CLCAs - A Family of Metalloproteases of Intriguing Phylogenetic Distribution and with Cases of Substituted Catalytic Sites

Anna Lenart[1], **Małgorzata Dudkiewicz**[2], **Marcin Grynberg**[3], **Krzysztof Pawłowski**[1,2]*

1 Department of Cellular and Molecular Neurobiology, Nencki Institute of Experimental Biology, Polish Academy of Sciences, Warsaw, Poland, **2** Faculty of Agriculture and Biology, Warsaw University of Life Sciences, Warsaw, Poland, **3** Department of Genetics, Institute of Biochemistry and Biophysics, Polish Academy of Sciences,Warsaw, Poland

Abstract

The zinc-dependent metalloproteases with His-Glu-x-x-His (HExxH) active site motif, zincins, are a broad group of proteins involved in many metabolic and regulatory functions, and found in all forms of life. Human genome contains more than 100 genes encoding proteins with known zincin-like domains. A survey of all proteins containing the HExxH motif shows that approximately 52% of HExxH occurrences fall within known protein structural domains (as defined in the Pfam database). Domain families with majority of members possessing a conserved HExxH motif include, not surprisingly, many known and putative metalloproteases. Furthermore, several HExxH-containing protein domains thus identified can be confidently predicted to be putative peptidases of zincin fold. Thus, we predict zincin-like fold for eight uncharacterised Pfam families. Besides the domains with the HExxH motif strictly conserved, and those with sporadic occurrences, intermediate families are identified that contain some members with a conserved HExxH motif, but also many homologues with substitutions at the conserved positions. Such substitutions can be evolutionarily conserved and non-random, yet functional roles of these inactive zincins are not known. The CLCAs are a novel zincin-like protease family with many cases of substituted active sites. We show that this allegedly metazoan family has a number of bacterial and archaeal members. An extremely patchy phylogenetic distribution of CLCAs in prokaryotes and their conserved protein domain composition strongly suggests an evolutionary scenario of horizontal gene transfer (HGT) from multicellular eukaryotes to bacteria, providing an example of eukaryote-derived xenologues in bacterial genomes. Additionally, in a protein family identified here as closely homologous to CLCA, the CLCA_X (CLCA-like) family, a number of proteins is found in phages and plasmids, supporting the HGT scenario.

Citation: Lenart A, Dudkiewicz M, Grynberg M, Pawłowski K (2013) CLCAs - A Family of Metalloproteases of Intriguing Phylogenetic Distribution and with Cases of Substituted Catalytic Sites. PLoS ONE 8(5): e62272. doi:10.1371/journal.pone.0062272

Editor: Shoba Ranganathan, Macquarie University, Australia

Received November 20, 2012; **Accepted** March 19, 2013; **Published** May 9, 2013

Funding: AL, MD, MG and KP were supported by the grant N N301 3165 33 from the Ministry of of Science and Higher Education, Republic of Poland. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: krzysztof_pawlowski@sggw.pl

Introduction

The protein sequence space, recently becoming sampled more and more densely thanks to genomic and metagenomic sequencing projects, has undoubtedly 'granular' features, and can be classified using various algorithms and classification systems [1,2]. Yet, it has also features of continuity, with very distant sequence similarities discovered between hitherto unrelated protein families, and local structural similarities found between members of different folds [3,4]. The protein sequence/structure space, arguably, is not fully sampled during evolution. Further, the still incomplete charting of the protein universe is expected to be biased by technology and prevailing research trends [5,6]. Even for the charted regions of the sequence space, many evolutionarily-justified similarity relationships are not obvious and often are found only after solving three-dimensional structures [7] or applying sophisticated bioinformatics approaches [8–10].

Here, we focus on a broad clan of metalloproteases that are a good example of protein class with such a dual granular and continuous features. The proteases, originally noted for their involvement in digestive processes, are now acknowledged for many crucial regulatory roles in cellular signalling in diverse biological processes, on cellular, tissue and organism scale, e.g. in cell proliferation and differentiation, inflammation, tissue remodelling, neurogenesis, angiogenesis, apoptosis, wound healing, blood coagulation [11–14]. Not surprisingly, proteases constitute an important class of drug targets [15–17]. Among the generic class of proteases, distinct clans have been identified using the catalytic mechanism and three-dimensional fold as the classifier [18,19]. The zinc ion-dependent zincin-like metalloproteases grouped in the MA clan in the MEROPS database include 37 families [20], while in the Pfam database there are 52 families of the Peptidase_MA clan, including also putative metalloproteases [21,22]. The proteins containing the zincin-like domains often feature complex domain composition reflecting their biological functions [23].

Here, we explore the realm of all proteins identified by the simple HExxH active site motif common to most MA clan member families and show topology features of the sequence similarity network of the clan families. Also, we show that the motif can be used as a prefilter for discovery of novel metalloproteases.

CLCAs are a protein family implicated in several pathologies in humans, including asthma, chronic obstructive pulmonary disease (COPD) and cancer [24,25]. Originally, they were believed to be calcium-activated chloride channels [26,27]. Despite their characterisation as putative metalloproteases several years ago [28], they attracted moderate interest. The current view is that they are involved in regulation of calcium-activated chloride currents [29]. Several members of the CLCA family have been characterised beyond any doubt as secreted zinc-dependent metalloproteases [30,31] that perform self-cleavage at a conserved site. Vertebrates possess several closely homologous CLCA genes (usually 3–6), the functional relationships between them are not fully elucidated. It is not known whether CLCAs possess other physiological substrates except themselves, whether they are cleaved by other proteases except themselves, and whether different CLCA proteins cleave each other [32,33]. It is now believed that activation of ion channels by CLCA proteins occurs via a direct protein-protein interaction between an ion channel molecule and the N-terminal fragment of a CLCA protein, an interaction possible only after CLCA self-cleavage [31].

Recently, cases of patchy phylogenetic distribution of homologues of human genes in prokaryotes have attracted some attention, [9,10,23]. Such distribution has been interpreted as potential sign of horizontal gene transfer (HGT) [34–37].

In this article, first, we survey the HExxH proteins and identify domain families with majority of members containing the motif. Second, we analyse the substituted HExxH motifs in families where they are conserved in most members. Third, we provide support for the hypothesis of eukaryote to prokaryote horizontal gene transfer in CLCA proteins. Lastly, we present the prokaryote-specific CLCA-like domain family (CLCA_X) and present an overall representation of sequence similarity topology of the zincin-like clan.

Results and Discussion

Survey of HExxH motif-containing proteins

The ubiquitous HExxH zinc-binding motif is a hallmark of zinc-dependent metalloproteases [38–41]. We surveyed the Trembl database and found 151223 occurrences of the motif compared to 80000 expected by chance (significant, almost twofold over-representation, p-value of the binomial text $<<10^{-10}$, see Methods). Then, we checked whether the occurrences of this motif were within the known Pfam protein domains (Pfam database version 24.0), or outside those.

After removal of redundancy in the hit sequence set at 90% sequence identity, the occurrence of the HExxH motif within Pfam domains was 47794 (versus 41946 expected) which makes up 52% of the occurrences, while the occurrence outside Pfam domains was 43395 (versus 49242 expected). Thus, since approx. 46% of the total length of Trembl proteins lie within the Pfam domains, within these domains the HExxH motif is found significantly more often (p-value of the binomial text $<<10^{-10}$, see Methods) than expected by chance, while outside of the domains it is found significantly less often than expected.

The regions of protein sequence databases unassigned to known protein domains (e.g. Pfam) can be unassigned for two reasons: first, they may constitute novel, yet undescribed domains, second, they may belong to special regions (e.g. transmembrane segments, low-complexity regions, disordered regions, unique variable regions and so forth). Hence, it can be expected that some HExxH motifs found here outside Pfam domains actually do occur in yet undiscovered protein domains, possibly in novel protease domains. Search for novel metalloprotease domains is out of the scope of this article, however the existence of yet undescribed zincin domains can be argued for by the rapid increase in the numbers of zincin domains described in domain databases (for example, in the Pfam database, the Peptidase_MA clan grew from 36 families in release 24, 2009, to 52 families in release 26, 2011 [42]. Also, recently, some novel zincin families have been characterised [23,30]. Examples of zincin-like metalloprotease domains, suspected but not described formally yet, include the family of the ddrB protein of *E. coli* bacteriophage P1 [43] and the CLCA_X family mentioned herein.

Since the HExxH motif is found in protein sequences twice as often as expected by chance, even if the HExxH metalloproteases had not been known, one would have expected some functional relevance of the motif. Yet, obviously, some of the occurrences have to be due to chance. One way of separating the functional HExxH motifs from the random ones is identifying those motifs that are conserved by evolution. To this extent, we sought protein domain families, for which significant fraction of family members had the motif present in a conserved position.

The Pfam domains are clearly split into those that have few (sporadic, random) occurrences of HExxH and those that possess the motif, as a rule, in a conserved position (see Fig. 1). Indeed, all the domains with majority of members having a HExxH motif can be predicted by sensitive sequence analysis methods (FFAS) to possess the structures of zincin-like metalloproteases (see Table 1). Thus, zincin-like fold can be predicted for eight structurally uncharacterised Pfam domains possessing the HExxH motif. One of them is the fatty acid desaturase domain (FA_desaturase, PF00487) that is present in proteins coded by eight human genes [44,45]. The remaining seven are DUF462, DUF922, DUF1025, DUF2248, DUF2342, DUF3267 and SprA-related (PF12118) domains. Metalloprotease function is most likely conserved in these families, as judged from the conservation of active site motifs (see Fig. 2).

Substituted HExxH sites in metalloproteases

Both among the HExxH proteins that are characterised as metalloproteases and those predicted as such, many domains (approximately half of the cases) have active site motifs occasionally "broken down", i.e. with substitutions at one of the critical positions, His, Glu or second His, see the protein families with less than 100% of conserved HExxH motif in Fig. 1. The domain families with significant fraction of substituted active sites occur in all domains of life, bacteria, eukaryotes and archaea alike.

Interestingly, the substituted motifs exhibit non-random substitution patterns (see Fig. 2). Both the first and the second histidine residues are significantly more often than expected by chance replaced with positively charged arginine and lysine. Generally in protein sequences, histidine is most often replaced by glutamine. The glutamate residue of the HExxH motif in the zincin-like proteins is most often replaced by glutamine, as generally in proteins. The biological role of lysine- or arginine-replaced histidines in the substituted active sites are not clear since these positively charged residues are not expected to participate in zinc ion binding. The replacement frequencies of the critical histidine and glutamate residues in substituted HExxH proteins deviate largely from the general replacement frequencies observed in proteins for these residues.

