFYLGAHLSLIP
15889057	Agрту	32 LVGILAAALAIYVVPAYLLS	27 EWLWTPVTVYFSLHGGIEHILFNLGLWMAFGAPVLR	0 IGTVRVLLWCISAASAFHGAALN
15891346	Agрту	36 QVVTIGLIVINLVVWLPTGV	34 PDDLIVVTYAFHLHDFWHLHGNMFLWVFGDNVEDA	0 LGHFRFLIYFVLCIAAGALPHGVFA
15894241	Cloac	141 MRVTWILVIVINFIVYGISAW	26 QYYRLITCMFLHAGITHIGANMYSLSYMSGYMLENI	0 YGKLYRYTAIYFISGITASFPYSIFS
15903945	Strpn	12 VTSFFLLVTLVFLMLVTA	25 EQVWRLLSAIFVHIGWEHFIVNMSLSYLVYGRQVEE	0 FGSKQFPFLYLLSGMMGNLFPVFSV
15966395	Simme	17 QVVTITLVIDFVAWLAIGP	34 PDEFITVTVSFLHGGFVHILGNMFLWVFGDNVEDA	0 LGHFRFLIYFVLCIAAGALHGLLE
16077528	Bacsu	15 YPVVTFILALQAVLWLFPSL	21 GEWWRLLTPLLHAGFTHLLFNSMSIFLFPALPAMER	0 LKGRFLLVYVAGSGIIGNIQTVYTE
16079543	Bacsu	177 PFTTYLFIALQILMFSALLEI	23 GEWWRLLTPLLHIGIAHLAFNTLALWSVGTAVERM	0 YGSGRFLIYVLAAGITGSLIASFVS
16126863	Caucr	12 NAPWALLVAAAVIIPHLLL	20 GRWTVLVVHVGWVHAINMNAFGLAFGAPVSRV	0 LGLNVGGGIFCLFVYVCGVIAGVG
16272560	Haein	9 GKITLITLALCVLIYLAQQL	19 SEVWRYISHTLVHLSNLHILFNLVWVYVIFGMIERT	0 FGSVKLLVYVAVASAIYVQYVNVVS
16332120	Synsp	13 LQSQFSIIVSFLAIPWLEI	20 EGLRGIVFAPLHADFGHLIANSVPPVVLAWLWMLQ	3 DFVIVTITIMVWVGLVWLIAPPNT
16800442	Lislin	182 PVTYTSPIGLIVAAAPLVVTF	23 GEWWRFLSPILHSGSLIHLASNAVMLYIVGAWAERI	0 YGKWRYLILILLGGICGNIASFALN
17231423	Nosspp	14 PFFTYGLIGMNVLVFLHEVS	25 GEWPTLTPVLLHGGWHLISNMVFLWVFGNNIEER	0 LGHKFYLIYVLCAGLAAALQVQV
17232329	Nosspp	14 PVTYGLIAANILAPLYEAN	33 PEWATLITSQFLHGGFLHLAGNMLLWVFGNNVEEK	0 LGHARYLLFVYLCAGLILASLQWYFS
consensus/90%		.....h...hhh..h...	.....hh...h.H.sh.HhhhN.h..h.hs..ht..	.....hhh..shhs..h..h..

Figure 1 (see legend on the previous page)

