Origin and early diversification of the papain family of cysteine peptidases

Dušan Kordiš and Vito Turk *

Posted Date: 22 June 2023

doi: 10.20944/preprints202306.1638.v1

Keywords: papain family; cysteine peptidases; phylogenomic analysis; evolution

Article

Origin and Early Diversification of the Papain Family of Cysteine Peptidases

Dušan Kordiš [1,*] and Vito Turk [2,3,*]

[1] Department of Molecular and Biomedical Sciences; J. Stefan Institute, Ljubljana, Slovenia
[2] Department of Biochemistry, Molecular and Structural Biology; J. Stefan Institute, Ljubljana, Slovenia
[3] Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia
* Correspondence: DK: dusan.kordis@ijs.si; VT: vito.turk@ijs.si.

Abstract: Peptidases of the papain family play a key role in protein degradation, regulated proteolysis, and the host-pathogen arms race. Although the papain family has been the subject of many studies, knowledge about its diversity, origin, and evolution in Eukaryota, Bacteria, and Archaea is limited; thus, we aimed to answer these long-standing questions. We traced the origin and expansion of the papain family through phylogenomic analysis, using sequence data from numerous prokaryotic and eukaryotic proteomes, transcriptomes, and genomes. We identified the full complement of the papain family in all prokaryotic and eukaryotic lineages. Analysis of the papain family provided strong evidence for its early diversification in the ancestor of eukaryotes. We found that the papain family has undergone complex and dynamic evolution through numerous gene duplications, which produced eight eukaryotic ancestral paralogous C1A lineages during eukaryogenesis. Different evolutionary forces operated on C1A peptidases, including gene duplication, horizontal gene transfer, and gene loss. This study challenges the current understanding of the origin and evolution of the papain family and provides valuable insights into their early diversification. The findings of this comprehensive study provide guidelines for future structural and functional studies of the papain family.

Keywords: papain family; cysteine peptidases; phylogenomic analysis; evolution

1. Introduction

The papain family (peptidase C1A family in InterPro (IPR013128), subfamily C1A peptidases in the MEROPS database (Db)) is the largest and best characterised group of cysteine peptidases, named after the first archetype, the plant cysteine protease papain. Members of this family are widely distributed in Archaea, Bacteria, Eukaryota and some viruses [1,2]. Papain-like peptidases are involved in numerous physiological and pathological processes, parasitic infections, and host defence. In parasitic protozoa, C1A peptidases participate in diverse processes, such as host cell and tissue invasion, encystation/excystation, catabolism of host proteins, and both stimulation and evasion of host immune responses [3]. In plants, C1A peptidases are involved in stress response, mobilisation of storage proteins during seed germination, induction of defence reactions, senescence, and regulation of cell death [4–6]. They are central hubs in plant immunity and are required for their resistance to various pathogens. At the same time, C1A peptidases are targeted by secreted pathogen effectors to suppress immune responses. Consequently, they are subject to a coevolutionary host-pathogen arms race [5]. The most studied mammalian C1A peptidases are human lysosomal cysteine cathepsins, which are essential for antigen processing [7], ageing, neurodegeneration [8], cancer [9,10], cardiovascular diseases [11], signalling [12,13], cell death [14], and inherited diseases [15]. Their activity can be regulated by gene expression, post-translational modifications, activation of inactive zymogens, accessibility to cleave peptide bonds, compartmentalisation, metal binding, and endogenous or exogenous inhibitors [16]. Dysregulation of C1A peptidase expression, localisation, or proteolytic activity can disrupt cellular homeostasis.

The first crystal structure of the papain family to be determined was papain from papaya (*Carica papaya*) [17]. The crystal structures of diverse human lysosomal cysteine cathepsins were determined during the 1990s, first with cathepsin B [18], followed by other cathepsins K [19], L [20,21], H [22], X [23], V [24], C [25], S [26], and F [27]. To date, more than 40 crystal structures of diverse papain family representatives have been determined [1]. The papain fold is composed of two domains: the left L-domain, which contains three α-helices, and the right R-domain, which contains a twisted β-sheet and two helices. The two domains are linked to each other, forming a deep active site cleft that acts as a substrate-binding groove in which Cys25 is positioned at the N-terminus (left domain) and His159 is positioned in the R-domain. Both residues form an ion pair. The binding sites between the substrate and the enzyme are the S2, S1, and S1' sites [28]. All cathepsins are monomers of approximately 30 kDa, with the exception of tetrameric cathepsin C [25,29] and dimeric cathepsin X [30]. Cathepsins differ in their specificity and tissue distribution [31]. Most cathepsins exhibit endopeptidase activity, whereas cath B, H, C, and X are the only known exopeptidases. However, cathepsins C and X are strict exopeptidases. Lysosomal cathepsins are synthesised as inactive zymogens. They are composed of propeptides that unfold at an acidic pH, thereby opening the active site of the enzyme [32]. Cathepsins are activated by autocatalytic processing [33–35] or by other proteases such as cathepsin C [36]. Equally important are cystatins, the cathepsin's endogenous proteinase inhibitors, which are the most investigated. Cystatins were divided into three families: stefins, cystatins, and kininogens [37,38]. Kininogens are composed of two inhibitory cystatin-like domains. They are divided into low molecular weight (LK) and high molecular weight (HK) kininogens. Both LK and HK kininogens bind to two molecules of cathepsins with high affinity, which is unique among cathepsins [39,40]. More information on cystatins can be found in [41,42].

The evolutionary analyses of the papain family started in the early 1990s in the pre-genomic era [43–45] and were based on a small sample of organisms and the limited diversity within the papain family that was available at that time. Since then, the number of representatives of the papain family has increased significantly, largely due to the accumulation of eukaryotic and prokaryotic transcriptomic and genomic sequences. Here, we aimed to obtain a comprehensive insight into the distribution, origin, and early diversification of the papain family in Eukaryota, Bacteria, and Archaea. Such an analysis could not have been previously performed because of the limited number of available eukaryotic and prokaryotic genomes. We used significantly expanded taxon sampling compared to previous studies, in which, for the first time, all major eukaryotic and prokaryotic lineages were represented [46,47]. In particular, we included data from previously unsampled eukaryotic lineages to represent all eukaryotic supergroups [46]. We traced the birth and expansion of the papain family through phylogenomic analysis, using publicly available information from numerous prokaryotic and eukaryotic proteomes, transcriptomes, and genomes. We found that the papain family expanded greatly during eukaryogenesis through massive gene innovation and diversification, which resulted in eight ancestral C1A lineages in the ancestor of eukaryotes. The papain family expanded further during eukaryotic evolution, especially through extensive gene duplications of the ancestral cathepsin L and B lineages. Together, we demonstrated that diversification of the papain family predates the origin of eukaryotes, and that a burst of innovation during eukaryogenesis led to a eukaryotic ancestor with a complex set of ancestral C1A lineages.

2. Results and Discussion

2.1. Papain family is highly represented in omics databases

Notably, the number of C1A peptidases in the proteomic, transcriptomic, and genomic databases is significantly different. Proteomic databases have the lowest numbers because many organisms lack proteomic data, but unannotated data exist for these organisms in transcriptomic or genomic databases. Although the papain family is well represented in the MEROPS database (release 12.4) with >17.000 sequences [1], much higher numbers of members of the papain family are present in general proteomic databases, such as InterPro and Superfamily. More than 60.000 C1A peptidases are present in InterPro database (IPR000668 - Peptidase C1A, papain C-terminal; IPR038765 - Papain-

 doi:10.20944/preprints202306.1638.v1

like cysteine peptidase family); however, vast numbers are present in transcriptomic and genomic databases. In this study, we obtained unbiased insights into the complete repertoire of the papain family across all major eukaryotic and prokaryotic lineages. A phylogenomic approach was used to analyse the papain family, in which we placed the proteome/transcriptome/genome data into an evolutionary context. To find new members of the papain family, we searched all publicly available proteomic, transcriptomic, and genomic databases of key eukaryotic taxa and/or lineages. We limited the search in the National Center for Biotechnology Information (NCBI) TSA transcriptomic and WGS genomic databases to specific taxonomic groups. Searching for C1A peptidases in the EukProt V3 transcriptomic database [48] was crucial because we obtained data for eukaryotic organisms and lineages that were not available at the NCBI Db. This approach was especially important for identifying new representatives of C1A orthologous gene families in a large number of diverse unicellular eukaryotes. Prokaryotes have only a small number of representatives of the papain family per species.