Several substitutions do occur in HExxH motifs at least twice as often as in proteins in general (see Table S1). The first histidine residue of the motif is approximately twofold more often than in an average protein substituted by Glu or Arg. The second histidine

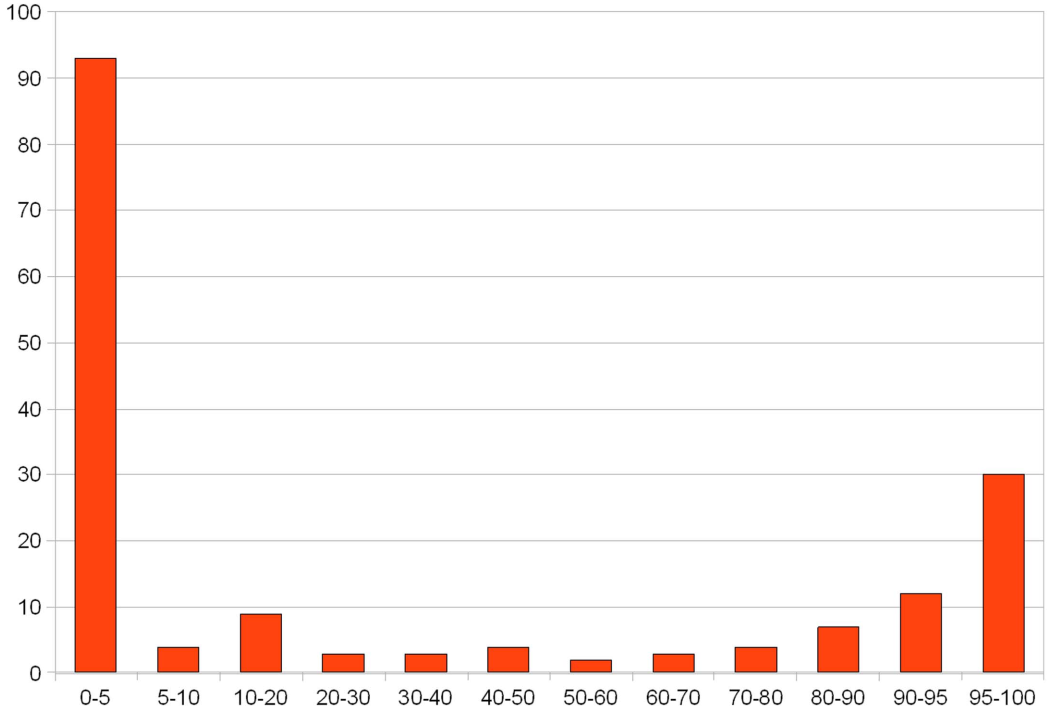


Figure 1. Histogram. Pfam protein domains binned by percentage of family members that possess the HExxH motif. Shown are only the domains with more than 50 occurrences of the motif included plus all Peptidase domains with at least one occurrence.
doi:10.1371/journal.pone.0062272.g001

Table 1. Structure predictions for domain families with majority of members possessing the HExxH motif.

Query	% of domain family members possessing the HExxH motif	FFAS *Z-score*	% sequence identity	top FFAS hit	First prediction of a zincin-like structure
CLCA_N (PF08434) Calcium-activated chloride channel	72	−6,3	11	d1kufa_ d.92.1.9 (A:) Snake venom metalloprotease [Trimeresurus mucrosquamatus], atrolysin E	[28]
SprA-related (PF12118) SprA-related family	95	−7,7	11	PF05569.4; Q8RPJ4_DESHA/7–280; BlaR1 peptidase M56	this article
FA_desaturase (PF00487) Fatty acid desaturase	37	−8,4	16	d1k7ia2 d.92.1.6 (A:18–258) Metalloprotease [Erwinia chrysanthemi]	this article
Metallopep (PF12044) Putative peptidase family	87	−12,3	20	d1c7ka_ d.92.1.1 (A:) Zinc protease [Streptomyces caespitosus]	Pfam annotation
MtfA (PF06167) Phosphoenolpyruvate: glucose-phosphotransferase regulator	100	−43,9	11	d1j7na2 d.92.1.14 (A:551–773) Anthrax toxin lethal factor, N- and C-terminal domains [Bacillus anthracis]	[7]
DUF2248 (PF10005) Uncharacterized protein conserved in bacteria	100	−13,9	10	d1j7na2 d.92.1.14 (A:551–773) Anthrax toxin lethal factor, N- and C-terminal domains [Bacillus anthracis]	this article
DUF2265 (PF10023) Predicted amin opeptidase	100	−7,7	11	d3b7sa3 d.92.1.13 (A:209–460) Leukotriene A4 hydrolase catalytic domain [Homo sapiens]	Pfam annotation
DUF462 (PF04315) Protein of unknown function	100	−7	13	d1j7na2 d.92.1.14 (A:551–773) Anthrax toxin lethal factor, N- and C-terminal domains [Bacillus anthracis]	this article
DUF922 (PF06037) Bacterial protein of unknown function	100	−7,4	10	d1kjpa d.92.1.2 (A:) Thermolysin [Bacillus thermoproteolyticus]	this article
DUF3267 (PF11667) Protein of unknown function	83	−10,8	12	d1asta_ d.92.1.8 (A:) Astacin [Astacus astacus)]	this article
DUF1025 (PF06262) Domain of unknown function	76	−43,4	34	d3e11a1 d.92.1.17 (A:1–113) Uncharacterized protein Acel_2062 [Acidothermus cellulolyticus]	this article
DUF2342 (PF10103) Uncharacterized conserved protein	43	−93,8	19	d3cmna1 d.92.1.16 (A:43–391) Uncharacterized protein Caur0242 [Chloroflexus aurantiacus]	this article

doi:10.1371/journal.pone.0062272.t001

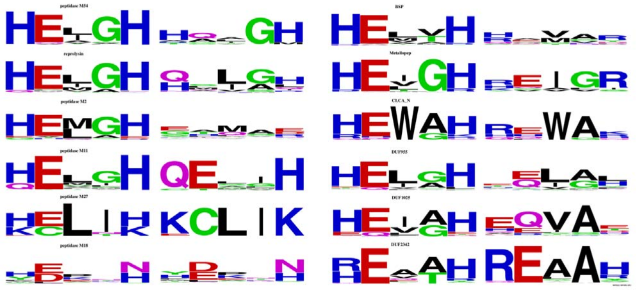


Figure 2. Sequence logos of substituted and conserved active site motifs in selected zincin-like families.
doi:10.1371/journal.pone.0062272.g002

residue is also approximately twofold more often replaced with Lys or Arg and almost fivefold more often - by Leu. The catalytic Glu is very often replaced by Ala, Leu and Gln (2-, 3- and 3-fold, respectively). These substitutions can be divided into two categories: first, missense mutations resulting from a single-nucleotide change in a codon: H→R, H→Q, H→L and E→Q, and second, amino acid substitutions requiring two nucleotide changes in the corresponding codon: H→K, H→E, E→L.

The common active site substitutions are not spread evenly among the HExxH metallopeptidase families (see Fig. 2). The H1→R change (substitution of first histidine of the motif by an arginine) is found often in CLCA_N peptidases and in BSPs (Basic Secretory Proteins). The H1→E change is found often in peptidases M2. The H5→L change occurs in reprolysins, while the H5→K change – in CLCA_N peptidases and reprolysins. The substitution E2→Q occurs in peptidases M54 and in BSPs (Basic Secretory Proteins) and the change E2→L is seen in reprolysins. Interestingly, only some of the unusually frequent active site substitutions are biochemically conservative, retaining the hydrophilic/charged properties of a residue (e.g. H→R, H→K, E→Q). Intriguingly, other frequent substitutions do change strongly the properties of the amino acid residue from hydrophilic and/or charged to hydrophobic (H→L, E→L). Thus, probably active site substitutions observed in HExxH proteins are driven by more than one biological mechanism. Some substitutions may allow a metalloprotease to partly retain its biochemical properties (e.g. zinc ion binding) while other changes may definitely abrogate the original activity.. The roles of inactive HExxH metalloproteases are not well-understood. The best studied are some of the mammalian ADAM family members that have HExxH substitutions such as LQxxL or HQxxH, and are involved in sperm-egg interactions in the fertilisation process [46,47]. Although the molecular function of the probably inactive proteases is mysterious, they can be expected to act as decoys, mimmicking other, proteolytically active ADAM paralogues, or as accessory proteins to their active counterparts.

The CLCA family

The CLCA proteins, defined herein as those possessing the CLCA_N peptidase domain (PF08134) are one of the zincin-like metalloprotease families identified in this study as having the active site motif substituted in a number of cases (see Fig. 2).

A survey of CLCA_N domains shows widespread presence throughout *Metazoa* including early branching Plocozoan (*Trichoplax*), with notable absence in some model organisms like *Drosophila* or *Caenorhabditis* [28]. No CLCA_N domains were found in other eukaryotic taxa including *Fungi*, plants or amoebae. An analysis of distribution of substituted and correct active site motifs in a phylogenetic tree of selected representatives of the CLCA_N domain suggests that loss of proper active site in the CLCA family occurred most likely independently five times in specific lineages (see Fig. 3) rather than once in an ancestral CLCA_N domain. Despite multiple occurrences of CLCA proteins in various organisms, they often represent lineage-specific expansions, e.g. human and Plocozoan multiple CLCAs all originate from single proteins specific to their respective lineages (see Fig. S1).

Interestingly, CLCA_N domains were also found in a number of prokaryotic genomes, including one archaeal genus (*Thermococcus*) and a handful of bacterial strains from several bacterial phyla. Strikingly, the CLCA_N-possessing strains are scattered throughout several main bacterial phyla (see Table 2 and Fig. 4). Thus, CLCA proteins are found in *Chloroflexi*, *Synergistetes*, alpha-, gamma- and delta-proteobacteria, but not in more than three five strains from each phylum.

The typical domain architecture of a metazoan CLCA protein, as shown in Figs. 4 and 5, is repeated in bacterial homologues from alpha-, gamma- and delta- Proteobacteria, as well as *Synergistetes*. Other prokaryotic CLCA proteins (archaeal and those from *Chloroflexi*) have the CLCA_N domain thrown into a different, species-specific domain architecture (see Fig. 5). Interestingly, the typical CLCA domain architecture is also present in two species, the gamma-proteobacterium *Teredinibacter* and the euryarcheon *Methanosarcina*, with the CLCA_N domain replaced by two unrelated enzymatic (peptidase) domains, Peptidase_M10 and CHAP, respectively [22,48].

Conservation of the active site motif and general sequence conservation in the CLCA_N family suggests its peptidase function is broadly conserved (see the multiple alignment in Fig. 6, top, and Fig. S2). The cysteine-rich domain following the CLCA_N core metalloprotease domain (approx. residues 200–260 in human CLCA1 protein, see the multiple alignments) has no detectable homologues in known proteins and may be hypothesized to be involved in stabilisation of the peptidase domain and/or substrate binding [31].