	TMH4	TMH5	TMH6
6325010 Sacce	4 VAGASGCWCFLLFAYYSFKESQI	9 DYSIPTLYTPLVLLVAIAVVI	2 SSWGFHFFGLCVGYAIGYKESWF
19075999 Schpo	6 IAGLSCWAFAFISASCVHSPQR	6 LFSIPAYCFPIIYLLIMTTIIV	2 ASFIGHASGAVMVGCTPMLGSI
21593075 Arath	12 AIGFSGILFMSMIVETSLSGVT	6 LFNVPKLYPWILLIVQQLLM	2 VSLGLHLCGLISGFSYSYGLRNF
19570079 Diodi	8 HLGFSGLVLFALYIESNSSGRD	5 AVKIPSKLYPWAMLLAHVVFV	2 SSFIGHFSGIVVGLFKYGLDI
18676811 Homsa	16 AVGFSGLVLFALKVLNHHYCPGG	5 GFPVFNRFACWVLLVAIHLFS	2 TSFAGHLGAILVGLMVTQGPLKK
18401578 Arath	17 AVGFSGLVLFAMKVVLNQABEDY	4 GILVPTKYAAWAEILLVQMFV	2 ASFLGHLGGILLAGIYIYKLGKSY
11498616 Arcfu	8 ALGASAAIFGVMGCLAIITAPEI	8 IPINIRTALLFAAYDFWMMV	10 VANIAHLAGLAVGLYGGKRLGR
6321538 Sacce	10 SLGASGALFGVLGCFYSYLFPHA	5 VFPVPGGAWVFLASVAVNAA	8 FDYAAHLGGSMGVLGYWYISKA
11066250 Homsa	8 SLGASGAIMTVLAAVCTKIPEG	6 LPMPTFTAGNALKAIAMDTA	8 FDHAAHLGGALFGIYVYTYGHEL
17647867 Drome	8 SLGASGAIMTVLAAVCTKIPEG	6 LPALTFSSAGAGIKVLMGIDFA	8 FDHAAHLGGAMFGIYVYTYGQA
18394631 Arath	22 GLGASGAVNAIMLLDIFLHPRA	6 FIPVPAMLLGIFLIGKDLIRI	6 ISGSAHLGGAAVAA-IAWARIRK
19112976 Schpo	60 LLGASGAVYATAAIFACLFPPYT	4 FFVYPVKAGIFMPLDFIAEYV	11 VAFDAHVSGTFFGVVSSLLFLPA
21295914 Anoga	8 SLGASGAIMGILAYVCSQYPTD	6 LPMYTFSSAGAAIKVIMGIDLA	8 FDHAAHLGGALFGLWFCHFGSQN
23270666 Arath	5 SVGASGAVFGLFAISVLVKMSW	8 LILGQFVIERVMEAAQASAGL	12 VNHAHLGSGALVGVVLLWLLSKF
7509358 Caeel	4 AVGFSQAQCGILAAVIVECCDN	8 WALVQHILVITLVLICGFIPIW	0 -VDNWAHLGFTIFGLLTTIIPFY
13375799 Homsa	4 EVGPAGSQGGLLACLFLVELFQS	8 KAFNLNSAIVLFLICGLLWP	0 -IDNIAHIFGFLSGLLAFAPLFPY
17647863 Drome	4 EVGFSASLSGVVLAASLIALLVWF	9 IALFKLLLLCSVLVIGITLPEY	1 LNFGLLACVIGGLLTMSLVFP
15240744 Arath	4 SVGASGALFGLGSMLESEFTN	8 AALTLFLVILLNLAIGLPH	0 -VDNFAHVGGVTFGLLGLPILAR
16944591 Neucr	5 TTGASGALFGIITALLLDDLYS	8 KDLFLIGLDIVISFVLGLLPG	0 -LDNFAHIGGLIAGLGLGICVLQS
8923409 Homsa	4 LVGASGGVYALMGYFMNVLVN	10 FRLLIILIIIVLDMGFALYRR	9 VSFAAHIAAGGFAGMSIGYTVFSC
17647865 Drome	4 LMGASGAVYAMLGSHVPHVLVN	8 ARIASLLIILLLSDVGGTTYHF	9 TSLEAHILGGGVAGILCGFIVYRR
17647866 Drome	4 LVGASGGVYALAAHLANIITLN	8 TQLGSSVIVFVCDLGYALYQ	12 VSYIAHLTGALAGLTIKGLVLKN
17644410 Drome	4 LVGASGGVYALAAQLASLLLN	8 IQLMAVILFVFCDLGYALYSR	12 VSYIAHMTGALAGISVGLLLLRQ
21264326 Homsa	4 VVGSAGVYALVSAHLANIIMN	10 LRMAVALICMSMEFGRVWLR	11 PSFVAHLGGVAVGITLVGVVLR
17933592 Drome	4 LAGASGGVYALTAHIAIIMN	8 VQLLAFVFCFDTLGTSVYRH	7 IGYVAHLGSAVAGLVGLVIRLN
17977674 Drome	4 LVGASGGVYALAAHLANVLLN	8 IKLLHILVFSVDFGFAIYAR	25 VSYVAHLGAGIAGLTIKGLVLKS
17553192 Caeel	4 LVGASGAVYALFAHVAIVLILN	8 IRVVLVFPVIFLDFGGAIHRR	8 VSHLAHIAAGVATGLFFGVVLYN
21297308 Anoga	4 LVGASGGVYALAAHLANVMLN	8 LRLLAIFLFAACDVGFAIYSR	11 VSYVAHLTGALAGLTIKGLVLKN
3219925 Schpo	11 SLGASGAIYAAIAAATSYFFPNA	6 LFPPIKIGVALLGLMAFDW	15 IDHAAHLGGGICFWLYKYGYST
21281444 Arath	4 SSSGAFFGLIGAMLSALAKN	8 SALAIIFTFTVNLGIFLFP	0 -IDNFANIGGPIGFLGLVPLLFK
15225245 Arath	10 SVGASGAIIFGLVGSVAEIVIRH	8 EDLMQIAQIITALNMMGLMSR	1 IDNWHIIGGLGTAMTLLGQP
15231701 Arath	12 AVGYSCVVFQWMTILSVKQRSS	6 LLSLPISEFAPFESLIFTSIIV	2 ASFLGHLGSLVGYAISWGLIGG
18312405 Pyrae	20 AVGASGAIISGVLGAYMVLVPHA	15 IP-AWAYIGFWFIYQLFYGAL	9 VAYFAHIGGF IAGALTALYRNR
15789622 Halsp	25 TLGASGAIITAAALMGVLTLLNPLG	7 IPMLPLATGLFAAYSIFVSG	8 VAQLAHLGAGLIGLGYGAKLRE
20093492 Metka	5 VVGAASAIIFGLGCLTMLRMS	6 IPMLALFAVLYAALALFVIQ	6 VAHAGHLVGMIVGGVLLALRPS
12226784 Metma	4 FGGLSGVLYGLLGHWCIFQYLA	3 AYRLPRGVVAMMLIWLVLCSL	10 TANGAHVGGLLVGLSGSGLLGLL
14520881 Pyrab	5 SGGASGGLFVIGALLSIEGVL	3 IQKALINALALFLINSIF---	2 VNIFAHVGGVLTGLVGLFYGYIV
14601690 Aeerp	22 AVGASGAIISGVLGAYALILPFS	15 SVPASIFIGFWFYQVLMGLA	9 IAFWAHVGGTLGVALAPLVDK
15669882 Metja	8 SVGASGAIIFGIMGALAILAPHL	8 IPVNVIRVAVIIFALIDLILP	6 IAHTHLGAGLITGLIFGKLLYRK
15790000 Halsp	14 AVGFSGVVFAFAGFALKYPLA	4 VAARDAISVFWRTLEPVTFA	13 VAVQGHLPGLLGAALAVLVLH
15897391 Sulso	5 SSGASGGIFGLLSYTYFYDFLK	4 GYGLVFLVSVFVSDLIF---	2 VNVVAHIGGLIGGIMYAVVYLLI
15920355 Sulto	5 SAGASGGIFGLPAYITVTDYLK	4 NQISILLVSVFVILSDTLPF	2 VDIWAHTGGITGLISLILFFKI
16081803 Theac	8 SYGQSGVVYGLMGSAASMAILD	34 SLALIFLLTFVFMALDKAFY	5 IDSFWHAMAFGSSAIIFPIISYT
15598282 Pseae	10 MVGASGAIIVGFPAALTVLEPNL	6 VPMRLKHALLFAVDFLVMVN	4 IAHTHLGSLFVGLMYGRYIKRM