2.2. Early diversification of the papain family in the ancestor of eukaryotes

By using diverse eukaryotic representatives of the papain family (e.g., human, invertebrate, plant, and protist C1A sequences) as queries for searching eukaryotic proteome, transcriptome, and genome databases, we recognised the conserved repertoire of the eight ancestral C1A lineages that are present in all eukaryotic supergroups (Tables 1 and 2). This distribution pattern demonstrated that eight ancestral eukaryotic paralogous C1A lineages were present in the last eukaryotic common ancestor (LECA). The eight ancestral eukaryotic paralogous C1A lineages are as follows: cathepsins B, C, X, L, H, F, 26/29 kDa peptidase, and type 1 long C1 peptidase. Maximum likelihood (ML) phylogenetic analysis of the papain family in key eukaryotic lineages has provided strong evidence for the early diversification of the papain family (Figures 1 and 2).

Table 1 Distribution of the eight ancestral eukaryotic C1A paralogous lineages in Eukaryota

	cath B	cath C	cath X	cath L	cath F	cath H	26/29 kDa peptidase	type 1 C1 long
Diaphoretickes	■	■	■	■	■	■	■	■
Chloroplastida	■	■	■	■	■	■	■	□
Glaucophyta	■	■	■	■	■	■	■	□
Rhodophyta	■	■	■	■	■	■	■	□
Picozoa	■	■	■	■	□	□	■	□
Cryptista	■	■	■	■	■	■	■	□
Haptophyta	■	■	■	■	■	■	■	■
Centroheliozoa	■	■	■	■	■	■	■	■
Provora	■	■	■	■	■	■	□	□
Hemimastigophora	■	■	■	■	■	□	■	■
Telonemia	■	■	■	■	■	□	□	□
Stramenopiles	■	■	■	■	■	■	■	■
Alveolata	■	■	■	■	■	■	■	■
Rhizaria	■	■	■	■	■	■	■	■
Amorphea	■	■	■	■	■	■	■	■
Amoebozoa	■	■	■	■	■	■	■	■
Obazoa	■	■	■	■	■	■	■	■
CRuMs	■	■	■	■	■	■	□	□
Ancyromonadida	■	■	■	■	■	□	■	□
Malawimonadida	■	■	■	■	■	□	■	□
"Excavata"	■	■	■	■	■	■	■	■
Discoba	■	■	■	■	■	■	■	■
Metamonada	■	■	■	■	■	□	■	■

Black square represents presence, while white square represents absence

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 June 2023 doi:10.20944/preprints202306.1638.v1

Table 2 Distribution of the eight ancestral eukaryotic C1A paralogous lineages and two evolutionary younger C1A lineages in Obazoa

	cath B	cath C	cath X	cath L	cath F	cath H	cath O	26/29 kDa peptidase	type 1 C1 long	VWFA-C1
Opisthokonta	■	■	■	■	■	■	■	■	■	■
Metazoa	■	■	■	■	■	■	■	■	□	■
Choanoflagellata	■	■	■	■	■	□	■	■	□	■
Filasterea	■	■	■	■	■	□	■	□	□	□
Tunicaraptor	■		■	■	■	□	■	□	■	□
Pluriformea	■	■	■	■	□	■	■	■	□	□
Ichthyosporea	■	■	■	■	■	□	■	■	□	□
Rotosphaerida	□	□	□	□	□	□	□	□	□	□
Fungi	□	□	□	□	□	□	□	□	□	■
Breviatea	■	■	■	■	■	□	□	■	■	□
Apusozoa	■	□	□	■	■	■	■	■	□	□
Metazoa										
Basal metazoans	■	■	■	■	■	■	■	■	□	■
Bilateria	■	■	■	■	■	■	■	■	□	■
Protostomia	■	■	■	■	■	■	■	■	□	■
Deuterostomia	■	■	■	■	■	■	■	■	□	■

Black square represents presence, while white square represents absence

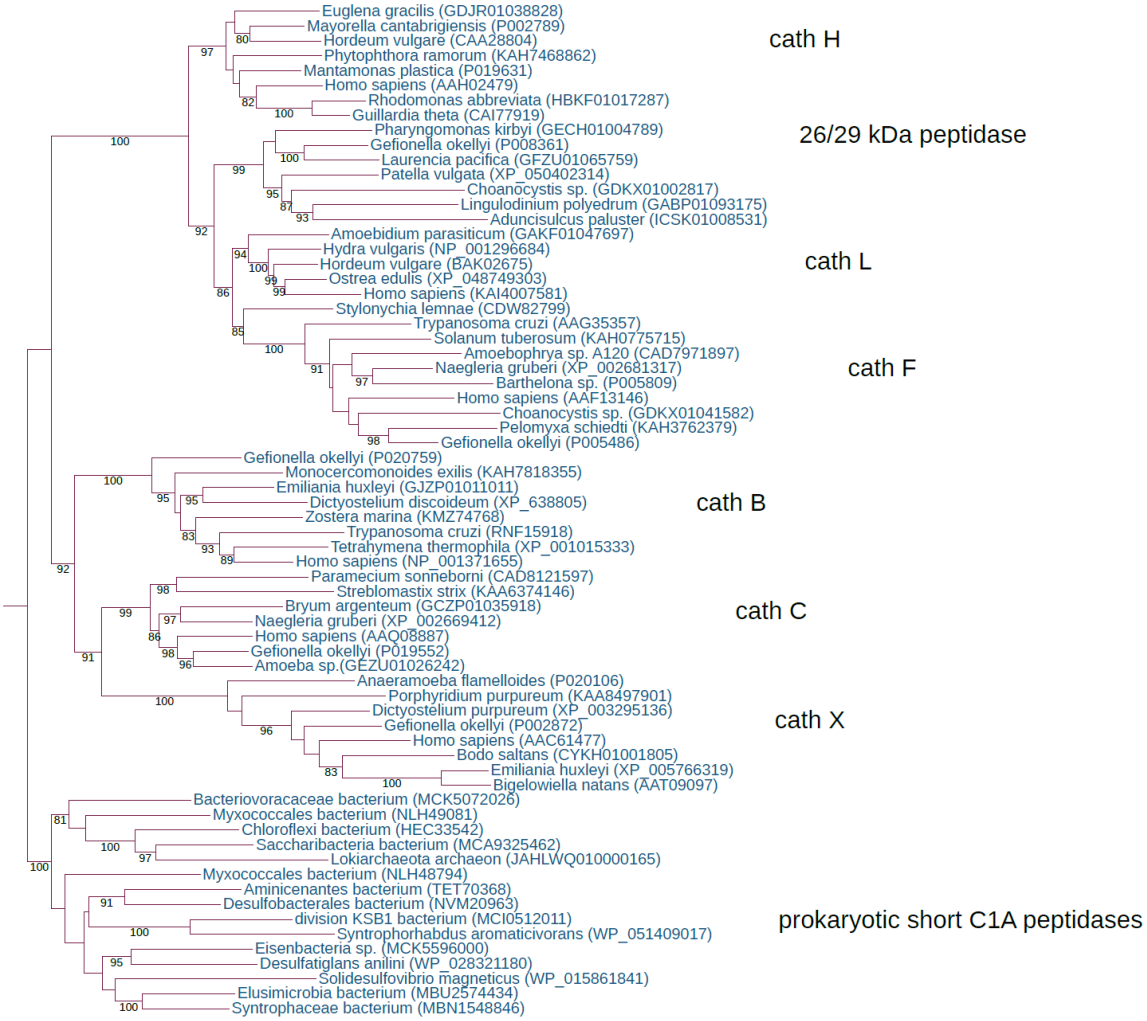

Figure 1. Early diversification of the papain family in the eukaryotic ancestor. The rooted ML tree shows the evolutionary relationships between the seven ancestral eukaryotic orthologous gene families, cathepsins B, C, X, L, H, F, and the 26/29 kDa peptidase. The best-fit model according to the Bayesian information criterion was WAG+I+G4. The ML tree represents bootstrap consensus following 1000 replicates. Sequences were obtained from GenBank, and species names and accession numbers are included.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 June 2023 doi:10.20944/preprints202306.1638.v1