The CLCA_N domains of the bacterial CLCA proteins are grouped together in a typical phylogenetic tree of the CLCA_N domains from different organisms (see Fig. 3), however, their precise phylogenetic relation to metazoan CLCA_N domains, hence their origin, remains unclear. The variability of sequences in the family precludes a reliable phylogenetic tree that could suggest the origin of the putative HGT event.

It is argued that phylogenetic analysis using DNA sequences, using the codon alignment derived from protein sequence alignment can be more reliable than the corresponding analysis using only protein sequences [49]. In the Fig. S3, two CLCA metalloprotease domain phylogenetic trees are presented side-by-side: one derived from mRNA sequences and the other derived from protein sequences, both built using the same protein sequence alignment. The nucleotide sequence tree has generally much better bootstrap values. The trees are generally similar in grouping sequences from major Metazoan taxa (e.g. vertebrates, insects, crustaceans), however they differ in placement of several highly diverged sequences that may have been subject to specific evolutionary pressures. Also, the two trees differ in placement of prokaryotic sequences whereas only the protein sequence tree groups all the prokaryotic sequences together. However, the low bootstrap values even for the nucleotide sequence tree preclude a reliable elucidation of the origin of the putative HGT of CLCA domain from Metazoa to prokaryotes. Also, because of low bootstrap values and the presence of highly divergent sequences (e.g. the lower branches in Fig. S3, upper part), the molecular

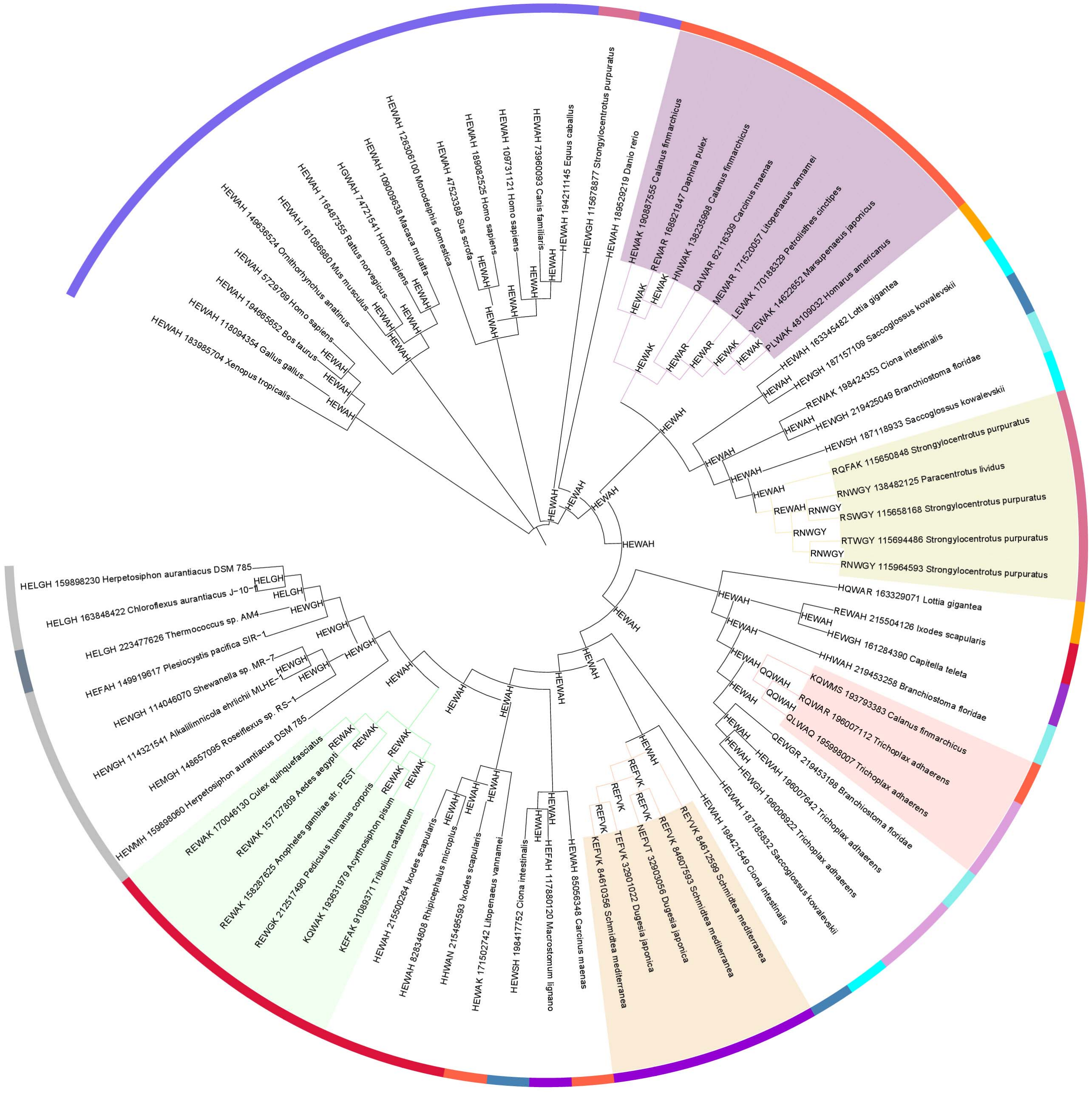

HEWAH 183985704 Xenopus tropicalis
HEWAH 118094354 Gallus gallus
HEWAH 194665652 Bos taurus
HEWAH 5729769 Homo sapiens
HEWAH 149636524 Ornithorhynchus anatinus
HEWAH 161086980 Mus musculus
HEWAH 116487355 Rattus norvegicus
HGWAH 74721541 Homo sapiens
HEWAH 109096638 Macaca mulatta
HEWAH 126306100 Monodelphis domestica
HEWAH 47523388 Sus scrofa
HEWAH 189082525 Homo sapiens
HEWAH 109731121 Homo sapiens
HEWAH 73960093 Canis familiaris
HEWAH 194211145 Equus caballus
HEWGH 115678877 Strongylocentrotus purpuratus
HEWAH 189529219 Danio rerio
HEWAK 190887555 Calanus finmarchicus
REWAR 168921847 Daphnia pulex
HNWAK 139235998 Calanus finmarchicus
QAWAR 62118309 Carcinus maenas
MEWAR 171520057 Litopenaeus vannamei
LEWAK 170188329 Petrolisthes cinctipes
YEWAK 14622652 Marsupenaeus japonicus
PLWAK 48109032 Homarus americanus
HEWAH 163345482 Lottia gigantea
HEWGH 187157109 Saccoglossus kowalevskii
REWAK 198424353 Ciona intestinalis
HEWGH 219425049 Branchiostoma floridae
HEWSH 187118933 Saccoglossus kowalevskii
RQFAK 115650848 Strongylocentrotus purpuratus
RNWGY 138482125 Paracentrotus lividus
RSWGY 115658168 Strongylocentrotus purpuratus
RTWGY 115694486 Strongylocentrotus purpuratus
RNWGY 115964593 Strongylocentrotus purpuratus
HQWAR 163329071 Lottia gigantea
REWAH 215504126 Ixodes scapularis
HEWGH 161284390 Capitella teleta
HHWAH 219453258 Branchiostoma floridae
KQWMS 193793383 Calanus finmarchicus
RQWAR 196007112 Trichoplax adhaerens
QLWAQ 195998007 Trichoplax adhaerens
QEWGR 219453198 Branchiostoma floridae
HEWAH 196007642 Trichoplax adhaerens
HEWGH 196006922 Trichoplax adhaerens
HEWAH 187185832 Saccoglossus kowalevskii
HEWAH 198421549 Ciona intestinalis
REYVK 84612599 Schmidtea mediterranea
REFVK 84607593 Schmidtea mediterranea
NEFVT 32903056 Dugesia japonica
TEFVK 32901022 Dugesia japonica
KEFVK 84610356 Schmidtea mediterranea
HEWAH 85056348 Carcinus maenas
HEFAH 117880120 Macrostomum lignano
HEWSH 198417752 Ciona intestinalis
HEWAK 171502742 Litopenaeus vannamei
HHWAN 215495593 Ixodes scapularis
HEWAH 82834808 Rhipicephalus microplus
HEWAH 215500264 Ixodes scapularis
KEFAK 91089371 Tribolium castaneum
KQWAK 193631979 Acyrthosiphon pisum
REWGK 212517490 Pediculus humanus corporis
REWAK 158287625 Anopheles gambiae str. PEST
REWAK 157127809 Aedes aegypti
REWAK 170046130 Culex quinquefasciatus
HEWMH 159898060 Herpetosiphon aurantiacus DSM 785
HEMGH 148657095 Roseiflexus sp. RS-1
HEWGH 114321541 Alkalilimnicola ehrlichii MLHE-1
HEWGH 114046070 Shewanella sp. MR-7
HEFAH 149919617 Plesiocystis pacifica SIR-1
HELGH 223477626 Thermococcus sp. AM4
HELGH 163848422 Chloroflexus aurantiacus J-10-fl
HELGH 159898230 Herpetosiphon aurantiacus DSM 785

Figure 3. Phylogenetic tree (ANCESCON, see Methods) of selected representatives of the CLCA_N domain. Locations of proteins with substituted and correct active site motifs. Also predicted active sites of ancestral sequences shown.
doi:10.1371/journal.pone.0062272.g003

clock is not applicable here for estimation of time of the likely horizontal gene transfer event recorded in the tree,

The habitats of CLCA_N domain-possessing prokaryotic strains are strikingly similar. These organisms are all aquatic and free living, usually aerobic, however they differ in temperature preferences and energy sources used (see Table 2). Taking together the taxonomic spread of CLCA proteins in prokaryotes and the conservation of their multi-domain composition, the most plausible evolutionary scenario seems to be that of horizontal gene transfer from eukaryotes (*Metazoa*) to bacteria. This direction of the transfer is most likely because of the ubiquity of CLCAs in *Metazoa* and their paucity in prokaryotes. Of note, an automated approach for detection of phylogenetically atypical genes in has identified a CLCA homologue from *Shewanella* as a HGT candidate [50]. Because all the CLCA-possessing organisms live in aquatic environments, it may be hypothesised that the horizontal gene transfer of a metazoan CLCA gene to bacterium occurred in an aquatic milieu.

Another argument in favour of a HGT scenario could have been conservation of CLCA genomic neighbourhoods in prokaryotes, however, no such conservation can be observed here. Genomic neighbourhood similarities are usually restricted to the species level. However, protein families annotated with some common functional themes can be found in the neighbourhoods, e.g. protease genes (COG1988, predicted membrane-bound metal-dependent hydrolases and COG4955, distant caspase

Table 2. Habitats and lifestyles of bacteria and archaea possessing CLCA proteins.