17549219 Ralso	9 IVGASGCVFGLLLAYVLVFRPR	9 PMPAHLFATYVALVELTLGIS	5 IAHFHHLGGMAGSGVLLNRW-LR
17549744 Ralso	5 VVGASGAVFVAGAVLVAIRQY	8 SKRLLTQIGLVLYSLVQGLT	3 VDNAHVGGILGGCLLACILPAR
17987022 Brume	6 LVGASGAIISGMMGAARYGFRR	21 LKPVILFVGVWFLINIVTGLY	9 IAWEAHIGGPIAGFFPIPLMDRP
19553712 Corgl	1 HIGASGLIYGLYGLIVRGIFN	3 KQFLGLVIFLAFYISGLFWGLL	5 VSWQGHLPFALGGIAGAGAFIAD
20806909 Thete	6 VVGASGAIAGIMGAYVFLPSPA	16 PIPAVVYLFWFLPTQLYSGMV	11 IAWWAHIGGPIISGVLLNRPFLRD
21220616 Strco	6 LVGASGAIISAVLGFAPLFLPRA	14 RFPAAVWLPFWVSLQWLAAGR	6 VAYLHVLVGLFGFAMVAVNRGR
21222264 Strco	5 TLGASGAIIFGLFGATAALVLR--	1 LNADMRPVLLVLSLIFTFT	3 ISWQAHVGGVAVGAVIYGYMLHA
21224370 Strco	6 LIGASGAIAGVGLGAYLVLYPRA	14 RLPAAVWLVGFVGLQAVYSSG	8 VAYVAHVGVFGVGMILIAVPLRR
21224996 Xanca	6 IIGASGAVSALIGTYLALFPGA	15 RVPAPPLIGAWAVLQVVFAYI	6 VAWSAHIAAGFVFGIYVGLYVRA
21230863 Xanca	12 SLGASGAVSALVFAVILKLPWT	7 PAPAIIYAVFYVGYSLWMDRR	4 INHSAHLGAAAFVGMFMLEPBR
21233650 Xanca	1 HLGASGVTHTGLMFLVFLGLLR	3 PAIATSMIAFLFYGGMLMTIL	5 VSWQSHLGGAVAGLGLAALLRLR
21675030 Chlte	6 TIGASGAIIFGVLAFGMMFPDR	8 PIKTKYFVAGYALIEFIMGLG	9 IAYFALHGGMLFYGIYIVRRNE
1168254 Prost	19 VGASGAIMGIAAASVYIYLKV	14 QKYQLYNLIAMTALTLINLQ	2 VDNAHIGGAIIGALISYAYLIV
13470470 Meslo	6 LVGASGAIISGMMGAARGFERT	21 SRGVVFLAVWMIINLATGLL	9 IAWEAHIGGPIAGFFPIPLMDRP
13473011 Meslo	6 LIGASGAIAGVYVAVLILYPRV	12 RIPAFIPLILWLVQVFMFAA	5 ISWACHIGGI IAGVALVVLRRS
15606530 Aquae	9 MVGASGAIISGVLGAYMFMFPHA	15 ELPAVIFIGLWFFIQLINGII	9 VAWYAHIGGPIITGYLLVDYFRKR
15607252 Myctu	5 TAGASGAVFGLFGATFMVVAR--	1 LHLDRVWVALIVINLAFTFL	3 ISWQGHVGGVLTGALVAATVYVA
15608477 Myctu	6 HIGASGLIFGLVAFLLVFLGLFV	3 WDIVIGLVFLVYGGIILGAM	8 VSWQGHLSGAVAGVAALYLSAP
15639966 TrepA	9 LLGASGSIIFALFLFVSMFPFTA	8 PIPAPLLIVGYLFIIDFLDF	4 VSHLTHLGLVFAWGYIRIRRGI
15640131 Vibch	3 FGGLSGVVYALAGYLWILGQRA	3 GLSIPRSLMGFMLIWLVLGYV	5 IANTHLAGLISGVLVAALWFSQR
15641983 Vibch	1 YVGLSGLTHLGLFAYYALNEALN	5 WLLVLGVIGKVAWEQWFGASA	9 VATEAHLAGLVGGLLAAGHCFL
15643350 Thema	4 SVGASGAIIFGLIGLFAAGFRK	3 FFMKPVTVGSLPILINLVVY	7 INNAHLGGFLSGMLLGYTMSFP
15643845 Thema	6 MVGASGAVSVMGAYVFLPFSY	15 EIPAFYILMIFWFIQVNLGLV	4 IAWWAHIGGPIYGMIVGYILRMR
15672152 Lacla	4 SAGASTSIFGLFAAVVGLFAFFT	4 LQOIGRMFTVLIVANVMNLF	4 VSIWAHIGGAIIGLLLSAIPAPK
15803931 Escco	3 FGGLSGVVYALMGYVNLRRGERD	3 GIYLQRGLIIFALIWVAGWF	6 MANGAHIAAGLAVGLAMAFVDSL
15806990 Deira	1 HLGASGLIFGLVAFLLVFLGLFV	3 LSVVIAVIFALYGGVWLVGL	5 ISWEHLGFPPIGGLVAALLLSK
15827590 Mycle	6 HIGVSGLIIFGLVAFLLVFLGLFV	3 W-DIIGCMFLFAYGGVLLGVM	8 VSWQGHLCGAIISGVVAALYLSAP
15837251 Xylfa	8 VLGASGGVFGLLMAYGMLFPNE	9 PMKARTFVILYGVIELLMGIT	5 VAHFTHLGGMFLGWLLIRYWRGQ
15837656 Xylfa	6 IIGASGAVSALIGSYLALFPGA	15 RVPAPPLIGFALLQVVFAYT	6 VAWSAHIAAGFVFGIYVGLYVRA
15838777 Xylfa	12 SLGASGAVSALVFAAVLLQWPA	7 PAPAIIYAVFYVGYSLWMDRR	4 INHSAHLGAAAFVGMFMLEPBR
15889057 Agrtu	6 LIGASGAVSALMGAAACRFAPV	22 NRTVLIFTLMLWFGNVLTAIG	10 IAWDAHVFGFLGLGFLFLDRP
15891346 Agrtu	6 LIGASGAVSALMGAAACRFAPV	12 PLPAIFLPLALWIGQQLMLAL	5 VSWGHLGGVILGALIMVIFMRFP
15894241 Cloac	4 SVGASGAIIFGLGAAIVFGEKFL	4 GKAFANMVGVPALNIFISFT	3 IDIFAHFGGLGGLVGVVSLGRT
15903945 Strpn	4 AAGASTSLYGLFAAIVLRYAT	4 IQQLQSYLTLFVNNIGSVL	3 ISLAGHIGGAVGGAFLAVIFPFR
15966395 Sinme	6 LIGASGAIISGVVAAYVLLHFKV	12 PLPAAIFLPAFWIGQQLMFLA	5 VSWAHVGGVILAGLVLLVLRPR
16077528 Bacsu	5 HVGASGAIIFGLFGVYLFMVLFR	4 GOEHSKMIITLAFVAVLMSFI	3 INMMAHLFGLCGGLLSPFCVQK
16079543 Bacsu	3 SAGASGAIIFGLGALLVYVSLN	4 LRTIGTNIIVIIINLGFQFA	3 IDNSHIGGLIGGFAVVAALGLP
16126863 Caucr	11 VVGASGAIAGLGMGAARTMDSA	8 GPRVLSLGLWLVNMLAVT	10 VAWEAHLIGFVAVGLIIGPFAR
16272560 Haein	3 FGGLSGVVYAVLGYVIRDKLN	2 LFDLPEGFMTLLVGIAGLFI	7 MGNAAHISGLVILWGFIDSKL
16332120 Synsp	1 TVGASILIFGVLGFLFRGNFQ	3 ASIVLSIVLVLYGSALWGLL	5 VSWQHLGFPPIGGAIAAWLIARE
16800442 Lisin	3 SVGASTAVFVMMGALLYLVVLK	4 AKTIGTSIASLVAINLLIDVF	3 IDIAGHIGGLVGGFLLAGALSPL
17231423 NossP	6 SLGASGAIISGVLGAYLIRFQA	15 SVPALVLIIGFFVQNVISGLV	14 VAYWAHIGGPIIFLAPIFGLF
17232329 NossP	6 SLGASGAIISGVLGAYLIRFQA	15 RVPAYFPLGFVFLQSQPYGLA	14 IAYWAHAGGPIFGALLGLPLGLF
consensus/90%	.hGhSsh.uhhhh.....	.....hh.hh.h.h....	hs..sHh.Ghh.Ghh.....