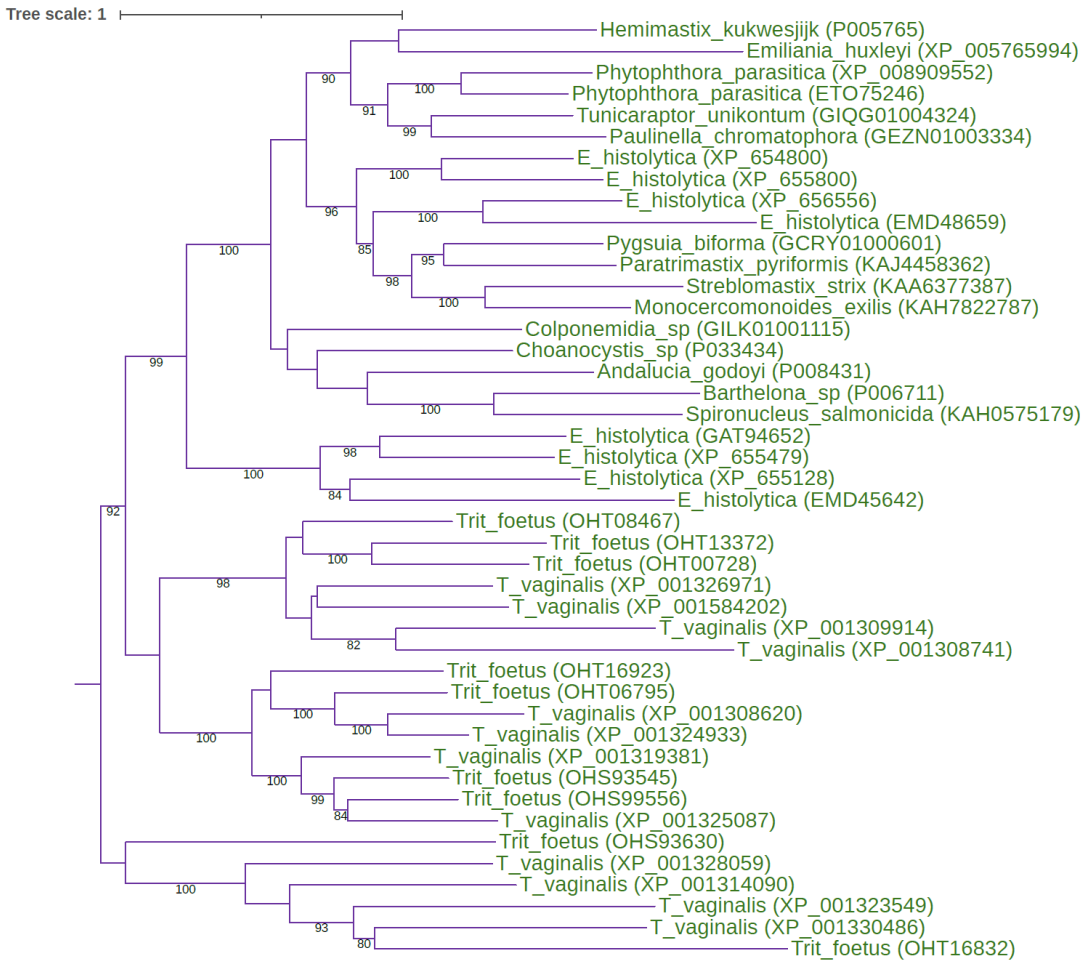

Figure 2. ML tree of the eukaryote-specific type 1 long C1 peptidases The midpoint-rooted ML tree shows the evolutionary relationships inside the eukaryote-specific type 1 long C1 peptidases. The best-fit model according to the Bayesian information criterion was PMB+F+I+G4. The ML tree represents the bootstrap consensus following 1000 replicates. Sequences were obtained from GenBank, and species names and accession numbers are included. Abbreviations: T_vaginalis: *Trichomonas vaginalis*; Trit_foetus: *Tritrichomonas foetus*; E_histolytica: *Entamoeba histolytica*.

Because most of these C1A lineages originated in the LECA by gene duplication, they are referred to as ancestral or ancient eukaryotic paralogs; some other C1As may have originated through horizontal gene transfer (HGT) from bacteria and are referred to as ancestral eukaryotic pseudoparalogs [49]. Thus, phylogenomic analysis of the papain family has provided strong evidence of its origin (i.e., where and when the lineages originated, and from which progenitor). In early evolutionary studies of the papain family, there was much speculation concerning the nature of their ancestor [44,45], and only two classes, namely cathepsins B and L, were recognised [43]. However, our exhaustive analysis of C1A peptidases showed that this widely accepted scheme of evolution of the papain family [43] is insufficient and needs to be revised. Due to either bias and/or a limited number of C1A peptidase representatives, the papain family can neither be adequately classified nor mapped to the origin of its numerous orthologous gene families. In contrast to eukaryotes, no widespread orthologous gene families exist in Bacteria and Archaea. Moreover, by using a diverse set of prokaryotic representatives of the papain family as queries for searching eukaryotic omics databases, we uncovered another hidden repertoire of short and long eukaryotic C1A peptidases. These C1A peptidases were horizontally acquired from bacteria independently several times into a few eukaryotic lineages, such as rotifers, green algae, stramenopiles, and fungi, which will be briefly described later.

 doi:10.20944/preprints202306.1638.v1

2.3. Distribution of the papain family in eukaryotes and prokaryotes

The phyletic distribution pattern of the papain family was analysed, and revealed the presence of this family in Bacteria, Archaea, and Eukaryota (Tables 1–3).

2.3.1. Papain family in eukaryotes

Phylogenomic analysis indicated widespread distribution of the papain family in eukaryotes. However, the distribution patterns of the eight ancestral eukaryotic paralogous C1A lineages in eukaryotes differed (Table 1). Cathepsins B, C, X, and L are widespread throughout eukaryotes (Table 1), with only a few losses in Opisthokonta. Cathepsins B, C, X, and L were lost in Fungi and Rotosphaerida, whereas cathepsins C and X were also lost in Apusozoa (Table 2). Cathepsin F is similarly widespread, with few losses in Picozoa, Pluriformea, Fungi, and Rotosphaerida (Tables 1 and 2). Cathepsin H is also present in major eukaryotic lineages but shows more losses in Picozoa, Hemimastigophora, Telonemia, Ancyromonadida, Malawimonadida, and Metamonada; whereas in Obazoa there was a remarkable number of losses in opisthokonts, namely in Choanoflagellata, Filasterea, Tunicaraptor, Ichthyosporea, Fungi, Rotosphaerida, and in Breviatea (Tables 1 and 2). The 26/29 kDa peptidase is widespread in eukaryotes, with only a few losses in Ancoracysta, Telonemia, and CRuMs, while in opisthokonts, few losses can be seen in Fungi, Rotosphaerida, Filasterea, and Tunicaraptor (Tables 1 and 2). The 26/29 kDa peptidase is widespread in Metazoa, with a loss in the ancestor of Theria (marsupials and placental mammals). It is still present in monotremes (platypus + echidna) but is absent in any other mammalian species. Type 1 C1 long peptidase (500–650 aa long) is a novel eukaryote-specific C1A peptidase with an unusual distribution in eukaryotes, since it was lost in multicellular lineages (animals and plants) (Tables 1 and 2). Although this orthologous gene family is present in all three major eukaryotic domains (Diaphoretickes, Amorphea, and "Excavata"), there were a number of losses in Archaeplastida (plants and relatives), Picozoa, Cryptista, Ancoracysta, Telonemia, CRuMs, Ancyromonadida, and Malawimonadida. In opisthokonts, this C1A peptidase is present only in Tunicaraptor and is absent in metazoans, Fungi, choanoflagellates, Filasterea, Pluriformea, Ichthyosporea, and Rotosphaerida. It is present in Breviatea but has been lost in Apusozoa; these are two sister lineages of opisthokonts and are representatives of the Obazoa domain (Table 2). The largest number of representatives of this novel C1A peptidase is found in the SAR supergroup (especially in oomycetes), Haptophyta, Amoebozoa, and Metamonada.

In addition to these eight ancestral eukaryotic C1A lineages, two orthologous gene families exist in Obazoa. The first is the well-known cathepsin O, which is widespread in metazoans and present throughout opisthokonts (Choanoflagellata, Filasterea, Tunicaraptor, Pluriformea, and Ichthyiosporea). Similar to other ancestral eukaryotic C1A lineages, cathepsin O was lost in Fungi and Rotosphaerida (Table 2). It is also present in Apusozoa but absent from Breviatea. A new, longer C1A peptidase that contains a vWFA domain (vWFA-C1 peptidase) is present in metazoans, and while it is quite widespread in basal metazoans and Protostomia, it is very rare in Deuterostomia (found only in some echinoderms and a few fishes) (Table 2). Besides Metazoa, it can be found only in Fungi and Choanoflagellata. This peptidase has a sparse distribution in other eukaryotic lineages and can be found only in a few isolated cases, namely in green algae, Cryptista, Centroheliozoa, the SAR supergroup, and Discoba. In these lineages, vWFA-C1 peptidases are present in only one or two species; thus, it is most likely that they were horizontally acquired.