Species, strain	phylum	environment	lifestyle	oxygen requirement	energy source	thermo-philic
Alkalilimnicola ehrlichii MLHE-1	γ-proteobacteria	aquatic	free living	facultative anaerobic	chemoautotroph	−
Aminomonas paucivorans	*Synergistetes*	aquatic/sewage	free living	anaerobic	chemoautotroph	−
Chloroflexus aurantiacus J-10-fl	*Chlorflexi*	aquatic/hot springs	free living	facultative aerobic	photoautotroph	+
Citreicella sp. SE45	α-proteobacteria	aquatic	free living	aerobic	chemoautotroph	−
Desulfobulbus propionicus DSM 2032	δ-proteobacteria	aquatic/marine sediments	free living	anaerobic	chemoautotroph	−
Herpetosiphon aurantiacus DSM 785	*Chlorflexi*	aquatic	free living	aerobic	chemotroph	−
Plesiocystis pacifica SIR-1	δ-proteobacteria	marine/aquatic	free living	aerobic	chemoheterotroph	−
Roseiflexus sp. RS-1	*Chlorflexi*	aquatic	free living	aerobic	phototroph	+
Shewanella sp. MR-4	γ-proteobacteria	aquatic	free living	facultative anaerobie	heterotroph	mesophile (30–40°C)
Thermococcus_sp._AM4	*Euryarchaeota (Archaea)*	aquatic	free living	facultative anaerobie	chemoautotroph	hyper-thermofile

doi:10.1371/journal.pone.0062272.t002

homologues). Also, present in the neighbourhoods are COG2199 (diguanylate cyclase with PAS/PAC sensor), COG VicK (histidine kinase), COG Baes (sensor histidine kinase), COG3899 (sensor histidine kinase), COG-NtrB (signal transduction histidine kinase, nitrogen specific). The protease and signalling domains present in bacterial CLCA neighbourhoods are reminiscent of the vertebrate CLCA extracellular protease functions, including the regulatory functions.

A homologous gene acquired by a host by the way of HGT and whose evolution therefore does not match the evolution of its host organism has been termed xenologue [51]. The hypothetical eukaryote-to-bacteria gene transfer described here is obviously not the first known case of eukaryote-derived xenologues in bacterial genomes. Recently, three-dimensional structures of two virulence factors from *Bacteroidetes* bacteria have been solved. Both three-dimensional structures turned out to be metalloproteases. Their structural features and sequence similarity relationships strongly suggested these proteins had been acquired from mammals by a bacterial pathogen [52,53]. Although the prevalence of HGT between eukaryotes and prokaryotes has only recently been appreciated [54], its crucial importance for the evolution of bacteria, archaea and viruses has been known for a decade and HGT is now established as one of key mechanisms modulating the classic Darwinian mechanisms of evolution [51,55,56].

The CLCA_X (CLCA-like) family

Among the distant subsignificant hits in PSI-BLAST sequence similarity searches for CLCA proteins, a recurring sequence could be noted: a 260 residue-long protein annotated as hypothetical protein Maqu_3852, NCBI gi:120556756, from gamma-proteobacterium *Marinobacter aquaeolei* VT8. It turned out to be a founding member of a large group of bacterial and viral proteins, termed herein CLCA_X. In the Pfam database, most of these could be assigned to the uncharacterised Pfam-B_1042 domain family. Similarity between CLCA_N protease domain and CLCA_X is significant, as proved by HHalign alignment between HHsenser-generated profiles of the two groups (alignment p-value 2.3E-05, see Fig. 7).

Charting the CLCA_X family using HHsenser brought 153 proteins, after CD-hit clustering at 95 sequence identity. Clustering at 70% sequence identity yielded 115 representative sequences. A substantial part of these, were annotated as viral proteins, in general, only 12 out of 115 proteins did not hit any mobile genomic elements (phage, prophage or plasmid sequences) in the ACLAME database (see list in Table S2).

For groups of distantly related protein families, phylogenetic trees are in general not feasible. The approximate topology of sequence similarity networks can be visualized by graph approaches utilising sets of pairwise similarities, e.g. CLANS [57]. In order to locate the CLCA_N and CLCA_X families within the context of zinc dependent metalloproteases of the zincin fold, we applied the CLANS clustering algorithm to the complete set of families of the Peptidase_MA clan and several more families that were identified as related to them (see Results, first section).

In the CLANS clustering graph, the CLCA_X group locates consistently as a sister group to CLCA_N (see Fig. 8). Even using various significance thresholds for the CLANS analysis, one obtains a consistent picture whereas CLCA_X is clustered together with CLCA_N and close to the central zincin-like families. The known protease family closest to CLCAs is the Peptidase_M64 (IgA peptidase) [58], a secreted protease present in many bacterial strains that have humans as hosts, including pathogenic bacteria. Then, the next Pfam family most similar to CLCA_X was PF04298 Zn_peptidase_2, an uncharacterized bacterial family, present mostly in *Firmicutes* and *Bacteroidetes*. Further, relatively closely to CLCA_X occurred the mostly bacterial uncharacterised family DUF955, and the ubiquitous Peptidase_M3 family of secreted proteases (e.g. neurolysins) present in prokaryotes and eukaryotes alike, including humans [59,60]. Among the above-mentioned metalloprotease families, those characterized (Peptidase_M64, Peptidase_M3 and CLCA_N) are known to act as secreted proteases.

CLCA_X proteins are found in several bacterial phyla, partly those that possess CLCA proteins with CLCA_N domains, namely alpha-, beta-, gamma-, delta-proteobacteria, as well as *Synergistetes*, *Spirochaetes* and *Firmicutes*. Also, *Caudovirales* viruses possess CLCA_X proteins. However, CLCA_N and CLCA_X domains

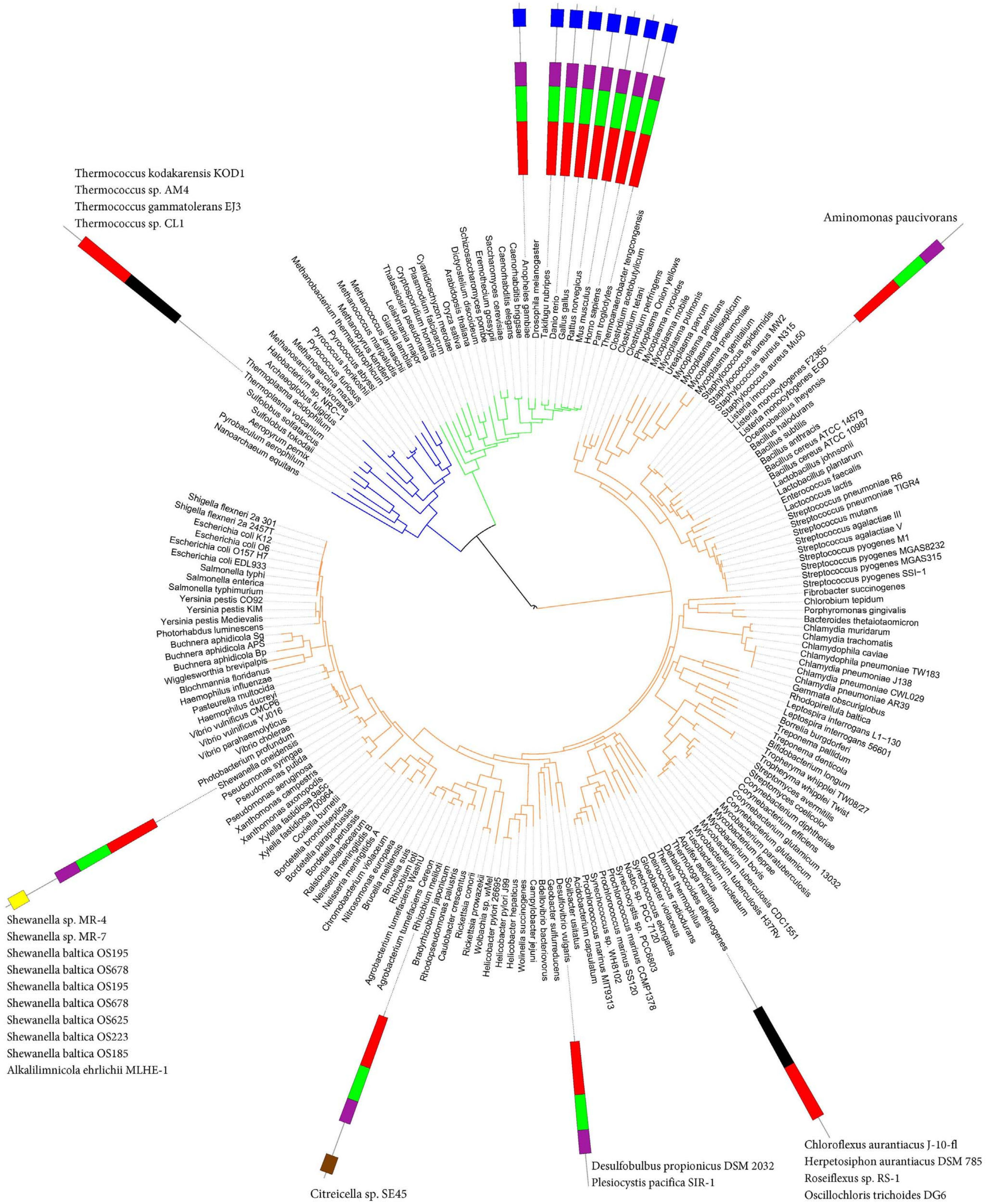


Figure 4. Tree of Life, i.e. representative species tree (adapted from iTOL [77]), with approximate locations of CLCA protein-possessing organisms shown. Schematic diagrams of domain architectures shown also: red, CLCA_N; green, von Willebrand factor type A; magenta, DUF1973; blue, fibronectin type III; black, other.
doi:10.1371/journal.pone.0062272.g004