Figure 1 (continued from the previous page)

*Drosophila* RHO7 [6], resides within a large, heterogeneous prokaryotic cluster (Figure 2). Within this subfamily, PARL and its orthologs from other animals and from fungi have distinct domain architecture, with an extra TMH added to the amino terminus of the core, whereas the rest have only the core (a carboxy-terminal TMH and a ubiquitin-associated domain are appended in one *Arabidopsis* protein; Figure 2). Thus, the existence of two distinct subfamilies of eukaryotic rhomboids is supported by features of domain architectures that appear to comprise shared derived characters. Within these two major eukaryotic subfamilies, evolution apparently proceeded by both ancient and more recent duplications. Several lineage-specific expansions of paralogs [23] are noticeable, in insects, mammals and plants (Figure 2).

Archaeal rhomboids are scattered over the phylogenetic tree, with two major clusters and, in addition, three isolated proteins joining different bacterial branches (Figure 2). There is no indication of an affinity between any of the archaeal and eukaryotic rhomboids. Although many of the bacterial rhomboids form phylogenetically coherent clusters corresponding to the established bacterial lineages, there are also several clusters that have an odd composition, such as the grouping of proteobacterial and Gram-positive species; some of these clusters are well supported by bootstrap (see clusters 1-4 in Figure 2).

Unexpected tree topologies often emerge due to artifacts of phylogenetic analysis methods. This concern is particularly serious for highly divergent families of membrane proteins, such as the rhomboids, in which parallel amino-acid substitutions are likely. Therefore we investigated the phylogeny of the rhomboid family in greater detail using several independent phylogenetic methods and the corresponding statistical tests. First, we assessed the robustness of the topology of the tree shown in Figure 2 using the Kishino-Hasegawa (KH) test whereby the clade of interest is forced into various positions on the tree and the likelihoods of the resulting topologies are estimated. Specifically, the KH test was used to evaluate two alternative topologies, in which the RHO and PARL subfamilies formed a clade, and two topologies, in which the RHO subfamily formed a clade with archaeal rhomboids (Figure 2 and Table 1). Each of these alternative topologies had a significantly lower likelihood than the original topology shown in Figure 2 (see Table 1).

**Table 1**

**Log-likelihood analysis of possible placements of selected branches of maximum likelihood trees for the proteins analyzed**

Tree*	Diff lnL <sup>†</sup>	SE <sup>‡</sup>	RELL-BP <sup>§</sup>
Original tree	0.0	-	0.9702
A → B	-18.9	10.2	0.0264
B → A	-46.6	14.6	0.0003
A → C	-30.3	12.8	0.0031
A → D	-47.9	15.6	0.0000

\*A-D, clades that were subjected to local rearrangements in the tree as indicated in Figure 2 and discussed in the text. <sup>†</sup>Difference of the log-likelihoods relative to the best tree. <sup>‡</sup>Standard error of Diff lnL. <sup>§</sup>Bootstrap probability of the given tree calculated using the RELL method (resampling of estimated log-likelihoods).

In addition, a tree of the rhomboid family was constructed using the Bayesian inference method, which has recently become a practical alternative to the more traditional methods of phylogenetic analysis [24,25]. The tree produced using the MRBAYES package [26] showed the same major clades as the tree in Figure 2 (data not shown); moreover, clustering of the RHO and PARL subfamilies of eukaryotic rhomboids with the respective prokaryotic clades was supported by high posterior probabilities (Figure 2).

We also attempted to construct a phylogenetic tree of the rhomboid family by using the maximum parsimony method [27]. The resulting tree contained the same major clades as the trees constructed using ML and MRBAYES; however, the number of parsimony-informative sites was insufficient to obtain high bootstrap support with this approach (data not shown).

We also tested alternative phylogenies using neighbor-joining search with constraint trees [27]. The alternative phylogenies reflected two distinct hypotheses: first, clustering of the RHO and PARL subfamilies of eukaryotic rhomboids with the prokaryotic rhomboid families as suggested by the tree topology in Figure 2; and second, monophyly of the eukaryotic rhomboids (Figure 3). The phylogenies corresponding to these alternative hypotheses were compared to the best phylogeny using three statistical tests (Table 2). The

**Figure 2** (see figure on the next page)

Phylogenetic tree of the rhomboid family. The sequences and their regions used to construct the tree are exactly those shown in Figure 1. The color coding and abbreviations are as in Figure 1. The two major eukaryotic subfamilies are denoted as RHO and PARL (see text) and four clusters containing unexpected (from a phylogenetic viewpoint) sets of species are denoted 1-4. The clades that were investigated in the KH test are denoted A through D. Although the tree is shown in a pseudorooted form for convenience, this is an unrooted tree. Internal nodes with at least 70% RELL bootstrap supported are denoted by black circles and nodes with a 50-70% support by blue circles. The posterior probabilities reported by the MRBAYES program are indicated for some key internal branches. Domain architectures are connected to the respective proteins by brackets or lines. The domain key is shown at the bottom of the figure.