2.3.2. Papain family in prokaryotes

Analysis of the distribution of the papain family in Bacteria has shown its widespread presence in all the major phyla. The only exception is the phylum Dyctioglomi, which has genome data for only two species and shows the absence of the papain family (Table 3). Similarly, the papain family is widespread in Archaea and is present in all the major archaeal lineages (Table 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 June 2023 doi:10.20944/preprints202306.1638.v1

Table 3 Papain superfamily is widespread in Bacteria and Archaea

Taxonomic group/phylum	Single domain C1A peptidases	Multidomain C1A peptidases
Bacteria	●	●
Acidobacteria	●	●
Aquificae	●	●
Atribacterota	●	●
Caldiserica/Cryosericota group	●	●
Calditrichaeota	●	●
Chrysiogenetes	●	●
Coprothermobacterota	●	●
Deferribacteres	●	●
Desulfobacterota	●	●
Dictyoglomi	○	○
Elusimicrobia	●	●
FCB group	●	●
-Bacteroidetes/Chlorobi group	●	●
-Fibrobacteres	●	●
-Gemmatimonadetes	●	●
Fusobacteria	●	●
Myxococcota	●	●
Nitrospinae/Tectomicrobia group	●	●
Nitrospirae	●	●
Pseudomonadota (=Proteobacteria)	●	●
PVC group	●	●
-Chlamydiae	●	●
-Lentisphaerae	●	●
-Planctomycetota	●	●
-Verrucomicrobia	●	●
Spirochaetes	●	●
Synergistetes	●	●
Terrabacteria group	●	●
-Actinomycetota	●	●
-Bacillota (=Firmicutes)	●	●
-Chloroflexi	●	●
-Cyanobacteria/Melainabacteria group	●	●
-Deinococcus-Thermus	●	●
-Tenericutes	●	●
Thermodesulfobacteria	●	●
Thermotogae	●	●
Bacteria candidate phyla	●	●
Archaea	●	●
Asgard group	●	●
Thermoplasmatota	●	●
DPANN group	●	●
Euryarchaeota	●	●
TACK group	●	●

Black circle represents presence, while white circle represents absence

The number of C1A peptidases in prokaryotic genomes is very low, they are mostly present as a single sequence, although occasionally up to three sequences may also exist. In some species, only complex multidomain proteins with a C1 domain are present, whereas others possess short C1A peptidases, which, from an evolutionary point of view, are more important as progenitors of eukaryotic C1A peptidases. Although numerous prokaryotic C1A peptidases are present in the form of complex multidomain proteins, they were not analysed in more detail here because they cannot be potential progenitors of eukaryotic C1A peptidases. Short C1A peptidases are widespread in both Bacteria and Archaea (Table 3). However, the simple presence/absence pattern can be misleading, as horizontal gene transfer is quite common between different bacteria and between Bacteria and Archaea [50,51]. Often, a spotty distribution of C1A peptidases in large bacterial phyla is observed. As evident from the MEROPS Db, the distribution of C1A peptidases in Bacteria is not uniform, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 June 2023 doi:10.20944/preprints202306.1638.v1

often only a single species possesses a C1A peptidase. However, the analysis of a large collection of bacterial genomes has often shown a different situation from that of the MEROPS Db data, since genome data covers a much larger taxonomic diversity than the MEROPS Db. Moreover, a brief analysis of the signal peptides in prokaryotic C1A peptidases showed that they are present in extracellular proteins but absent in intracellular proteins.

2.4. Diverse evolutionary forces are reshaping the papain family

Evolution of the C1A peptidases (Table 4) is driven by several forces, the major players are gene duplication, horizontal gene transfer, and gene loss.

Table 4 Evolutionary forces acting on the C1A peptidases

Evolutionary forces	Archaea	Bacteria	Eukaryota	Viruses
Gene duplication	○	○	●	○
Gene loss	●	●	●	●
Horizontal gene transfer	●	●	●	●
Functional diversification	○	○	●	○
Gene/domain fusion	●	●	●	○
C1 domain duplication	○	○	●	○
Loss of active site	○	○	●	○
Change of the peptidase class	○	○	●	○
Lineage-specific expansion	○	○	●	○
Coevolution of peptidases with their inhibitors (arms race)	●	●	●	●
Alternative splicing	○	○	●	○
Pseudogenization	○	○	●	○
Expression divergence	○	○	●	○

Black circle represents presence, while white circle represents absence

2.4.1. Gene duplication

Gene duplication within the papain family is almost restricted to eukaryotes, as only a few cases can be found in prokaryotes. It is well known that gene duplication in prokaryotes is not as common as in eukaryotes and that the major force for prokaryotic adaptation is the acquisition of novel genes by horizontal gene transfer [50,51]. Numerous gene duplications in the eukaryotic ancestor resulted in the emergence of the eight ancestral eukaryotic C1A lineages. A number of bursts in functional diversification, as evidenced by large lineage-specific expansions resulting in large multigene families, have occurred mostly in the cathepsin L orthologous gene family in unicellular eukaryotes, land plants, invertebrates, and in some lineages of placental mammals (e.g., placental cathepsins in rodents) [52–55]. In eukaryotes, the cathepsin B orthologous gene family has remained either as a single gene or as small multigene families that sometimes undergo bursts in functional diversification, such as in aphids [56] or in plants [57]. In contrast to cathepsin B, the cathepsin L orthologous gene family has experienced more complex and dynamic evolution through numerous gene duplications, the majority of which are species-specific. The consequence of these large lineage-specific expansions is that cathepsins L are the most numerous representatives of the papain family.

 doi:10.20944/preprints202306.1638.v1

In two rotifer species, subsequent gene duplications of horizontally acquired genes from cyanobacteria generated large lineage-specific expansions of C1A peptidases, from tens to over hundred sequences per species. Additionally, evolutionarily younger C1A orthologous gene families originated in vertebrates (cathepsins W, K, S, and L2) by gene duplications [58].

2.4.2. Horizontal gene transfer

HGT is a very common process in prokaryotes [50,51,59–61]; thus, it is not surprising that HGT is well represented in the prokaryotic part of the papain family. Many cases of horizontally acquired C1A peptidases can be observed in prokaryotes, and the most obvious cases are present in Archaea. The easiest way to recognise HGT is through a homology search. Unusually high levels of sequence conservation between distantly related organisms are a clear indication of HGT [50,51]. In prokaryotes, HGT occurs in two directions: between diverse bacterial taxa, and between Bacteria and Archaea. We found that Archaea possess 44% Archaea-specific short C1A peptidases, while 56% were acquired through HGT from Bacteria (Suppl. file 1). In Archaea, we found that the majority of the HGT cases of short C1A peptidases are concentrated in methanogenic Archaea (in Euryarchaeota – in the Stenosarchaea and Methanomada groups). Surprisingly, in the evolutionary very important Asgardarchaeota (closest relatives of eukaryotes) [62], we found that the vast majority of short C1A peptidases were horizontally acquired from Bacteria. Based on our numerous homology searches between diverse bacterial taxa, we found that the extent of HGT within Bacteria is similar to that observed in Archaea, although Bacteria possess more of the short C1A peptidases. The biological reason for such a high level of HGT in Archaea and Bacteria is related to the ecology of these organisms, as they prevail in microbial mats and biofilms, where HGT is very common among diverse taxonomic lineages of microbes [63].

We found at least five separate and independent cases of HGT of C1A peptidases from diverse bacteria to eukaryotes (Suppl. file 2). The first case was a recent HGT from cyanobacteria to two rotifer species. The second case was the HGT of short bacterial C1-terB peptidases to the Ascomycota fungi. The third case was the HGT of long bacterial C1A peptidases to a few green algae and two chytrid fungi. The fourth case was the HGT of a long C1A peptidase from Streptomycetes to Ascomycota fungi. The fifth case involves the HGT of short C1A peptidases from bacteria to the SAR supergroup (mostly dinoflagellates). In contrast to the above-described HGT cases, eukaryotes have very few cases of HGT between different eukaryotic lineages. The most obvious case occurs in plant fungal pathogens, where diverse plant cathepsins were horizontally acquired from plant hosts (Suppl. file 2).