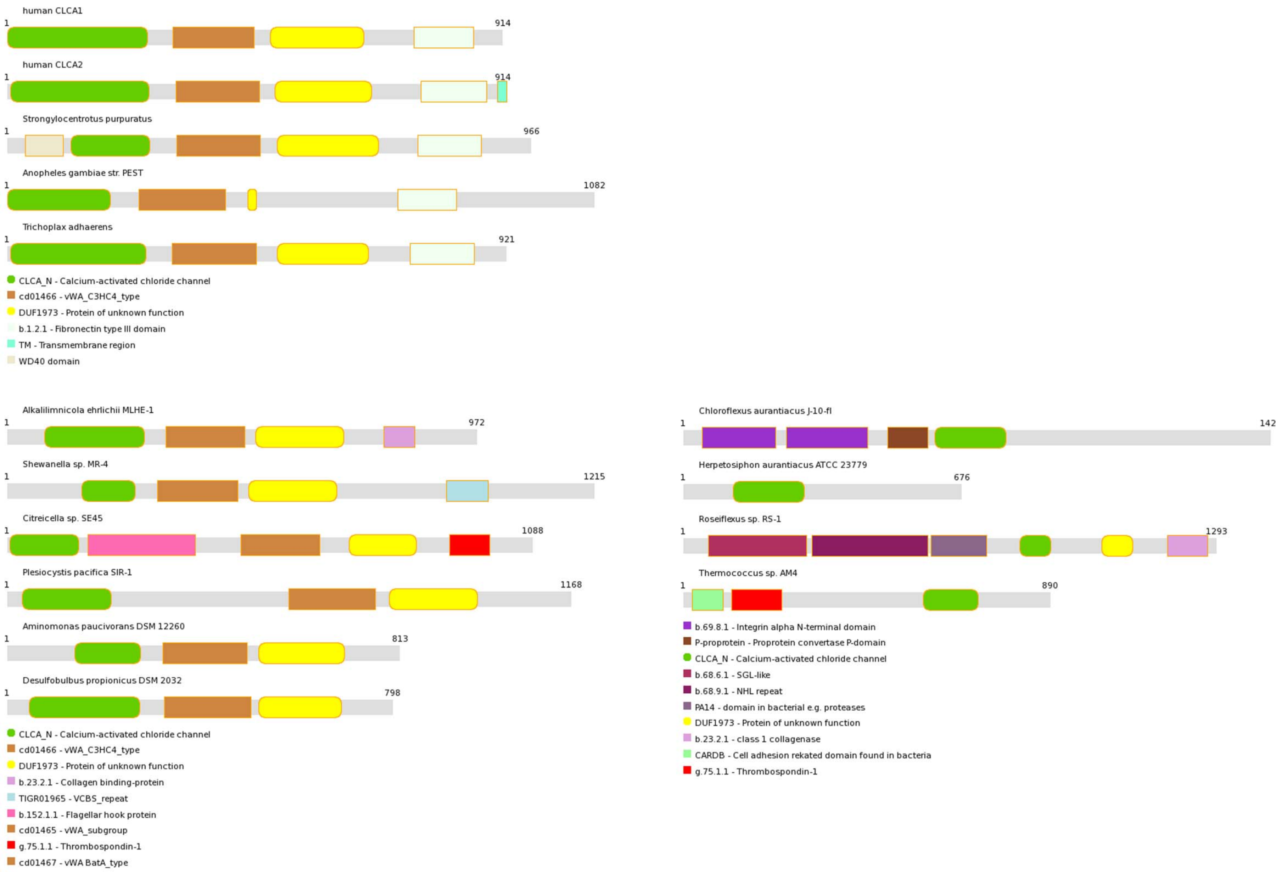

human CLCA1
1
914
human CLCA2
1
914
Strongylocentrotus purpuratus
1
966
Anopheles gambiae str. PEST
1
1082
Trichoplax adhaerens
1
921
CLCA_N - Calcium-activated chloride channel
cd01466 - vWA_C3HC4_type
DUF1973 - Protein of unknown function
b.1.2.1 - Fibronectin type III domain
TM - Transmembrane region
WD40 domain
Alkalilimnicola ehrlichii MLHE-1
1
972
Shewanella sp. MR-4
1
1215
Citreicella sp. SE45
1
1088
Plesiocystis pacifica SIR-1
1
1168
Aminomonas paucivorans DSM 12260
1
813
Desulfobulbus propionicus DSM 2032
1
798
CLCA_N - Calcium-activated chloride channel
cd01466 - vWA_C3HC4_type
DUF1973 - Protein of unknown function
b.23.2.1 - Collagen binding-protein
TIGR01965 - VCBS_repeat
b.152.1.1 - Flagellar hook protein
cd01465 - vWA_subgroup
g.75.1.1 - Thrombospondin-1
cd01467 - vWA BatA_type
Chloroflexus aurantiacus J-10-fl
1
1424
Herpetosiphon aurantiacus ATCC 23779
1
676
Roseiflexus sp. RS-1
1
1293
Thermococcus sp. AM4
1
890
b.69.8.1 - Integrin alpha N-terminal domain
P-proprotein - Proprotein convertase P-domain
CLCA_N - Calcium-activated chloride channel
b.68.6.1 - SGL-like
b.68.9.1 - NHL repeat
PA14 - domain in bacterial e.g. proteases
DUF1973 - Protein of unknown function
b.23.2.1 - class 1 collagenase
CARDB - Cell adhesion rekated domain found in bacteria
g.75.1.1 - Thrombospondin-1

Figure 5. Protein domain architectures (Pfam) of selected CLCA proteins.
doi:10.1371/journal.pone.0062272.g005

are not found in the same species with the exception of *Shewanella* sp. MR-7. The sequence conservation within the CLCA_X family and similarity to CLCA_N domains suggests a conserved protease function and similar active site architectures.

The CLCA_X active site conforms to a consensus HExxHxxxD, only somewhat similar to the CLCA consensus, which is HExxHxxxGxxDEY (see Fig. 6, bottom, and Fig. S4), whereas either the aspartate or the second glutamate residue in the latter motif has been proposed as a likely additional ligand of the zinc ion [31]. The conserved aspartate in the CLCA_X active site motif may perform the same role. The CLCA active site motif is remarkably similar to the active site of known Peptidase_M64 proteins, the HExxHxxxxLxDEY motif [58,61]. In a recently solved structure of a Peptidase_M64 metalloprotease (PDB code 3P1V), a second zinc ion is present, liganded by E of the conserved DEY motif, and by three strongly conserved cysteines, located 90–100 residues away from the HExxH motif towards the C-terminus. Thus, such a role for the DEY motif in the CLCA_N can be postulated.

Although many CLCA_X proteins are long (e.g. more than half of the CLCA_X proteins are longer than 500 residues, more than sixty are 1000 residues or longer), almost none of them contains any known protein domains, suggesting possible existence of completely novel protease auxiliary domains.

Conclusions

The CLCA proteins receive continued attention due to their medical and biological relevance [62–64]. The details of the catalytic metalloprotease activity of CLCA are being elucidated, as well as mechanisms of ion channel activation [31].

Here, we argue for an atypical evolutionary scenario of HGT, from multicellular eukaryotes to bacteria and archaea. Such transfers have been described previously, yet they involved *Wolbachia*, intracellular parasites of *Drosophila* [54,65]. A CLCA protein from the bacterium *Shewanella* has been identified previously as a putative HGT gene [50]. Although the horizontal gene transfer of CLCA genes is only a hypothesis, it seems to be the best explanation of the phylogenetic CLCA distribution observed. Thus, study of distant prokaryotic homologues of CLCA may shed light on its biological functions in *Metazoa*, including humans.

We also argue that the current catalogues of proteins families, including enzymes, are still incomplete and insufficient, as shown by our metalloprotease structure and function prediction for eight uncharacterised Pfam families (see Table 1). The number of HExxH motifs and HExxH motif-containing families suggests that other HExxH metalloprotease families may remain undicovered. We also expect that prevalence and functional importance of inactive homologues of known enzymes may be under-appreciated.

Methods

Survey of HExxH proteins

The Trembl database, version 04.2010, counting 10706472 sequences was used, and the HExxH proteins were extracted using PrositeScan [66]. The proteins with motifs identified were