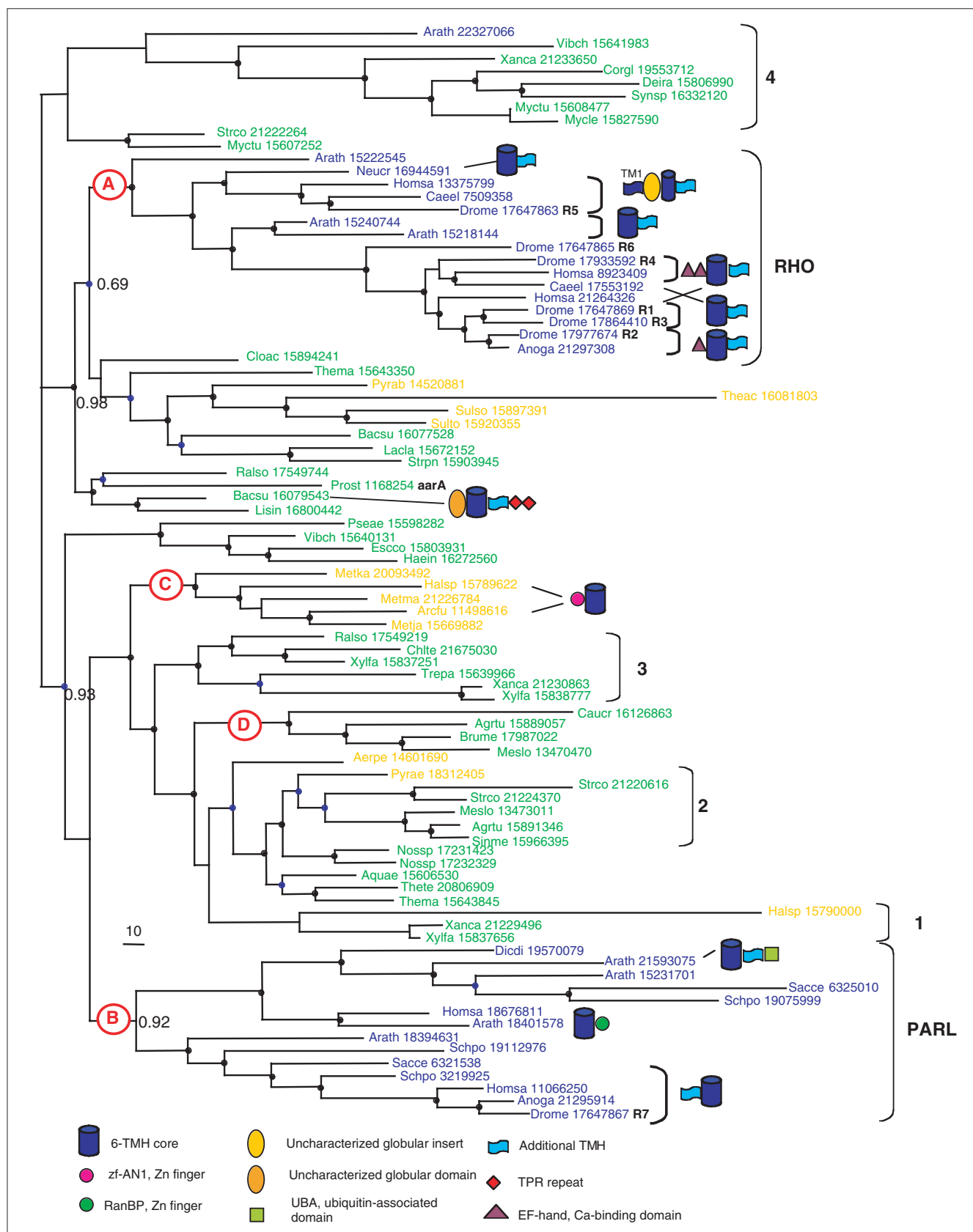
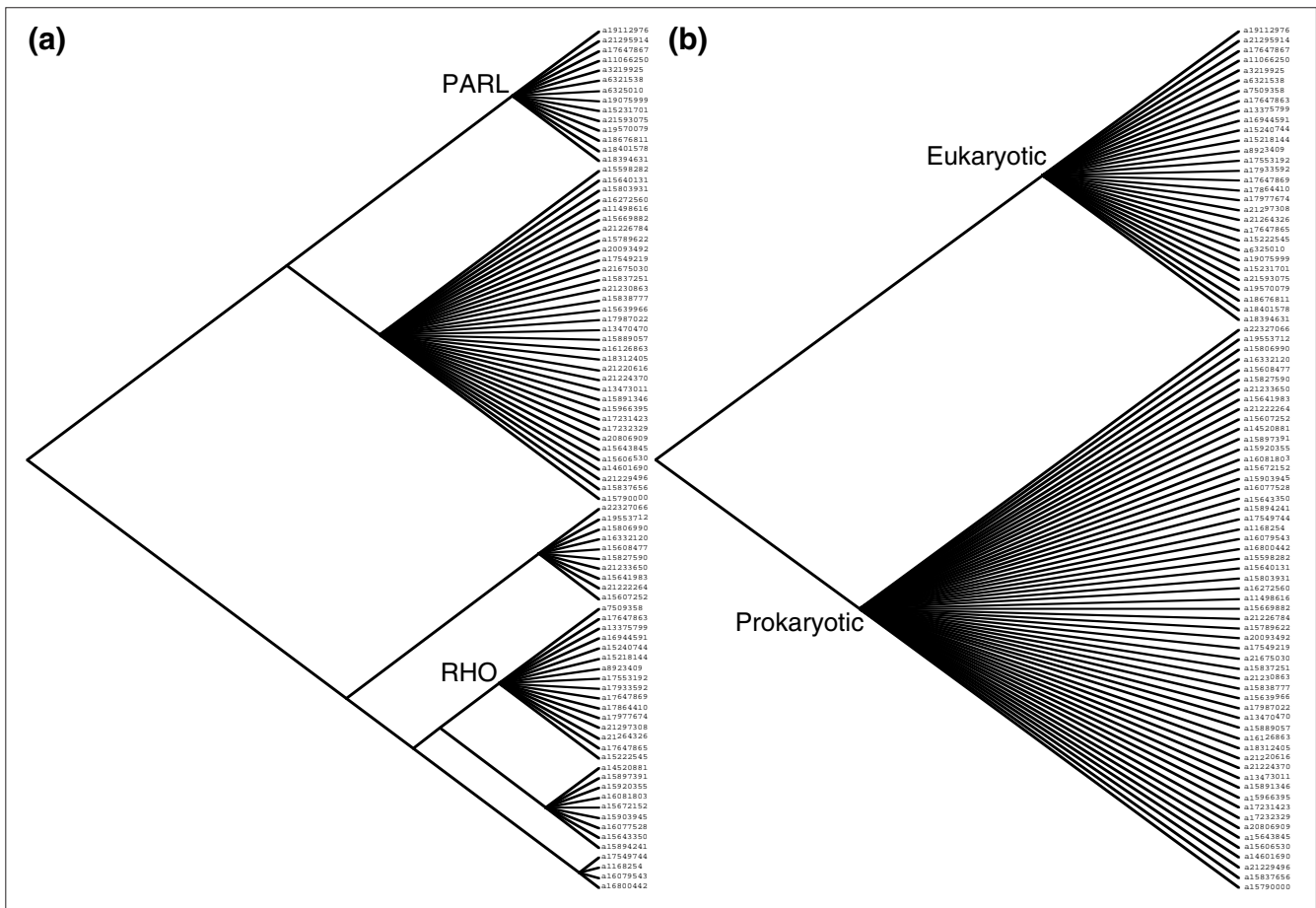


Figure 2 (see legend on the previous page)



**Figure 3**  
 Hypothesis-specific constraint tree for the rhomboid family. **(a)** Hypothesis 1, polyphyletic origin of eukaryotic rhomboids from prokaryotic progenitors. The RHO and PARL subfamilies are denoted; the remaining clusters include prokaryotic rhomboids designated as in Figure 2 (with 'a' added to the GI number). Within each cluster, the branches were collapsed into a multifurcation. **(b)** Hypothesis 2, monophyletic origin of eukaryotic rhomboids. All eukaryotic and prokaryotic sequences were collapsed into the two respective clusters. The trees are unrooted, although shown in a pseudorooted form.

hypothesis 1 tree was not significantly different from the best tree under any of these tests whereas the hypothesis 2 tree was significantly ( $p < 0.05$ ) worse than the best tree according to each of the tests (Table 2).