We also found several examples of HGT of C1A peptidases from diverse eukaryotes to the DNA viruses, as well as from bacteria to DNA viruses. In the case of HGT of C1A peptidases from eukaryotes to DNA viruses, lepidopteran cathepsin F (also called V-cath peptidase) was acquired by nucleopolyhedroviruses and granuloviruses (Baculoviridae). Cathepsin B was independently acquired horizontally by lymphocystisviruses (Iridoviridae), ascoviruses (Ascoviridae), and algal phaeoviruses (Phycodnaviridae). Short bacterial C1A peptidases (xylellain-type) were independently horizontally acquired by very large DNA viruses (Mimiviridae, Megaviricetes), and by Myoviridae and Siphoviridae (Caudoviricetes) (Suppl. file 2). Most of these HGT cases can also be found in the MEROPS Db, although the genome databases show a larger number of cases.

2.4.3. Gene loss

Gene loss within eukaryotes can be easily recognised since we inferred the ancestral state of the papain family from the presence of eight ancestral eukaryotic C1A peptidases within the same genome. We demonstrated that diverse ancestral C1A peptidases have been lost several times in numerous eukaryotic genomes or in entire taxonomic groups (Tables 1 and 2). We also inferred ancestral states for all eukaryotic supergroups (Table 5). Demonstrating the ancestral state of the papain family is important for the recognition of several independent cases of gene loss in diverse orthologous gene families in distinct eukaryotic taxonomic groups. Some taxonomic groups with very large genome data coverage, such as fungi, have lost all ancestral eukaryotic C1A peptidases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 June 2023 doi:10.20944/preprints202306.1638.v1

The evidence for these C1A losses is based on the analysis of the complete genomes. Gene loss in the papain family can cause problems in the correct interpretation of evolutionary history. However, the large sampling of diversity in the papain family throughout eukaryotes and prokaryotes has enabled us to correctly interpret their evolutionary history.

Table 5 Ancestral states for the eight ancestral eukaryotic paralogous C1A peptidases in major eukaryotic supergroups

	cath B	cath C	cath X	cath L	cath F	cath H	26/29 kDa peptidase	type 1 C1 long
Diaphoretickes	■	■	■	■	■	■	■	■
Amorphea	■	■	■	■	■	■	■	■
CRuMs	■	■	■	■	■	■	□	□
Ancyromonadida	■	■	■	■	■	□	■	□
Malawimonadida	■	■	■	■	■	□	■	□
Discoba	■	■	■	■	■	■	■	■
Metamonada	■	■	■	■	■	□	■	■
LECA	■	■	■	■	■	■	■	■

Black square represents presence, while white square represents absence

2.5. Origin and early diversification of the papain family

From the distribution pattern of C1A peptidases in Archaea and Bacteria, we can infer that they were present in the last universal common ancestor (LUCA) [64] (Table 6). However, HGT may present a problem when inferring the origin of the papain family. However, in Archaea, we found that 44% of the short C1A peptidases were Archaea-specific, but many more unique Bacteria-specific C1A peptidases were present in Bacteria. Thus, despite the vast amount of HGT in prokaryotes, we found sufficient number of short C1A peptidases in Archaea and Bacteria that support their presence in LUCA. The ancestor of the papain family is related to the short prokaryotic C1A peptidases.

Table 6 C1A peptidases that were present in LUCA, FECA, LECA and in the key prokaryotic lineages

	C1A	cath B	cath C	cath X	cath L	cath F	cath H	26/29 kDa peptidase	type 1 C1 long
LECA	□	■	■	■	■	■	■	■	■
FECA	■	□	□	□	□	□	□	□	□
α-Proteobacteria	■	□	□	□	□	□	□	□	□
Cyanobacteria	■	□	□	□	□	□	□	□	□
δ-Proteobacteria	■	□	□	□	□	□	□	□	□
Archaea	■	□	□	□	□	□	□	□	□
LUCA	■	□	□	□	□	□	□	□	□

Black square represents presence, while white square represents absence

Analysis of the presence of C1A peptidases in Asgardarchaeota, which are assumed to be the closest relatives of eukaryotes [62], revealed the presence of several representatives; however, most were horizontally acquired from diverse bacterial taxa. We also found that short C1A peptidases are quite rare in diverse Alphaproteobacteria, especially in Rickettsiales, which are assumed to be the progenitors of mitochondria. In the progenitors of chloroplasts, the Cyanobacteria, we identified a diverse collection of short C1A peptidases. In the syntrophy model of eukaryogenesis [65], Deltaproteobacteria was proposed as an additional taxon involved in eukaryogenesis, together with an archaeal host and a mitochondrial ancestor. We found that Deltaproteobacteria have a diverse collection of short C1A peptidases. Regarding the problems associated with the abundance of HGT events among prokaryotes and the rarity of C1A peptidases in the key taxa important for eukaryogenesis (Asgardarchaeota and Alphaproteobacteria), we may infer that in the first eukaryotic common ancestor (FECA), C1A peptidases may also have been acquired through HGT from some bacterial taxa. In the case of the tripartite syntrophy model, C1A peptidase could have been obtained from Deltaproteobacteria. Owing to the intense HGT in Asgardarchaeota [62], we believe that the C1A peptidase was horizontally acquired from bacteria (either in the bacterial symbiont or in the

archaeal host). Therefore, the ancestor of eukaryotic C1A peptidases was most likely of bacterial origin. Regardless of the origin of the prokaryotic C1A peptidase in FECA (archaea-specific, bacteria-specific, or horizontally acquired bacterial C1A), there is no doubt that only a single ancestor of eukaryotic C1A peptidases existed in FECA. During the long transition period from FECA to LECA [66], additional C1A peptidases may have been horizontally acquired from diverse bacterial taxa. During the transition period between FECA and LECA, intensive reshaping of the C1A repertoire occurred. While eight ancestral C1A lineages were present in LECA, only a single C1A ancestor was present in FECA. Regarding the differences among ancestral C1A paralogs, we can infer a simple origin for the four ancestral C1A lineages that possess the I29 domain. From their progenitor, a series of gene duplications was responsible for the emergence of cathepsins L, F, H, and 26/29 kDa peptidases. In the case of cathepsins B, C, and X, there is also a large structural sequence divergence among them; therefore, we cannot exclude the possibility of their origin through separate HGT events. However, an alternative possibility, which is also supported by the phylogenetic analysis (Figure 1), is that they originated through a few gene duplications. In this case, diverse propeptides were acquired separately (e.g., the exclusion domain in cathepsin C). In the case of cathepsin B, it is evident that some eukaryotic lineages may possess simpler forms without the propeptide domain, which can also be a consequence of the higher sequence divergence that prevents the recognition of this protein domain. Thus, during the transition period from FECA to LECA, many C1A lineages originated, some of which may have been lost. Surprisingly, prokaryotes have few C1A peptidases per species that are present either as single-domain C1As or complex multidomain proteins. From the ancestral eukaryotic C1A repertoire, it is evident that intensive functional diversification is responsible for generating such a large diversity of papain family in the LECA.