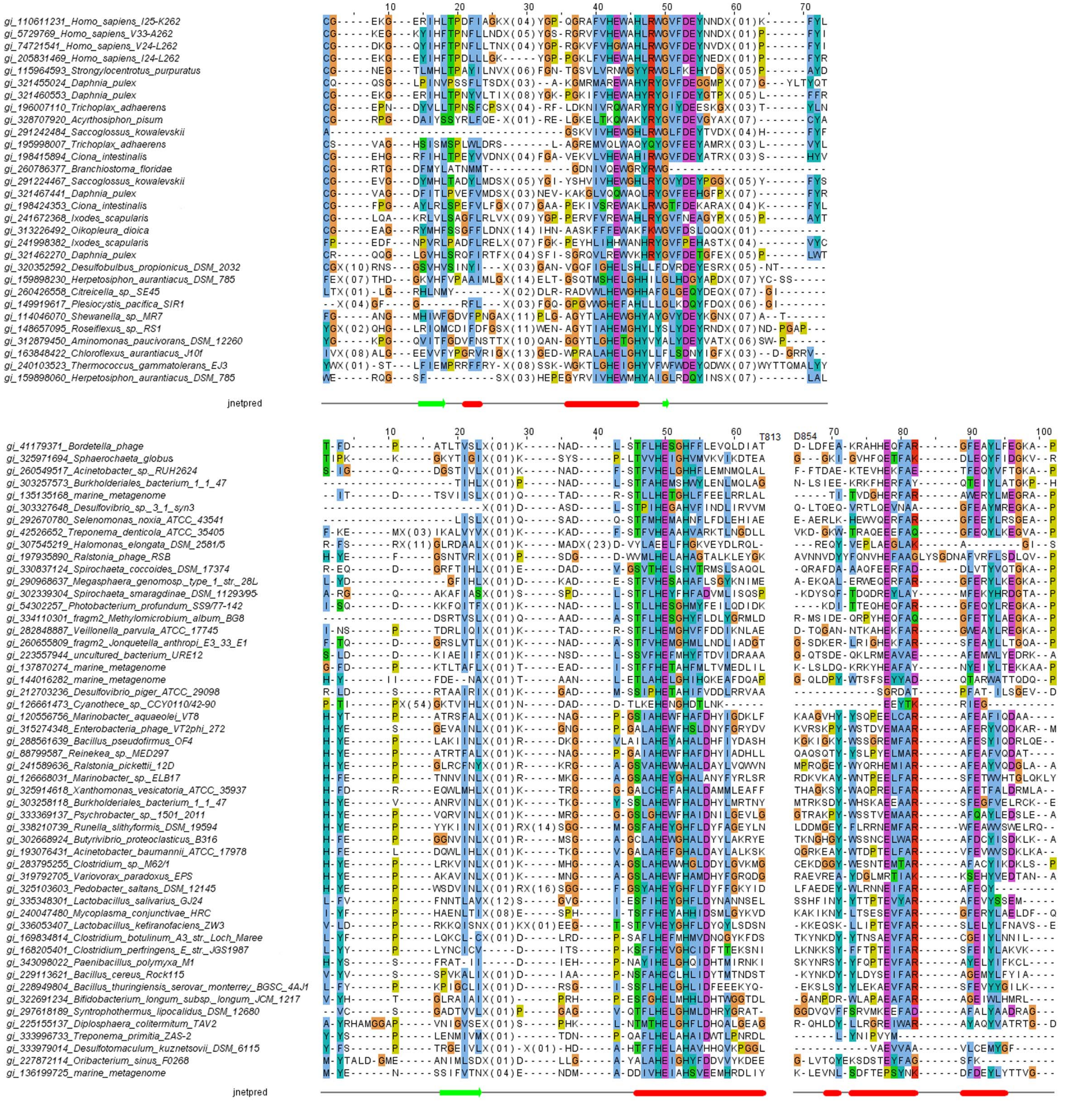
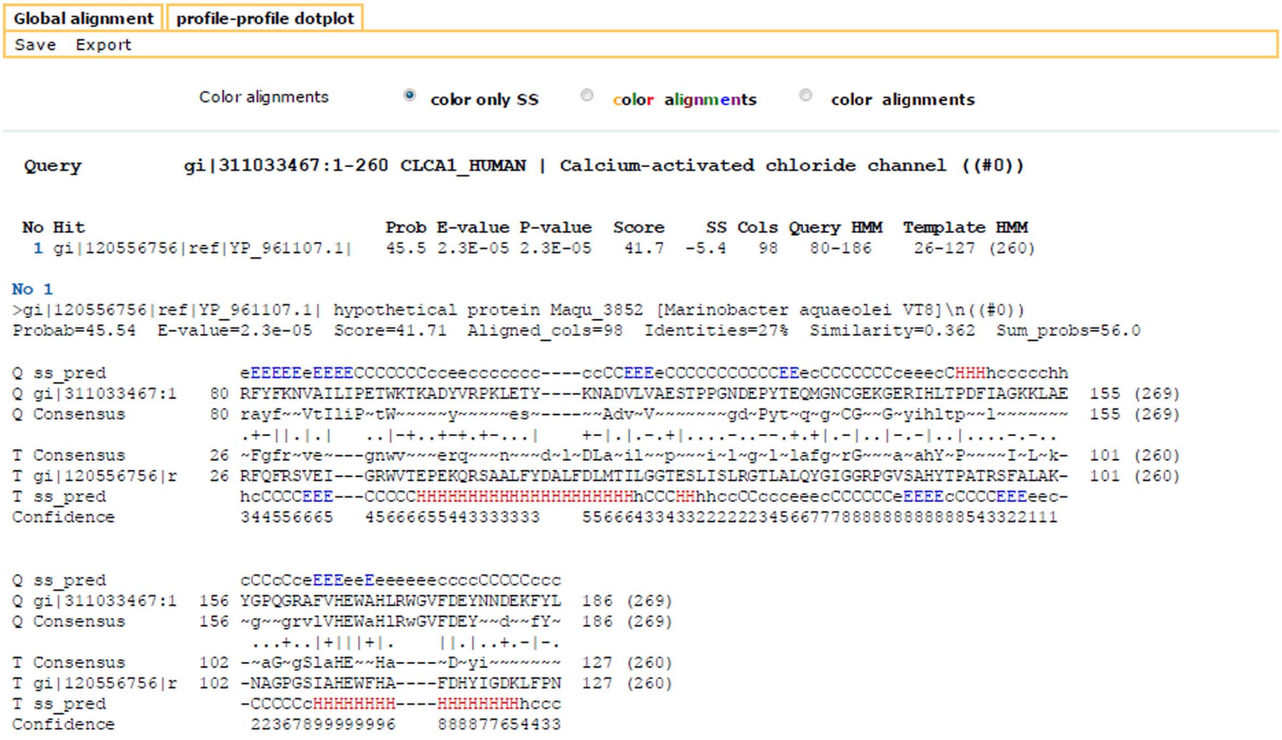

10
20
30
40
50
60
70
gi_110611231_Homo_sapiens_I25-K262
gi_5729769_Homo_sapiens_V33-A262
gi_74721541_Homo_sapiens_V24-L262
gi_205831469_Homo_sapiens_I24-L262
gi_115964593_Strongylocentrotus_purpuratus
gi_321455024_Daphnia_pulex
gi_321460553_Daphnia_pulex
gi_196007110_Trichoplax_adhaerens
gi_328707920_Acyrthosiphon_pisum
gi_291242484_Saccoglossus_kowalevskii
gi_195998007_Trichoplax_adhaerens
gi_198415894_Ciona_intestinalis
gi_260786377_Branchiostoma_floridae
gi_291224467_Saccoglossus_kowalevskii
gi_321467441_Daphnia_pulex
gi_198424353_Ciona_intestinalis
gi_241672368_Ixodes_scapularis
gi_313226492_Oikopleura_dioica
gi_241998382_Ixodes_scapularis
gi_321462270_Daphnia_pulex
gi_320352592_Desulfobulbus_propionicus_DSM_2032
gi_159898230_Herpetosiphon_aurantiacus_DSM_785
gi_260426558_Citreicella_sp._SE45
gi_149919617_Plesiocystis_pacifica_SIR1
gi_114046070_Shewanella_sp._MR7
gi_148657095_Roseiflexus_sp._RS1
gi_312879450_Aminomonas_paucivorans_DSM_12260
gi_163848422_Chloroflexus_aurantiacus_J10f
gi_240103523_Thermococcus_gammatolerans_EJ3
gi_159898060_Herpetosiphon_aurantiacus_DSM_785
jnetpred
10
20
30
40
50
60
T813
D854
70
80
90
100
gi_41179371_Bordetella_phage
gi_325971694_Sphaerochaeta_globus
gi_260549517_Acinetobacter_sp._RUH2624
gi_303257573_Burkholderiales_bacterium_1_1_47
gi_135135168_marine_metagenome
gi_303327648_Desulfovibrio_sp._3_1_syn3
gi_292670780_Selenomonas_noxia_ATCC_43541
gi_42526652_Treponema_denticola_ATCC_35405
gi_307545219_Halomonas_elongata_DSM_2581/5
gi_197935890_Ralstonia_phage_RSB
gi_330837124_Spirochaeta_coccoides_DSM_17374
gi_290968637_Megasphaera_genomosp._type_1_str._28L
gi_302339304_Spirochaeta_smaragdinae_DSM_11293/95
gi_54302257_Photobacterium_profundum_SS9/77-142
gi_334110301_fragm2_Methylomicrobium_album_BG8
gi_282848887_Veillonella_parvula_ATCC_17745
gi_260655809_fragm2_Jonquetella_anthropi_E3_33_E1
gi_223557944_uncultured_bacterium_URE12
gi_137870274_marine_metagenome
gi_144016282_marine_metagenome
gi_212703236_Desulfovibrio_piger_ATCC_29098
gi_126661473_Cyanothece_sp._CCY0110/42-90
gi_120556756_Marinobacter_aquaeolei_VT8
gi_315274348_Enterobacteria_phage_VT2phi_272
gi_288561639_Bacillus_pseudofirmus_OF4
gi_88799587_Reinekea_sp._MED297
gi_241589636_Ralstonia_pickettii_12D
gi_126668031_Marinobacter_sp._ELB17
gi_325914618_Xanthomonas_vesicatoria_ATCC_35937
gi_303258118_Burkholderiales_bacterium_1_1_47
gi_333369137_Psychrobacter_sp._1501_2011
gi_338210739_Runella_slithyformis_DSM_19594
gi_302668924_Butyrivibrio_proteoclasticus_B316
gi_193076431_Acinetobacter_baumannii_ATCC_17978
gi_283795255_Clostridium_sp._M62/1
gi_319792705_Variovorax_paradoxus_EPS
gi_325103603_Pedobacter_saltans_DSM_12145
gi_335348301_Lactobacillus_salivarius_GJ24
gi_240047480_Mycoplasma_conjunctivae_HRC
gi_336053407_Lactobacillus_kefiranofaciens_ZW3
gi_169834814_Clostridium_botulinum_A3_str._Loch_Maree
gi_168205401_Clostridium_perfringens_E_str._JGS1987
gi_343098022_Paenibacillus_polymyxa_M1
gi_229113621_Bacillus_cereus_Rock115
gi_228949804_Bacillus_thuringiensis_serovar_monterrey_BGSC_4AJ1
gi_322691234_Bifidobacterium_longum_subsp._longum_JCM_1217
gi_297618189_Syntrophothermus_lipocalidus_DSM_12680
gi_225155137_Diplosphaera_colitermitum_TAV2
gi_333996733_Treponema_primitia_ZAS-2
gi_333979014_Desulfotomaculum_kuznetsovii_DSM_6115
gi_227872114_Oribacterium_sinus_F0268
gi_136199725_marine_metagenome
jnetpred

Figure 6. Multiple sequence alignments of representative CLCA_N sequences (top) and representative CLCA_X sequences (bottom). Only regions around the predicted HexxH active site shown. Predicted secondary structures shown (jnetpred). Full versions of the alignments shown in Figs. S2 and S4, respectively.
doi:10.1371/journal.pone.0062272.g006

screened against the Pfam database (version 24.0) [67]. Fold predictions for HExxH rich domain families were performed using the FFAS server [68,69].

The CLANS algorithm [57] was run with five iterations of PSI-BLAST, using the BLOSUM45 substitution matrix and inclusion threshold of 0.001 on nr90 and env90 sequence databases. For the graphs, PSI-BLAST similarity relations with significance of P-value below 0.1 were considered, alternatively, thresholds of 0.01, 1E-5 and 1E-10 were used. The following protein families were included in the analysis: 52 families belonging to the Pfam Peptidase_MA clan, and also 12 families identified herein as similar to zincins (see Table 1): CLCA_N (PF08434), SprA-related (PF12118), FA_desaturase (PF00487), Metallopep (PF12044), MtfA (PF06167), DUF2248 (PF10005). DUF2265 (PF10023), DUF462 (PF04315), DUF922 (PF06037), DUF3267 (PF11667), DUF1025 (PF06262), DUF2342 (PF10103), and finally the CLCA_X family described herein.