The concordance of the results obtained with several independent methods for phylogenetic tree construction and statistical analysis specifically aimed at testing the alternative hypothesis of monophyletic origin of eukaryotic rhomboids shows strong support for the major aspects of the tree topology in Figure 2 and, in particular, for the polyphyly of eukaryotic rhomboids.

The phylogenetic tree of the rhomboid family shown in Figure 2 and supported by the additional tests described above follows neither the 'standard model' scenario [28,29], with the major split between the archaeo-eukaryotic and bacterial lineages nor the 'mitochondrial' scenario, which postulates acquisition of a gene by eukaryotes from the

pro-mitochondrial endosymbiont. Neither can this tree be explained by postulating a small number of lineage-specific gene losses. The parsimonious interpretation of the rhomboid family tree seems to be that the evolutionary history of this family had been replete with horizontal gene transfer (HGT) and lineage-specific gene loss events. In particular, in spite of the presence of rhomboids in the majority of modern life forms from all three primary superkingdoms, phylogenetic analysis suggests that this family has not been inherited from the last universal common ancestor (LUCA). Instead, the tree topology seems to indicate that this family emerged in some bacterial lineage and afterwards had been widely disseminated by HGT, and then lost in some lineages. Both archaea and eukaryotes seem to have acquired rhomboids on several independent occasions. In particular, at least two HGT events seem to have contributed to the origin of eukaryotic rhomboids, one of them yielding the RHO subfamily and the other one the PARL subfamily, with a possible additional HGT in plants (Figures 2,3).



**Table 2****Statistical comparisons of the best neighbor-joining tree with the hypothesis 1 and hypothesis 2 trees**

## Kishino-Hasegawa test

Tree	Length	Length difference	SD (difference)	t	p*
Best	4951	-			
Hypothesis 1	4966	15	11.9	1.26	0.211
Hypothesis 2	4974	23	10.8	2.12	0.036

## Templeton (Wilcoxon signed-ranks) test

Tree	Length	Rank sums	N	z	p*
Best	4951	-			
Hypothesis 1	4966	1418.0 -997.0	69	-1.33	0.185
Hypothesis 2	4974	1244.5 -708.5	62	-1.97	0.048

## Winning-sites (sign) test

Tree	Length	Counts	p*
Best	4951		
Hypothesis 1	4966	36 -33	0.810
Hypothesis 2	4974	40 -22	0.031

\*Probability of getting a more extreme test statistic under the null hypothesis of no difference between the two trees (two-tailed test).

Given the broad phyletic representation of both subfamilies of eukaryotic rhomboids, both the RHO subfamily and the PARL subfamily must have been acquired through HGT at an early stage of eukaryotic evolution, definitely before the divergence of the major crown-group lineages. This early epoch in eukaryotic evolution is thought to have been dominated by HGT from multiple bacterial symbionts [30,31].

An alternative to this multiple-HGT scenario is that LUCA already had multiple, paralogous rhomboids, which evolved by a series of ancient gene duplications, and the odd topology of the phylogenetic tree is due primarily to differential loss of these ancient paralogs. Although this cannot be ruled out formally, this hypothesis implies the existence of an elaborate signaling system in LUCA and, accordingly, suggests that LUCA was a complex organism, which might have had as many genes as modern bacteria. Theoretical analysis of evolutionary scenarios constructed on the basis of the phyletic patterns of COGs by applying the parsimony principle shows that the complexity of the inferred gene set of LUCA critically depends on the relative rates of gene loss and HGT at the early stages of evolution [32]. A complex

LUCA with around 2,000 genes is predicted only when one assumes that the rate of gene loss is an order of magnitude greater than the rate of HGT. However, explicit reconstruction of the gene set of LUCA under the assumption of equal rates of gene loss and HGT leads to a hypothetical genome that consists of only around 600 genes but appears to be 'compatible with life', that is, it includes genes responsible for most, if not all, essential cellular functions [32]. We currently believe that this is the most realistic, albeit inevitably imprecise, reconstruction of LUCA's gene set. With respect to the rhomboid family and other families whose phylogenetic trees show similar patterns, this makes the multiple-HGT interpretation the scenario of choice. Further theoretical, comparative-genomic and experimental analyses aimed at determining relative rates of gene loss and HGT will help in a more objective assessment of the validity of this argument.

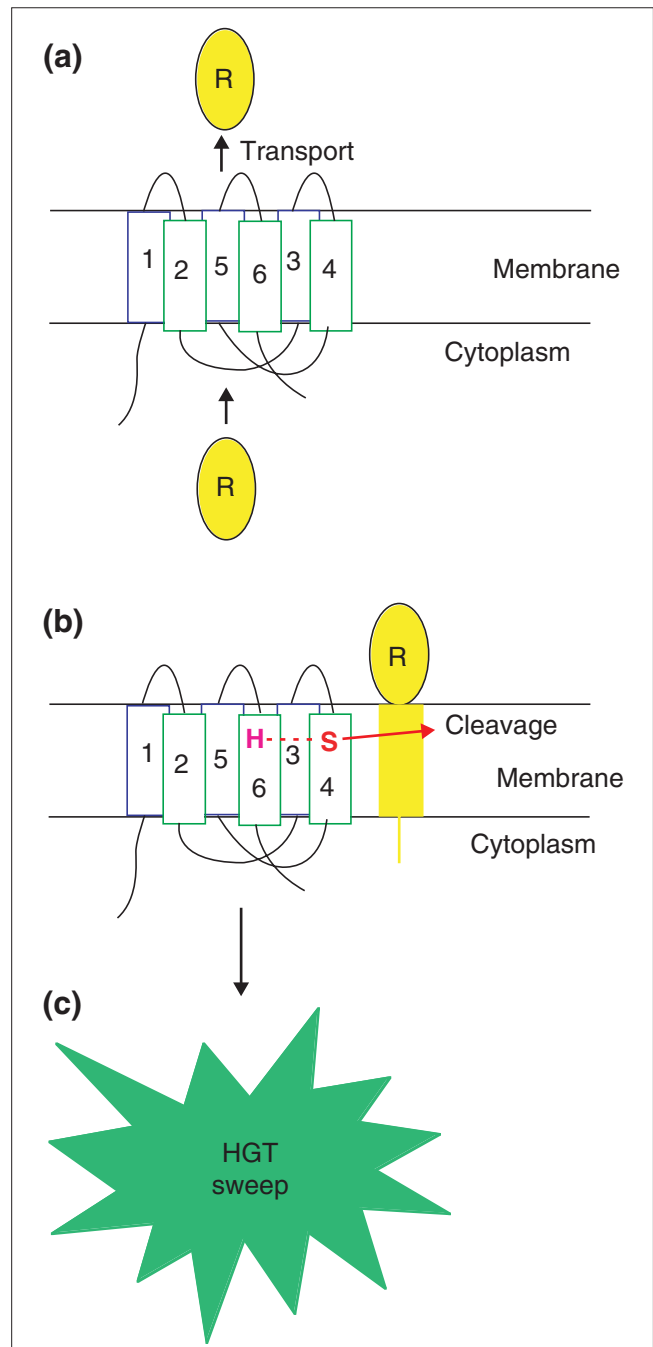
The multiple-HGT interpretation of the evolutionary history of the rhomboid family, while supported by the above argument, seems, at least at first glance, distinctly counter-intuitive, given that this family is nearly ubiquitous among extant life