2.6. Structure-function relationships in the papain family

The discovery of eight ancestral paralogous C1A lineages is of utmost importance for the recognition of the structure-function relationships within the papain family. Since we know what the ancestral eukaryotic C1A lineages are, we can also use the known data on their functions to make some generalisations. Most of these eukaryotic ancestral C1A lineages have been functionally and biochemically studied in metazoans (including parasites), plants, and protists (mostly in diverse pathogens) [3–6]. Although they have different functions in diverse eukaryotic lineages, these functions are related to lifestyle. In most cases, C1A peptidases are involved in protein degradation, regulate proteolysis, and largely participate in the host-pathogen arms race. In the latter case, they are either used in attack (C1A peptidases of bacteria and eukaryotic pathogens and parasites) or defence (in multicellular eukaryotes), where the papain family represents an important part of the innate immune system [66]. Most of the ancestral C1A lineages have been functionally and biochemically characterised; the only exception is the type 1 long C1 peptidase, for which no such data exist. However, genomic data showed that these peptidases are present in the form of large multigene families (Figure 2). Since they are common in protist pathogens, we can infer that they probably play an important role in host-pathogen arms race, most likely in host invasion. Additionally, the functions of bacterial ancestral-type short C1A peptidases are still scarce and limited to two species [68,69]. 3D structures are available for six eukaryotic ancestral paralogous C1A lineages, for cathepsins B [18], C [25], X [23], L/papain [17,20], F [27], and H [22], while no structures are available for the "insect 26/29 kDa peptidase" or the "type 1 long C1 peptidase", although high-quality models can be obtained with the AlphaFold [70] (Figure 3). Very recently, a LolA/EPDR domain was found in the N-terminal part of the 26/29 kDa peptidases [71]. These peptidases were formed during eukaryogenesis by the fusion of a LolA/EPDR protein with the ancestral cathepsin L. The LolA fold is a barrel-like fold comprising an 11-stranded antiparallel β-sheet with a short helix located within its centre. Although the hydrophobic cavity of the LolA fold represents a possible binding site for the lipid moiety of lipoproteins, its role in the 26/29 kDa peptidases might be different.

 doi:10.20944/preprints202306.1638.v1

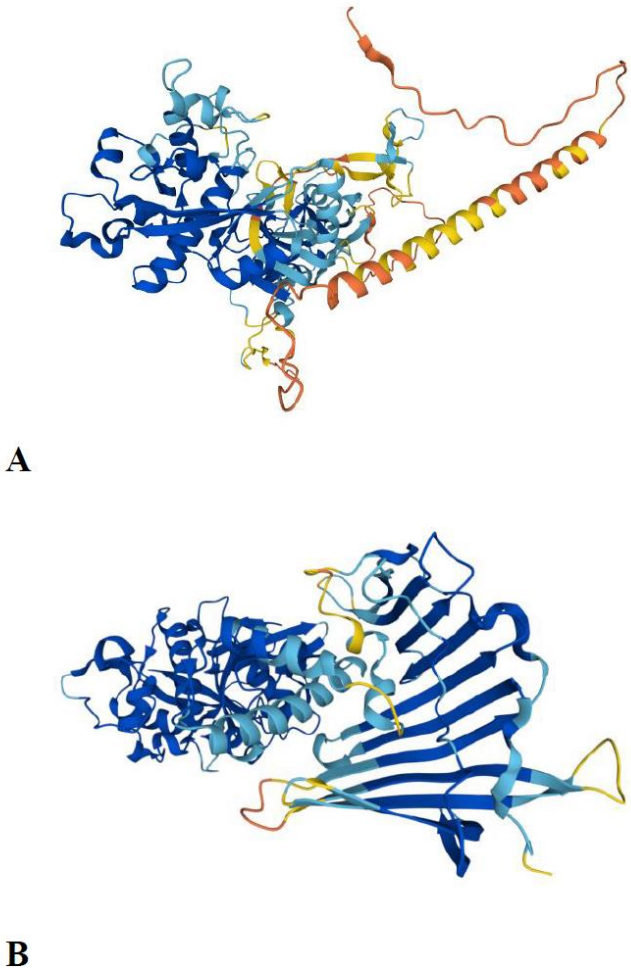

Figure 3. AlphaFold predicted structures of the two ancestral C1A lineages A) AlphaFold-predicted structure of the type 1 long C1 peptidase from *Phytophthora infestans* (A0A833SVU0_PHYIN). The transmembrane domain is located at positions 32–53. B) AlphaFold-predicted structure of the 26/29 kDa peptidase from *Aedes albopictus* (midgut cysteine proteinase 2; A0A182HAH0_AEDAL).

The discovery of the repertoire of ancestral eukaryotic C1A peptidases is important for directing structural studies of the papain family, as its structural diversity in eukaryotes is already well known. Even in prokaryotes, two 3D structures are available for single domain (or short) C1A peptidases [68,69], and one structure is available for multidomain C1A peptidase, where the lectin domain is fused with the C1A domain [72]. For evolutionarily younger eukaryotic C1A lineages (e.g., cathepsin O and vWFA-C1), high-quality models can be easily produced using AlphaFold [70]. Thus, structure-function relationships are known for the majority of eukaryotic ancestral C1A peptidase lineages. It is also important to solve these issues for less-studied C1As or the new C1A ancestral lineage that has not yet been analysed from a structure-function point of view. Thus, the structure-function knowledge of the papain family is in its "ripe form", as it is more or less completely solved.

3. Materials and Methods

3.1. Data mining

All database searches were performed online and were completed in January 2023. The databases analysed were the nonredundant (NR), TSA, WGS, and microbial and eukaryotic genome databases at the NCBI (http://www.ncbi.nlm.nih.gov). Diverse taxon-specific eukaryotic and prokaryotic proteome, transcriptome, and genome databases were searched using the NCBI website.

 doi:10.20944/preprints202306.1638.v1

Attempts were made to identify novel representatives of the papain family in diverse proteomic databases, such as MEROPS (merops.sanger.ac.uk) [1], Superfamily (supfam.org), and InterPro (www.ebi.ac.uk/interpro/). Crucial eukaryotic taxa and lineages not available in the NCBI databases were searched for C1A peptidases in the EukProt V3 transcriptome database (https://evocellbio.com/eukprot/) [48]. To identify all the available representatives of the papain family, database searches were performed iteratively. Comparisons were performed using the TBLASTN and BLASTP tools [73], with an E-value cutoff set to 10^{-5} and default settings for other parameters. Diverse eukaryotic (e.g., human, invertebrate, and eukaryotic pathogens), prokaryotic (bacterial and archaeal), and viral C1A peptidases have been used as queries. The C1 domain in newly discovered representatives of the papain family was identified using the NCBI CDD domain database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). The Translate program (http://www.expasy.org/tools/dna.html) was used to translate DNA sequences.

3.2. Phylogenetic analysis

Due to the high number of sequences available from the papain family, only representatives of the C1A peptidases from eukaryotic supergroups and a few prokaryotic representatives of the papain family were included in the analyses. In this way the stabile backbone phylogeny was obtained. Protein sequences were aligned using Clustal Omega [74]. Phylogenetic trees were reconstructed using the ML method [75]. The reliability of the resulting topologies was evaluated using 1000 bootstrap replications. Diverse prokaryotic representatives of the papain family were used as an outgroup. Phylogenetic analyses were performed using the program IqTree [75], whereas tree visualisation was performed using iTOL v6 [76].

4. Conclusions

We performed a comprehensive phylogenomic analysis of the papain family, using extensive proteomic, transcriptomic, and genomic data from the Archaea, Bacteria, and Eukaryota, and obtained new insights into the origin and evolution of C1A peptidases. In contrast to the widely accepted view that eukaryotes possess only two ancestral C1A lineages, the cathepsin B and L classes [43], our study shows that such a view is limited. Here, we demonstrated that eight ancestral eukaryotic paralogous C1A peptidase lineages were present in the ancestor of eukaryotes. These eight ancestral eukaryotic C1A peptidase lineages are cathepsins B, C, X, L, H, F, 26/29 kDa peptidase, and type 1 long C1 peptidase and are present in all eukaryotic supergroups. The key findings of our study report that the papain family was present in the LUCA and that this family was already highly diversified in the LECA. Altogether, our study provides an in-depth understanding of the diversity and evolution of the papain family.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions DK and VT conceived the study design. DK collected the sequence data and performed the bioinformatic, evolutionary, and phylogenomic analyses. Both authors wrote the manuscript and read and approved the final version of it.

Funding DK was supported by grant P1-0207 and VT by grants P1-0140 and J1-2473 from the Slovenian Research Agency.

Acknowledgments We thank Dr. Gregor Gunčar for the recognition of the LolA-like fold in the insect 26/29 kDa peptidase.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement The data presented in this study are available in the article and supplementary material. All new sequence data for the papain family will be sent to the MEROPS database.

Conflicts of Interest The authors declare no conflicts of interest. The funders had no role in the design of the study; collection, analysis, or interpretation of data; writing of the manuscript; or the decision to publish the results.

Abbreviations FECA: first eukaryotic common ancestor; HGT, horizontal gene transfer; LECA, last eukaryotic common ancestor; LUCA, last universal common ancestor; Db, database; cath, cathepsin.