Color alignments ● color only SS ○ color alignments ○ color alignments

```
 Query        gi|311033467:1-260 CLCA1_HUMAN | Calcium-activated chloride channel ((#0))


 No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
  1 gi|120556756|ref|YP_961107.1|   45.5 2.3E-05 2.3E-05   41.7  -5.4   98   80-186    26-127 (260)

No 1
>gi|120556756|ref|YP_961107.1| hypothetical protein Maqu_3852 [Marinobacter aquaeolei VT8]\n((#0))
Probab=45.54  E-value=2.3e-05  Score=41.71  Aligned_cols=98  Identities=27%  Similarity=0.362  Sum_probs=56.0

Q ss_pred              eEEEEEeEEEECCCCCCcceeccccccc----ccCCEEEeCCCCCCCCCCCEEecCCCCCCCceeecCHHHhccccchh
Q gi|311033467:1   80 RFYFKNVAILIPETWKTKADYVRPKLETY----KNADVLVAESTPPGNDEPYTEQMGNCGEKGERIHLTPDFIAGKKLAE  155 (269)
Q Consensus       80 rayf~~VtIliP~tW~~~~~y~~~~~es~----~~Adv~V~~~~~~~~gd~Pyt~q~g~CG~~G~yihltp~~l~~~~~~~  155 (269)
                       .+-||.|.|   ..|-+..+-+.+-...|    +-|.|.-.+|....-..--.+.+|.-|..|-.-|..|....-.-..
T Consensus       26 ~Fgfr~ve~---gnwv~~~erq~~~n~~~d~l~DLa~il~~p~~~i~l~g~l~lafg~rG~~~a~ahY~P~~~~I~L~k-  101 (260)
T gi|120556756|r   26 RFQFRSVEI---GRWVTEPEKQRSAALFYDALFDLMTILGGTESLISLRGTLALQYGIGGRPGVSAHYTPATRSFALAK-  101 (260)
T ss_pred              hcCCCCEEE---CCCCCHHHHHHHHHHHHHHHHHHHHHhCCCHHhhccCCccceeecCCCCCCeEEEEcCCCCEEEeec-
Confidence         344556665   456666554433333333    5566643343322222234566777888888888888543322111


Q ss_pred              cCCcCceEEEeeEeeeeeeccccCCCCCccc
Q gi|311033467:1  156 YGPQGRAFVHEWAHLRWGVFDEYNNDEKFYL  186 (269)
Q Consensus      156 ~g~~grvlVHEWaHlRwGVFDEY~~d~~fY~  186 (269)
                      ...+..|+|||+|.    ||.|..+.-|-.
T Consensus      102 -~aG~gSlaHE~~Ha----~D~yi~~~~~~~~  127 (260)
T gi|120556756|r  102 -NAGPGSIAHEWFHA----FDHYIGDKLFPN  127 (260)
T ss_pred              -CCCCCcHHHHHHHH----HHHHHHHHhccc
Confidence          22367899999996    888877654433
```

Figure 7. HHalign alignment between HHsenser-generated profiles of CLCA_N and CLCA_X protease domains.
doi:10.1371/journal.pone.0062272.g007

Survey of the CLCA_N homologues. Building representative sets of CLCA_N and CLCA_X sequences

The following sequences were used as seeds for five separate HHsenser [70] searches: residues 1–260 of human CLCA1 (gi|311033467), residues 1–312 of putative outer membrane adhesin like protein from *Shewanella* sp. MR-4 (gi|113971723), residues 575–875 of unnamed protein product from *Spirochaeta coccoides* DSM 17374 (gi|330837124), full length sequence (260 residues) of hypothetical protein Maqu_3852 from *Marinobacter aquaeolei* VT8 (gi|120556756), residues 700–1000 of Bbp10 protein from *Bordetella* phage BPP-1 (gi|41179371). Results from the first two and last three searches were combined into CLCA_N and CLCA_X sequence sets. The sets were cleared of redundant entries using the CD-HIT program [71] at 95 and 70% sequence identity levels, creating thus full and representative sets. HHsenser was ran on combined nr and env_nr (environmental sequences) databases using standard parameters. The CLCA_N full and representative sequence sets contained 160 and 92 sequences, respectively, while the CLCA_X sequence sets contained 153 and 115 sequences, respectively.

The five HHsenser run seeds were significantly similar, as judged by the FFAS profile-profile algorithm [72]: CLCA1 human vs *Shewanella*: Zscore −56.3, 12% sequence identity over 255 residues; *Marinobacter* Maqu_3852 vs *Bordetella* phage BPP-1: Zscore −33.6, 13% sequence identity over 182 residues; *Marinobacter* Maqu_3852 vs *Spirochaeta gi|330837124*: Zscore −32.6, 13% sequence identity over 181 residues. Similarities of the CLCA_N and CLCA_X families were also confirmed by HHalign [73] using HHsenser-generated multiple sequence alignments for human CLCA1 and *Marinobacter* protein Maqu_3852 as input. HHalign alignment was significant (E-value 3E-05) and covered 98 residues (see Fig. 7).

Phylogenetic tree reconstruction

A phylogenetic tree of the metazoan CLCA_N domains was built in order to establish the origin of CLCA_N domains with substituted active sites. For preparation of multiple sequence alignment, sequences of CLCA_N domains from various groups of organisms vertebrates, invertebrates and *Prokaryota* were manually selected. These sequences possessed both the correct and substituted active sites. The T-coffee [74] program with standard parameters was used to build the alignment. The obtained alignment was refinement using the G-blocks algorithm [75] to eliminate poorly aligned positions and divergent regions. The G-blocks option allow less strict flanking positions was used. The refined alignment was used to create phylogenetic tree using the ANCESCON program [76]. This algorithm provided reconstructed sequences for the tree root and all the internal nodes. The maximum likelihood method for estimaion of substitution rate factors was applied for estimation of the likelihood of residues at a site given a tree.

Other methods

For visualization of the reconstructed phylogenetic tree, the *on-line* tool iTOL [77] was applied. The sequence logos were created using the aligned Pfam seed sequences for the protein domains studied and the Weblogo tool, weblogo.berkeley.edu [78].

Identification of similarities to phage and viral proteins was performed using Blast queries on the ACLAME database [79].

The aminoacid substitution frequencies of residues in the HExxH motif were compared against the corresponding frequencies observed generally in proteins, as encoded in the PAM250 matrix [80]. Protein domains were identified using the Pfam database Pfam HMM tool [42].

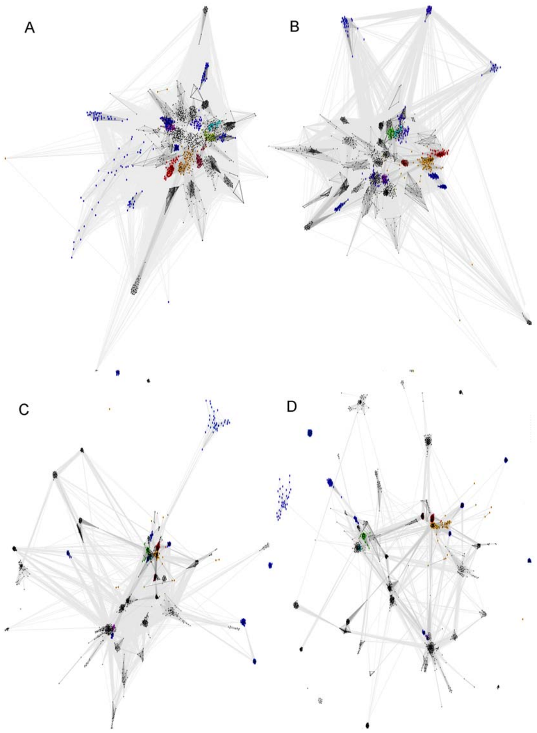


Figure 8. CLANS sequence similarity network for zincin-like proteins. Pfam clan Peptidase_MA and other zincin-like proteins included (see Methods). Four different BLAST E-value thresholds used for CLANS clustering. Pfam clan Peptidase_MA proteins: Matrixin (Peptidase_M10): green, Peptidase_M1: magenta, Reprolysin: cyan, Zn_peptidase_2: brown. Others in the Peptidase_MA clan: black Proteins not included in the Pfam clan Peptidase_MA: CLCA_N: red, CLCA_X: orange, Others outside the Peptidase_MA clan: blue. A) Relations with significance of P-value below 0.1, B) P-value below 0.01, C) P-value below 1E-5 and D) P-value below 1E-10.
doi:10.1371/journal.pone.0062272.g008

Supporting Information

Figure S1 Phylogenetic tree of human and Plocozoan CLCA_N domains.
(PDF)

Figure S2 Multiple sequence alignment of the full sequence set of CLCA_N domains (see Methods). Full version of upper part of Fig. 6.
(PDF)

Figure S3 Phylogenetic tree (PhyML, see Methods) of selected representatives of the CLCA_N domain. Upper part: tree built using nucleotide sequences. Lower part: tree built using protein sequences. Both trees were built starting from the same protein sequence alignment. Branches with bootstrap values above 50% shown in green, Human sequences highlighted in blue, prokaryotic ones highlighted in red.
(PDF)

Figure S4 Multiple sequence alignment of the full sequence set of CLCA_X domains (see Methods). Full version of lower part of Fig. 6.
(PDF)

Table S1 Replacement frequencies of the critical H and E site residues in substituted HExxH motifs in the domains of the Peptidase_MA clan as defined in the Pfam database divided by corresponding replacement frequencies in proteins in general (as derived from the PAM250 substitution matrix). First column: replacement position within the HExxH motif. Values above 2 or below 0.5 in bold.
(DOC)

Table S2 Plasmid, virus and prophage BLAST hits for CLCA_X proteins (QueryID) obtained in the ACCLAME database.
(XLS)

Author Contributions

Conceived and designed the experiments: KP. Performed the experiments: AL MD MG KP. Analyzed the data: AL MD MG KP. Contributed reagents/materials/analysis tools: AL MD MG KP. Wrote the paper: KP.