forms. Indeed, when attempts are made to construct parsimonious evolutionary scenarios on the basis of phyletic patterns alone [31-33], there is no chance that such a widespread family is not assigned to LUCA. It should be realized, however, that these approaches are inherently probabilistic, and extensive HGT can fool them [34]. For the rhomboid family, the multiple-HGT mode of evolution seems to be particularly plausible. It seems likely that the ultimate ancestor of the rhomboid family evolved from a nonenzymatic integral membrane protein, probably a transporter that might have been involved in an early primitive form of export of signaling peptides in bacteria. The protease active center might have evolved in such a transporter by chance emergence of the suitable catalytic amino acids within two or three of the TMHs (Figure 4). This would enable the transition from simple transport to the RIP mode of controlled export of signaling molecules. Emergence of RIP could have conferred a major selective advantage on the respective bacteria and might have resulted in an evolutionary sweep whereby the gene carrying this trait was repeatedly fixed, rather than eliminated, after HGT. In terms of the evolution of sequence itself, the requirements for the conservation of the protease activity apparently 'locked' the rhomboid family in a regime of relatively slow evolution, which ensures significant sequence similarity between all family members (Figure 1). The scenario of origin from non-catalytic transporters might potentially apply to other integral membrane enzymes, including intramembrane proteases involved in RIP, such as presenilins and their homologs [14,15] and the archaeo-eukaryotic signal peptide peptidase [35].

### Conclusions

The rhomboid family might be the most widespread and conserved group of integral membrane proteins. In and by itself, this would suggest that this family is part of the gene repertoire of LUCA. However, phylogenetic analysis suggests a different scenario, one of emergence in a bacterial lineage with subsequent multiple, independent HGT events and gene losses. Although caution is due in the evolutionary interpretation of phylogenetic trees for large families, particularly when membrane proteins with a relatively small number of conserved positions, such as the rhomboids, are involved, the multiple-HGT scenario seemed to be supported by several methods of tree analysis and statistical tests.

Eukaryotes probably acquired their two major rhomboid subfamilies, RHO and PARL, as the result of two independent, early HGT events. These events, which might have introduced RIP as a means of intercellular communication, could have been pivotal in the evolution of eukaryotic multicellularity along the lines discussed previously with regard to the apparent bacterial origin of key components of eukaryotic programmed cell death machinery [36]. Subsequent evolution of rhomboids in eukaryotes proceeded by lineage-specific expansion of paralogs [23] followed by



**Figure 4**

A hypothetical scenario for the origin and dissemination of the rhomboid family proteases. The figure schematically shows the proposed three stages of evolution of the rhomboid family. In (a), the progenitor of the rhomboid family functions as a transporter for a regulatory peptide in some bacterial lineage. In (b), the catalytic site of the intramembrane protease evolves, allowing the switch to RIP as the mechanism of the regulatory peptide release. In (c), the emergence of RIP is followed by a burst of HGT. R, regulatory peptide. The transmembrane helices of rhomboid are designated as in Figure 1; their topology in the membrane is based on that proposed in [7]. The catalytic histidine and serine are shown and connected by a dotted line to indicate the proposed charge-relay system of the protease; possible ancillary catalytic residues are not shown.

diversification through the addition of an extra TMH in different positions relative to the catalytic core, some limited domain accretion (see Figure 2) and sequence divergence.

Phylogenetic analysis of the rhomboid family described here carries a general message for studies aimed at the reconstruction of ancestral life forms, particularly LUCA. Although most of the (nearly) ubiquitous protein families probably do derive from LUCA, explicit phylogenetic analysis is required to ascertain this in each case.

## Materials and methods

The nonredundant (NR) protein sequence database at the National Center for Biotechnology Information (NIH, Bethesda) was searched iteratively using the PSI-BLAST program with multiple starting queries [37]. PSI-BLAST was normally run with expectation (E) value of 0.01 as the cut-off for inclusion of sequences into the position-specific scoring matrix. Multiple alignments of protein sequences were constructed using the ClustalW program [38] and manually adjusted on the basis of the examination of PSI-BLAST search outputs and the superposition of the predicted TMHs, which were identified using the programs TMPred [39] and TMAP [40].

Phylogenetic trees were built using the least-squares method [41] implemented in the FITCH program of the PHYLIP package [42], with subsequent local rearrangement using the PROTML program of the MOLPHY package to obtain the maximum likelihood tree [43]. The reliability of the tree topology was assessed using the RELI (resampling of estimated log-likelihoods) bootstrap method of MOLPHY, with 10,000 replications [44]. Alternative placements of selected clades in maximum-likelihood trees were compared by using the rearrangement optimization method (Kishino-Hasegawa test) as implemented in the ProtML program [43-45]. Maximum parsimony trees were constructed using the heuristic search option of PAUP\* [27]. In addition, trees were constructed by Bayesian inference using the Markov chain Monte Carlo method as implemented in the MRBAYES package [24,26]. The complete alignment information, including columns with gaps, was used for the MRBAYES analysis.

Constraint trees for phylogenetic hypothesis testing were generated using the TreeView program [46]. Constraint trees were imported into PAUP\* [27] and subjected to neighbor-joining search to generate the phylogenies corresponding to alternative hypotheses. These phylogenies were compared using the KH [45], Templeton (Wilcoxon signed-ranks) [47] and Winning-sites (sign) [48] tests implemented in PAUP\*.