References

1. Rawlings, N.D.; Barrett, A.J.; Thomas, P.D.; Huang, X.; Bateman, A.; Finn, R.D. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **2018**, *46(D1)*, D624-D632.
2. Rawlings, N.D.; Bateman, A. Origins of peptidases. *Biochimie* **2019**, *166*, 4-18.
3. Sajid, M.; McKerrow, J.H. Cysteine proteases of parasitic organisms. *Mol. Biochem. Parasitol.* **2002**, *120*, 1-21.
4. Shindo, T.; Van der Hoorn, R.A. Papain-like cysteine proteases: key players at molecular battlefields employed by both plants and their invaders. *Mol. Plant Pathol.* **2008**, *9*, 119-125.
5. Misas-Villamil, J.C.; van der Hoorn, R.A.; Doehlemann, G. Papain-like cysteine proteases as hubs in plant immunity. *New Phytol.* **2016**, *212*, 902-907.
6. van der Hoorn, R.A.; Jones, J.D. The plant proteolytic machinery and its role in defence. *Curr. Opin. Plant Biol.* **2004**, *7*, 400-407.
7. Unanue, E.R.; Turk, V.; Neefjes, J. Variations in MHC Class II Antigen Processing and Presentation in Health and Disease. *Annu. Rev. Immunol.* **2016**, *34*, 265-297.
8. Stoka, V.; Turk, V.; Turk, B. Lysosomal cathepsins and their regulation in aging and neurodegeneration. *Ageing Res. Rev.* **2016**, *32*, 22-37.
9. Gocheva, V.; Zeng, W.; Ke, D.; Klimstra, D.; Reinheckel, T.; Peters, C.; Hanahan, D.; Joyce, J.A. Distinct roles for cysteine cathepsin genes in multistage tumorigenesis. *Genes Dev.* **2006**, *20*, 543-556.
10. Olson, O.C.; Joyce, J.A. Cysteine cathepsin proteases: regulators of cancer progression and therapeutic response. *Nat. Rev. Cancer* **2015**, *15*, 712-729.
11. Zhang, X.; Luo, S.; Wang, M.; Shi, G.P. Cysteinyl cathepsins in cardiovascular diseases. *Biochim. Biophys. Acta Proteins Proteom.* **2020**, *1868*, 140360.
12. Turk, B.; Stoka, V. Protease signalling in cell death: caspases versus cysteine cathepsins. *FEBS Lett.* **2007**, *581*, 2761-2767.
13. Turk, B.; Turk, D.; Turk, V. Protease signalling: the cutting edge. *EMBO J.* **2012**, *31*, 1630-1643.
14. Repnik, U.; Stoka, V.; Turk, V.; Turk, B. Lysosomes and lysosomal cathepsins in cell death. *Biochim. Biophys. Acta* **2012**, *1824*, 22-33.
15. Ketterer, S.; Gomez-Auli, A.; Hillebrand, L.E.; Petrera, A.; Ketscher, A.; Reinheckel, T. Inherited diseases caused by mutations in cathepsin protease genes. *FEBS J.* **2017**, *284*, 1437-1454.
16. López-Otín, C.; Bond, J.S. Proteases: multifunctional enzymes in life and disease. *J. Biol. Chem.* **2008**, *283*, 30433-30437.
17. Drenth, J.; Jansonius, J.N.; Koekoek, R.; Swen, H.M.; Wolthers, B.G. Structure of papain. *Nature* **1968**, *218*, 929-932.
18. Musil, D.; Zucic, D.; Turk, D.; Engh, R.A.; Mayr, I.; Huber, R.; Popovic, T.; Turk, V.; Towatari, T.; Katunuma, N.; et al. The refined 2.15 A X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity. *EMBO J.* **1991**, *10*, 2321-2330.
19. McGrath, M.E.; Klaus, J.L.; Barnes, M.G.; Brömme, D. Crystal structure of human cathepsin K complexed with a potent inhibitor. *Nat. Struct. Biol.* **1997**, *4*, 105-109.
20. Fujishima, A.; Imai, Y.; Nomura, T.; Fujisawa, Y.; Yamamoto, Y.; Sugawara, T. The crystal structure of human cathepsin L complexed with E-64. *FEBS Lett.* **1997**, *407*, 47-50.
21. Guncar, G.; Pungercic, G.; Klemencic, I.; Turk, V.; Turk, D. Crystal structure of MHC class II-associated p41 Ii fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S. *EMBO J.* **1999**, *18*, 793-803.
22. Guncar, G.; Podobnik, M.; Pungercar, J.; Strukelj, B.; Turk, V.; Turk, D. Crystal structure of porcine cathepsin H determined at 2.1 A resolution: location of the mini-chain C-terminal carboxyl group defines cathepsin H aminopeptidase function. *Structure* **1998**, *6*, 51-61.
23. Guncar, G.; Klemencic, I.; Turk, B.; Turk, V.; Karaoglanovic-Carmona, A.; Juliano, L.; Turk, D. Crystal structure of cathepsin X: a flip-flop of the ring of His23 allows carboxy-monopeptidase and carboxy-dipeptidase activity of the protease. *Structure* **2000**, *8*, 305-313.
24. Somoza, J.R.; Zhan, H.; Bowman, K.K.; Yu, L.; Mortara, K.D.; Palmer, J.T.; Clark, J.M.; McGrath, M.E. Crystal structure of human cathepsin V. *Biochemistry* **2000**, *39*, 12543-12551.
25. Turk.; D, Janjić, V.; Stern, I.; Podobnik, M.; Lamba, D.; Dahl, S.W.; Lauritzen, C.; Pedersen, J.; Turk, V.; Turk, B. Structure of human dipeptidyl peptidase I (cathepsin C): exclusion domain added to an endopeptidase framework creates the machine for activation of granular serine proteases. *EMBO J.* **2001**, *20*, 6570-6582.