References

1. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36(Database issue): D419–425.
2. de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, et al. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 39(Database issue): D427–434.
3. Sadreyev RI, Kim BH, Grishin NV (2009) Discrete-continuous duality of protein structure space. Curr Opin Struct Biol 19(3): 321–328.
4. Alva V, Remmert M, Biegert A, Lupas AN, Soding J (2010) A galaxy of folds. Protein Sci 19(1): 124–130.
5. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. Embo J 5(4): 823–826.
6. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, et al. (2009) Exploration of uncharted regions of the protein universe. PLoS Biol 7(9): e1000205.
7. Xu Q, Gohler AK, Kosfeld A, Carlton D, Chiu HJ, et al. (2012) Structure of Mlc Titration Factor A (MtfA/YeeI) Reveals a Prototypical Zinc Metallopeptidase Related to Anthrax Lethal Factor. J Bacteriol 194(11): 2987–2999.
8. Bateman A, Coggill P, Finn RD (2010) DUFs: families in search of function. Acta Crystallogr Sect F Struct Biol Cryst Commun 66(Pt 10): 1148–1152.
9. Pawlowski K, Muszewska A, Lenart A, Szczepinska T, Godzik A, et al. (2010) A widespread peroxiredoxin-like domain present in tumor suppression- and progression-implicated proteins. BMC Genomics 11: 590.
10. Dudkiewicz M, Szczepinska T, Grynberg M, Pawlowski K (2012) A novel protein kinase-like domain in a selenoprotein, widespread in the tree of life. PLoS One 7(2): e32138.
11. Neurath H, Walsh KA (1976) Role of proteolytic enzymes in biological regulation (a review). Proc Natl Acad Sci U S A 73(11): 3825–3832.
12. Page-McCaw A, Ewald AJ, Werb Z (2007) Matrix metalloproteinases and the regulation of tissue remodelling. Nat Rev Mol Cell Biol 8(3): 221–233.
13. Turk B, Stoka V (2007) Protease signalling in cell death: caspases versus cysteine cathepsins. FEBS Lett 581(15): 2761–2767.
14. Turk B, Turk du SA, Turk V (2012) Protease signalling: the cutting edge. Embo J 31(7): 1630–1643.
15. Turk B (2006) Targeting proteases: successes, failures and future prospects. Nat Rev Drug Discov 5(9): 785–799.
16. Cudic M, Fields GB (2009) Extracellular proteases as targets for drug development. Curr Protein Pept Sci 10(4): 297–307.
17. Steuber H, Hilgenfeld R (2010) Recent advances in targeting viral proteases for the discovery of novel antivirals. Curr Top Med Chem 10(3): 323–345.
18. Hartley BS (1960) Proteolytic enzymes. Annu Rev Biochem 29: 45–72.
19. Rawlings ND, Barrett AJ (1993) Evolutionary families of peptidases. Biochem J 290 (Pt 1): 205–218.
20. Rawlings ND, Barrett AJ, Bateman A (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 40(Database issue): D343–350.
21. Gomis-Ruth FX (2009) Catalytic domain architecture of metzincin metalloproteases. J Biol Chem 284(23): 15353–15357.
22. Tallant C, Marrero A, Gomis-Ruth FX (2010) Matrix metalloproteinases: fold and function of their catalytic domains. Biochim Biophys Acta 1803(1): 20–28.

23. Nakjang S, Ndeh DA, Wipat A, Bolam DN, Hirt RP (2012) A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. PLoS One 7(1): e30287.
24. Loewen ME, Forsyth GW (2005) Structure and function of CLCA proteins. Physiol Rev 85(3): 1061–1092.
25. Winpenny JP, Marsey LL, Sexton DW (2009) The CLCA gene family: putative therapeutic target for respiratory diseases. Inflamm Allergy Drug Targets 8(2): 146–160.
26. Pauli BU, Abdel-Ghany M, Cheng HC, Gruber AD, Archibald HA, et al. (2000) Molecular characteristics and functional diversity of CLCA family members. Clin Exp Pharmacol Physiol 27(11): 901–905.
27. Eggermont J (2004) Calcium-activated chloride channels: (un)known, (un)loved? Proc Am Thorac Soc 1(1): 22–27.
28. Pawlowski K, Lepisto M, Meinander N, Sivars U, Varga M, et al. (2006) Novel conserved hydrolase domain in the CLCA family of alleged calcium-activated chloride channels. Proteins-Structure Function and Bioinformatics 63(3): 424–439.
29. Patel AC, Brett TJ, Holtzman MJ (2009) The role of CLCA proteins in inflammatory airway disease. Annu Rev Physiol 71: 425–449.
30. Bothe MK, Mundhenk L, Kaup M, Weise C, Gruber AD (2011) The murine goblet cell protein mCLCA3 is a zinc-dependent metalloprotease with autoproteolytic activity. Mol Cells 32(6): 535–541.
31. Yurtsever Z, Sala-Rabanal M, Randolph DT, Scheaffer SM, Roswit WT, et al. (2012) Self-cleavage of Human CLCA1 Protein by a Novel Internal Metalloprotease Domain Controls Calcium-activated Chloride Channel Activation. J Biol Chem 287(50): 42138–42149.
32. Mundhenk L, Johannesson B, Anagnostopoulou P, Braun J, Bothe MK, et al. (2012) MCLCA3 does not Contribute to the Calcium-Activated Chloride Conductance in Mouse Airways. Am J Respir Cell Mol Biol 47(1): 87–93.
33. Bothe MK, Mundhenk L, Beck CL, Kaup M, Gruber AD (2012) Impaired autoproteolytic cleavage of mCLCA6, a murine integral membrane protein expressed in enterocytes, leads to cleavage at the plasma membrane instead of the endoplasmic reticulum. Mol Cells 33(3): 251–257.
34. Keeling PJ (2009) Functional and ecological impacts of horizontal gene transfer in eukaryotes. Curr Opin Genet Dev 19(6): 613–619.
35. Poole AM (2009) Horizontal gene transfer and the earliest stages of the evolution of life. Res Microbiol 160(7): 473–480.
36. Anderson MT, Seifert HS (2011) Opportunity and means: horizontal gene transfer from the human host to a bacterial pathogen. MBio 2(1): e00005–00011.
37. Liu H, Fu Y, Li B, Yu X, Xie J, et al. (2011) Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes. BMC Evol Biol 11(1): 276.
38. McKerrow JH (1987) Human fibroblast collagenase contains an amino acid sequence homologous to the zinc-binding site of Serratia protease. J Biol Chem 262(13): 5943.
39. Matthews BW (1988) Structural basis of the action of thermolysin and related zinc peptidases. Accounts of Chemical Research 21(9): 333–340.
40. Bode W, Gomis-Ruth FX, Stockler W (1993) Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the 'metzincins'. FEBS Lett 331(1–2): 134–140.
41. Hooper NM (1994) Families of zinc metalloproteases. FEBS Lett 354(1): 1–6.
42. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40(Database issue): D290–301.
43. Lobocka MB, Rose DJ, Plunkett G, 3rd, Rusin M, Samojedny A, et al. (2004) Genome of bacteriophage P1. J Bacteriol 186(21): 7032–7068.
44. Lattka E, Illig T, Koletzko B, Heinrich J (2010) Genetic variants of the FADS1 FADS2 gene cluster as related to essential fatty acid metabolism. Curr Opin Lipidol 21(1): 64–69.
45. Blanchard H, Legrand P, Pedrono F (2011) Fatty Acid Desaturase 3 (Fads3) is a singular member of the Fads cluster. Biochimie 93(1): 87–90.
46. Nishimura H, Cho C, Branciforte DR, Myles DG, Primakoff P (2001) Analysis of loss of adhesive function in sperm lacking cyritestin or fertilin beta. Dev Biol 233(1): 204–213.
47. Oh JS, Han C, Cho C (2009) ADAM7 is associated with epididymosomes and integrated into sperm plasma membrane. Mol Cells 28(5): 441–446.
48. Bateman A, Rawlings ND (2003) The CHAP domain: a large family of amidases including GSP amidase and peptidoglycan hydrolases. Trends Biochem Sci 28(5): 234–237.
49. Kong L, Ranganathan S (2008) Tandem duplication, circular permutation, molecular adaptation: how Solanaceae resist pests via inhibitors. BMC Bioinformatics 9 Suppl 1: S22.
50. Podell S, Gaasterland T, Allen EE (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. BMC Bioinformatics 9: 419.
51. Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55: 709–742.
52. Goulas T, Arolas JL, Gomis-Ruth FX (2011) Structure, function and latency regulation of a bacterial enterotoxin potentially derived from a mammalian adamalysin/ADAM xenolog. Proc Natl Acad Sci U S A 108(5): 1856–1861.
53. Cerda-Costa N, Guevara T, Karim AY, Ksiazek M, Nguyen KA, et al. (2011) The structure of the catalytic domain of Tannerella forsythia karilysin reveals it is a bacterial xenologue of animal matrix metalloproteinases. Mol Microbiol 79(1): 119–132.
54. Dunning Hotopp JC, Clark ME, Oliveira DC, Foster JM, Fischer P, et al. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317(5845): 1753–1756.
55. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784): 299–304.
56. Koonin EV, Wolf YI (2012) Evolution of microbes and viruses: a paradigm shift in evolutionary biology? Front Cell Infect Microbiol 2: 119.
57. Frickey T, Lupas A (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20(18): 3702–3704.
58. Kosowska K, Reinholdt J, Rasmussen LK, Sabat A, Potempa J, et al. (2002) The Clostridium ramosum IgA proteinase represents a novel type of metalloendopeptidase. J Biol Chem 277(14): 11987–11994.
59. Knight CG, Dando PM, Barrett AJ (1995) Thimet oligopeptidase specificity: evidence of preferential cleavage near the C-terminus and product inhibition from kinetic analysis of peptide hydrolysis. Biochem J 308 (Pt 1): 145–150.
60. Brown CK, Madauss K, Lian W, Beck MR, Tolbert WD, et al. (2001) Structure of neurolysin reveals a deep channel that limits substrate access. Proc Natl Acad Sci U S A 98(6): 3127–3132.
61. Joint Center for Structural Genomics (2010) Crystal structure of a metallo-endopeptidases (BACOVA_00663) from Bacteroides ovatus at 1.93 A resolution. 3P1V. PDB Protein Data Bank. Available at http://www.rcsb.org/pdb/explore.do?structureId = 3p1v. Accessed 16 April 2013.
62. Iwashita H, Fujimoto K, Morita S, Nakanishi A, Kubo K (2012) Increased human Ca2+-activated Cl- channel 1 expression and mucus overproduction in airway epithelia of smokers and chronic obstructive pulmonary disease patients. Respir Res 13(1): 55.
63. Tanikawa C, Nakagawa H, Furukawa Y, Nakamura Y, Matsuda K (2012) CLCA2 as a p53-inducible senescence mediator. Neoplasia 14(2): 141–149.
64. Walia V, Yu Y, Cao D, Sun M, McLean JR, et al. (2012) Loss of breast epithelial marker hCLCA2 promotes epithelial-to-mesenchymal transition and indicates higher risk of metastasis. Oncogene 31(17): 2237–2246.
65. Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A 99(22): 14280–14285.
66. Gattiker A, Gasteiger E, Bairoch A (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. Appl Bioinformatics 1(2): 107–108.
67. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38(Database issue): D211–222.
68. Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9(2): 232–241.
69. Jaroszewski L, Li Z, Cai XH, Weber C, Godzik A (2011) FFAS server: novel features and applications. Nucleic Acids Res 39 Suppl 2: W38–44.
70. Soding J, Remmert M, Biegert A, Lupas AN (2006) HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res 34(Web Server issue): W374–378.
71. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13): 1658–1659.
72. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33(Web Server issue): W284–288.
73. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7): 951–960.
74. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1): 205–217.
75. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17(4): 540–552.
76. Cai W, Pei J, Grishin NV (2004) Reconstruction of ancestral protein sequences and its applications. BMC Evol Biol 4: 33.
77. Letunic I, Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 39(Web Server issue): W475–478.
78. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(6): 1188–1190.
79. Leplae R, Lima-Mendez G, Toussaint A (2010) ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res 38(Database issue): D57–61.
80. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of Protein Sequence and Structure. Washington: National Biomedical Research Foundation. pp. 345–352.