## Acknowledgements

L.P. is supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

## References

1. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29**:22-28.
2. Sturtevant MA, Roark M, Bier E: **The *Drosophila* rhomboid gene mediates the localized formation of wing veins and interacts genetically with components of the EGF-R signaling pathway.** *Genes Dev* 1993, **7**:961-973.
3. Sturtevant MA, Roark M, O'Neill JW, Biehs B, Colley N, Bier E: **The *Drosophila* rhomboid protein is concentrated in patches at the apical cell surface.** *Dev Biol* 1996, **174**:298-309.
4. Guichard A, Biehs B, Sturtevant MA, Wickline L, Chacko J, Howard K, Bier E: **rhomboid and Star interact synergistically to promote EGFR/MAPK signaling during *Drosophila* wing vein development.** *Development* 1999, **126**:2663-2676.
5. Mushegian AR, Koonin EV: **Sequence analysis of eukaryotic developmental proteins: ancient and novel domains.** *Genetics* 1996, **144**:817-828.
6. Pellegrini L, Passer BJ, Canelles M, Lefterov I, Ganjei JK, Fowlkes BJ, Koonin EV, D'Adamio L: **PAMP and PARL, two novel putative metalloproteases interacting with the COOH-terminus of Presenilin-1 and -2.** *J Alzheimers Dis* 2001, **3**:181-190.
7. Urban S, Lee JR, Freeman M: ***Drosophila* rhomboid-1 defines a family of putative intramembrane serine proteases.** *Cell* 2001, **107**:173-182.
8. Klambt C: **EGF receptor signalling: roles of star and rhomboid revealed.** *Curr Biol* 2002, **12**:R21-R23.
9. Guichard A, Roark M, Ronshaugen M, Bier E: **brother of rhomboid, a rhomboid-related gene expressed during early *Drosophila* oogenesis, promotes EGF-R/MAPK signaling.** *Dev Biol* 2000, **226**:255-266.
10. Wasserman JD, Urban S, Freeman M: **A family of rhomboid-like genes: *Drosophila* rhomboid-1 and roughoid/rhomboid-3 cooperate to activate EGF receptor signaling.** *Genes Dev* 2000, **14**:1651-1663.
11. Brown MS, Ye J, Rawson RB, Goldstein JL: **Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans.** *Cell* 2000, **100**:391-398.
12. Urban S, Freeman M: **Intramembrane proteolysis controls diverse signalling pathways throughout evolution.** *Curr Opin Genet Dev* 2002, **12**:512-518.
13. Wolfe MS, Xia W, Ostaszewski BL, Diehl TS, Kimberly WT, Selkoe DJ: **Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and gamma-secretase activity.** *Nature* 1999, **398**:513-517.
14. Steiner H, Kostka M, Romig H, Basset G, Pesold B, Hardy J, Capell A, Meyn L, Grim ML, Baumeister R, et al.: **Glycine 384 is required for presenilin-1 function and is conserved in bacterial polytopic aspartyl proteases.** *Nat Cell Biol* 2000, **2**:848-851.
15. Sreekumar KR, Aravind L, Koonin EV: **Computational analysis of human disease-associated genes and their protein products.** *Curr Opin Genet Dev* 2001, **11**:247-257.
16. Rather PN, Orosz E: **Characterization of *aarA*, a pleiotropic negative regulator of the 2'-N-acetyltransferase in *Providencia stuartii*.** *J Bacteriol* 1994, **176**:5140-5144.
17. Rather PN, Ding X, Baca-DeLancey RR, Siddiqui S: ***Providencia stuartii* genes activated by cell-to-cell signaling and identification of a gene required for production or activity of an extracellular factor.** *J Bacteriol* 1999, **181**:7185-7191.
18. Gallio M, Kylsten P: ***Providencia* may help find a function for a novel, widespread protein family.** *Curr Biol* 2000, **10**:R693-R694.
19. Urban S, Schlieper D, Freeman M: **Conservation of intramembrane proteolytic activity and substrate specificity in prokaryotic and eukaryotic rhomboids.** *Curr Biol* 2002, **12**:1507-1512.
20. Gallio M, Sturgill G, Rather P, Kylsten P: **A conserved mechanism for extracellular signaling in eukaryotes and prokaryotes.** *Proc Natl Acad Sci USA* 2002, **99**:12208-12213.
21. Esser K, Tursun B, Ingenhoven M, Michaelis G, Pratz E: **A novel two-step mechanism for removal of a mitochondrial signal sequence involves the mAAA complex and the putative rhomboid protease Pcp1.** *J Mol Biol* 2002, **323**:835-843.
22. **COGS: phylogenetic classification of proteins encoded in complete genomes** [<http://www.ncbi.nlm.nih.gov/COG>]

23. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes.** *Genome Res* 2002, **12**:1048-1059.
24. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP: **Bayesian inference of phylogeny and its impact on evolutionary biology.** *Science* 2001, **294**:2310-2314.
25. Huelsenbeck JP, Larget B, Miller RE, Ronquist F: **Potential applications and pitfalls of bayesian inference of phylogeny.** *Syst Biol* 2002, **51**:673-688.
26. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**:754-755.
27. Swofford DL: *PAUP\*: Phylogenetic Analysis Using Parsimony (\*and other methods)*. Sunderland, MA: Sinauer; 1998.
28. Brown JR, Doolittle WF: **Archaea and the prokaryote-to-eukaryote transition.** *Microbiol Mol Biol Rev* 1997, **61**:456-502.
29. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proc Natl Acad Sci USA* 1990, **87**:4576-4579.
30. Doolittle WF: **You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes.** *Trends Genet* 1998, **14**:307-311.
31. Koonin EV, Galperin MY: *Sequence - Evolution - Function. Computational Approaches in Comparative Genomics*. Boston: Kluwer; 2002.
32. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: **Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes.** *BMC Evol Biol* 2003, **3**:2.
33. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Res* 2002, **12**:17-25.
34. Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer.** *Mol Biol Evol* 2002, **19**:2226-2238.
35. Weihofen A, Binns K, Lemberg MK, Ashman K, Martoglio B: **Identification of signal peptide peptidase, a presenilin-type aspartic protease.** *Science* 2002, **296**:2215-2218.
36. Koonin EV, Aravind L: **Origin and evolution of eukaryotic apoptosis: the bacterial connection.** *Cell Death Differ* 2002, **9**:394-404.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
38. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
39. Hofmann K, Stoffel W: **TMbase - A database of membrane spanning protein segments.** *Biol Chem Hoppe-Seyler* 1993, **374**:166.
40. Persson B, Argos P: **Prediction of membrane protein topology utilizing multiple sequence alignments.** *J Protein Chem* 1997, **16**:453-457.
41. Fitch WM, Margoliash E: **Construction of phylogenetic trees.** *Science* 1967, **155**:279-284.
42. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266**:418-427.
43. Adachi J, Hasegawa M: *MOLPHY: Programs for Molecular Phylogenetics*. Tokyo: Institute of Statistical Mathematics; 1992.
44. Kishino H, Miyata T, Hasegawa M: **Maximum likelihood inference of protein phylogeny and the origin of chloroplasts.** *J Mol Evol* 1990, **31**:151-160.
45. Kishino H, Hasegawa M: **Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea.** *J Mol Evol* 1989, **29**:170-179.
46. Page RD: **TreeView: an application to display phylogenetic trees on personal computers.** *Comput Appl Biosci* 1996, **12**:357-358.
47. Templeton AR: **Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the humans and apes.** *Evolution* 1983, **37**:221-244.
48. Prager EM, Wilson AC: **Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences.** *J Mol Evol* 1988, **27**:326-335.