26. Turkenburg, J.P.; Lamers, M.B.; Brzozowski, A.M.; Wright, L.M.; Hubbard, R.E.; Sturt, S.L.; Williams, D.H. Structure of a Cys25-->Ser mutant of human cathepsin S. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 451-455.
27. Somoza, J.R.; Palmer, J.T.; Ho, J.D. The crystal structure of human cathepsin F and its implications for the development of novel immunomodulators. *J. Mol. Biol.* **2002**, *322*, 559-568.
28. Turk, B.; Turk, D.; Turk, V. Lysosomal cysteine proteases: more than scavengers. *Biochim. Biophys. Acta* **2000**, *1477*, 98-111.
29. Dolenc, I.; Turk, B.; Pungercic, G.; Ritonja, A.; Turk, V. Oligomeric structure and substrate induced inhibition of human cathepsin C. *J. Biol. Chem.* **1995**, *270*, 21626-21631.
30. Dolenc, I.; Štefe, I.; Turk, D.; Taler-Verčič, A.; Turk, B.; Turk, V.; Stoka, V. Human cathepsin X/Z is a biologically active homodimer. *Biochim. Biophys. Acta Proteins Proteom.* **2021**, *1869*, 140567.
31. Turk, V.; Stoka, V.; Vasiljeva, O.; Renko, M.; Sun, T.; Turk, B.; Turk, D. Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochim. Biophys. Acta* **2012**, *1824*, 68-88.
32. Jerala, R.; Zerovnik, E.; Kidric, J.; Turk, V. pH-induced conformational transitions of the propeptide of human cathepsin L. A role for a molten globule state in zymogen activation. *J. Biol. Chem.* **1998**, *273*, 11498-11504.
33. Rozman, J.; Stojan, J.; Kuhelj, R.; Turk, V.; Turk, B. Autocatalytic processing of recombinant human procathepsin B is a bimolecular process. *FEBS Lett.* **1999**, *459*, 358-362.
34. Turk, B.; Bieth, J.G.; Björk, I.; Dolenc, I.; Turk, D.; Cimerman, N.; Kos, J.; Colic, A.; Stoka, V.; Turk, V. Regulation of the activity of lysosomal cysteine proteinases by pH-induced inactivation and/or endogenous protein inhibitors, cystatins. *Biol. Chem. Hoppe Seyler* **1995**, *376*, 225-230.
35. Vasiljeva, O.; Dolinar, M.; Pungercar, J.R.; Turk, V.; Turk, B. Recombinant human procathepsin S is capable of autocatalytic processing at neutral pH in the presence of glycosaminoglycans. *FEBS Lett.* **2005**, *579*, 1285-1290.
36. Dahl, S.W.; Halkier, T.; Lauritzen, C.; Dolenc, I.; Pedersen, J.; Turk, V.; Turk, B. Human recombinant pro-dipeptidyl peptidase I (cathepsin C) can be activated by cathepsins L and S but not by autocatalytic processing. *Biochemistry* **2001**, *40*, 1671-1678.
37. Barrett, A.J. The cystatins: a diverse superfamily of cysteine peptidase inhibitors. *Biomed. Biochim. Acta* **1986**, *45*, 1363-1374.
38. Turk, V.; Bode, W. The cystatins: protein inhibitors of cysteine proteinases. *FEBS Lett.* **1991**, *285*, 213-219.
39. Turk, B.; Stoka, V.; Björk, I.; Boudier, C.; Johansson, G.; Dolenc, I.; Colic, A.; Bieth, J.G.; Turk, V. High-affinity binding of two molecules of cysteine proteinases to low-molecular-weight kininogen. *Protein Sci.* **1995**, *4*, 1874-1880.
40. Turk, B.; Stoka, V.; Turk, V.; Johansson, G.; Cazzulo, J.J.; Björk, I. High-molecular-weight kininogen binds two molecules of cysteine proteinases with different rate constants. *FEBS Lett.* **1996**, *391*, 109-112.
41. Turk, V.; Stoka, V.; Turk, D. Cystatins: biochemical and structural properties, and medical relevance. *Front. Biosci.* **2008**, *13*, 5406-5420.
42. Kordiš, D.; Turk, V. Phylogenomic analysis of the cystatin superfamily in eukaryotes and prokaryotes. *BMC Evol. Biol.* **2009**, *9*, 266.
43. Karrer, K.M.; Peiffer, S.L.; DiTomas, M.E. Two distinct gene subfamilies within the family of cysteine protease genes. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 3063-3067.
44. Hughes, A.L. Evolution of cysteine proteinases in eukaryotes. *Mol. Phylogenet. Evol.* **1994**, *3*, 310-321.
45. Berti, P.J.; Storer, A.C. Alignment/phylogeny of the papain superfamily of cysteine proteases. *J. Mol. Biol.* **1995**, *246*, 273-283.
46. Burki, F.; Roger, A.J.; Brown, M.W.; Simpson, A.G.B. The New Tree of Eukaryotes. *Trends Ecol. Evol.* **2020**, *35*, 43-55.
47. Hug, L.A.; Baker, B.J.; Anantharaman, K.; Brown, C.T.; Probst, A.J.; Castelle, C.J.; Butterfield, C.N.; Hernsdorf, A.W.; Amano, Y.; Ise, K.; et al. A new view of the tree of life. *Nat. Microbiol.* **2016**, *1*, 16048.
48. Richter, D.J.; Berney, C.; Strassert, J.F.H.; Poh, Yu-Ping; Herman, E.K.; Muñoz-Gómez, S.A.; Wideman, J.G.; Burki, F.; de Vargas, C. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community J.* **2022**, *2*, e56.
49. Makarova, K.S.; Wolf, Y.I.; Mekhedov, S.L.; Mirkin, B.G.; Koonin, E.V. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* **2005**, *33*, 4626-4638.
50. Arnold, B.J.; Huang, I.T.; Hanage, W.P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* **2022**, *20*, 206-218.
51. Gophna, U.; Altman-Price, N. Horizontal gene transfer in Archaea - from mechanisms to genome evolution. *Annu. Rev. Microbiol.* **2022**, *76*, 481-502.
52. Richau, K.H.; Kaschani, F.; Verdoes, M.; Pansuriya, T.C.; Niessen, S.; Stüber, K.; Colby, T.; Overkleeft, H.S.; Bogyo, M.; Van der Hoorn, R.A. Subclassification and biochemical analysis of plant papain-like cysteine proteases displays subfamily-specific characteristics. *Plant Physiol.* **2012**, *158*, 1583-1599.

53. Liu, J.; Sharma, A.; Niewiara, M.J.; Singh, R.; Ming, R.; Yu, Q. Papain-like cysteine proteases in *Carica papaya*: lineage-specific gene duplication and expansion. *BMC Genomics* **2018**, *19*, 26.
54. Robinson, M.W.; Dalton, J.P.; Donnelly, S. Helminth pathogen cathepsin proteases: it's a family affair. *Trends Biochem. Sci.* **2008**, *33*, 601-608.
55. Mason, R.W. Emerging functions of placental cathepsins. *Placenta* **2008**, *29*, 385-390.
56. Rispe, C.; Kutsukake, M.; Doublet, V.; Hudaverdian, S.; Legeai, F.; Simon, J.C.; Tagu, D.; Fukatsu, T. Large gene family expansion and variable selective pressures for cathepsin B in aphids. *Mol. Biol. Evol.* **2008**, *25*, 5-17.
57. McLellan, H.; Gilroy, E.M.; Yun, B.W.; Birch, P.R.; Loake, G.J. Functional redundancy in the Arabidopsis cathepsin B gene family contributes to basal defence, the hypersensitive response and senescence. *New Phytol.* **2009**, *183*, 408-418.
58. Zhou J, Zhang YY, Li QY, Cai ZH. Evolutionary History of Cathepsin L (L-like) Family Genes in Vertebrates. *Int. J. Biol. Sci.* **2015**, *11*, 1016-1025.
59. Akanni, W.A.; Siu-Ting, K.; Creevey, C.J.; McInerney, J.O.; Wilkinson, M.; Foster, P.G.; Pisani, D. Horizontal gene flow from Eubacteria to Archaebacteria and what it means for our understanding of eukaryogenesis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2015**, *370*, 20140337.
60. Sheinman, M.; Arkhipova, K.; Arndt, P.F.; Dutilh, B.E.; Hermsen, R.; Massip, F. Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. *Elife* **2021**, *10*, e62719.
61. Beiko, R.G.; Harlow, T.J.; Ragan, M.A. Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14332-14337.
62. Spang, A.; Stairs, C.W.; Dombrowski, N.; Eme, L.; Lombard, J.; Caceres, E.F.; Greening, C.; Baker, B.J.; Ettema, T.J.G. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* **2019**, *4*, 1138-1148.
63. Bolhuis, H.; Cretoiu, M.S.; Stal, L.J. Molecular ecology of microbial mats. *FEMS Microbiol. Ecol.* **2014**, *90*, 335-350.
64. Weiss, M.C.; Preiner, M.; Xavier, J.C.; Zimorski, V.; Martin, W.F. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet.* **2018**, *14*, e1007518.
65. López-García, P.; Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat. Microbiol.* **2020**, *5*, 655-667.
66. McInerney, J.O.; O'Connell, M.J.; Pisani, D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* **2014**, *12*, 449-455.
67. Brix, K.; Dunkhorst, A.; Mayer, K.; Jordans, S. Cysteine cathepsins: cellular roadmap to different functions. *Biochimie* **2008**, *90*, 194-207.
68. Leite, N.R.; Faro, A.R.; Dotta, M.A.; Faim, L.M.; Gianotti, A.; Silva, F.H.; Oliva, G.; Thiemann, O.H. The crystal structure of the cysteine protease Xylellain from *Xylella fastidiosa* reveals an intriguing activation mechanism. *FEBS Lett.* **2013**, *587*, 339-344.
69. Gong, X.; Zhao, X.; Zhang, W.; Wang, J.; Chen, X.; Hameed, M.F.; Zhang, N.; Ge, H.; Structural characterization of the hypothetical protein Lpg2622, a new member of the C1 family peptidases from *Legionella pneumophila*. *FEBS Lett.* **2018**, *592*, 2798-2810.
70. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583-589.
71. Wei, Y.; Xiong, Z.J.; Li, J.; Zou, C.; Cairo, C.W.; Klassen, J.S.; Privé, G.G. Crystal structures of human lysosomal EPDR1 reveal homology with the superfamily of bacterial lipoprotein transporters. *Commun. Biol.* **2019**, *2*, 52.
72. Bradshaw, W.J.; Kirby, J.M.; Thiyagarajan, N.; Chambers, C.J.; Davies, A.H.; Roberts, A.K.; Shone, C.C.; Acharya, K.R. The structure of the cysteine protease and lectin-like domains of Cwp84, a surface layer-associated protein from *Clostridium difficile*. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, *70*, 1983-1993.
73. Gertz, E.M.; Yu, Y.K.; Agarwala, R.; Schäffer, A.A.; Altschul, S.F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* **2006**, *4*, 41.
74. Sievers, F.; Higgins, D.G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **2018**, *27*, 135-145.
75. Trifinopoulos, J.; Nguyen, L.T.; von Haeseler, A.; Minh, B.Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 2016, *44(W1)*, W232-235.
76. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 2021, *49(W1)*, W293-W296.