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

Anna Lenart¹, Małgorzata Dudkiewicz², Marcin Grynberg³, Krzysztof Pawlowski¹²*¹

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, zinccins, 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 zinccin-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 zinccins are not known. The CLCAs are a novel zinccin-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 plages and plasmids, supporting the HGT scenario.

Citation: Lenart A, Dudkiewicz M, Grynberg M, Pawlowski 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

Copyright: © 2013 Lenart et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: AL, MD, MG and KP were supported by the grant N N301 3165 33 from the Ministry 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 remodeling, 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 zinccin-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 zinccin-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...
The Widespread CLCA Metalloprotease Family

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-estimated, p-value of the binomial text < 2 \times 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 56 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, Ghu 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 Ghu or Arg. The second histidine...
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 murostatus], atrolysin E   |                                            |
| SprA-related (PF12118) SprA-related family                           | 95                                                     | −7.7         | 11                  | PF05569; Q8RPJ4_DESHA/7—280; BlaR1 peptidase M56                                            | this article                              |
| FA_desaturase (PF00487) Fatty acid desaturase                         | 37                                                     | −8.4         | 16                  | d1k7a2 d.92.1.6 (A:18–258) Metalloprotease [Enwina chrysanthemi]                           | this article                              |
| Metallogep (PF12044) Putative peptidase family                        | 87                                                     | −12.3        | 20                  | d1c7ka_ . d.92.1.1 (A:) Zinc protease [Streptomyces caesipitosus]                          |                                            |
| Mtfa (PF06167) Phosphoenolpyruvate: glucos-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 aminopeptidase                           | 100                                                    | −7.7         | 11                  | d3b7s3 a. d.92.1.13 (A:209–460) Leukotriene A4 hydrolyase 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                  | d1kpa 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 Acel2062 [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 aurantiaci]    | this article                              |

doi:10.1371/journal.pone.0062272.t001
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 metallopeptidase 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 metallopeptases 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, mimicking 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 metallopeptidase 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 archael 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 Chlorella, Synechocystis, 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 Synechocystis. Other prokaryotic CLCA proteins (archael and those from Chlorella) have the CLCA_N domain thrown into a different, species-specific domain architecture (see Fig. 5). Interestingly, the typical CLCA_N domain architecture is also present in two species, the gamma-proteobacterium Tendersibacter and the euryarcheon Methanosaurox, with the CLCA_N domain replaced by two unrelated enzymatic (peptidase) domains, Pepptidase_M10 and CHAP, respectively [22, 40].

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
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...
The CLCA_X (CLCA-like) family

Classic Darwinian mechanisms of evolution [51,55,56]. HGT is now established as one of key mechanisms modulating the bacteria, archaea and viruses has been known for a decade and between eukaryotes and prokaryotes has only recently been bacterial pathogen [52,53]. Although the prevalence of HGT structural features and sequence similarity relationships strongly
dimensional structures turned out to be metalloproteases. Their functions.
CLCA extracellular protease functions, including the regulatory 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 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 xenologes in bacterial genomes. Recently, three-dimensional structures of two virulence factors from Bacteroides 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 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-proteobacteria, as well as alpha-, beta-, gamma- and delta-proteobacteria, as well as Synergistetes, Spirochaetes and Firmicutes. Also, Caudovirales viruses possess CLCA_X proteins. However, CLCA_N and CLCA_X domains 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-NirB (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 xenologes in bacterial genomes. Recently, three-dimensional structures of two virulence factors from Bacteroides 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].

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 CLCA 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 was PF04298 Zn_peptidase_2, an uncharacterized bacterial family, present mostly in Fungi and Bacteroides. Further, relatively closely to CLCA 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 structures turned out to be metalloproteases. Their structural features and sequence similarity relationships strongly
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
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 undiscovered. 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
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.

Figure 6. Multiple sequence alignments of representative CLCA_N sequences (top) and representative CLCA_X sequences (bottom). Only regions around the predicted HexxxH 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

The Widespread CLCA Metalloprotease Family

Peptidase_MA clan, and also 12 families identified herein as similar to zincins (see Table 1): CLCA_N (PF08434), SprA-related (PF12118), FA_desaturase (PF04087), 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.

PLOS ONE | www.plosone.org 9 May 2013 | Volume 8 | Issue 5 | e62272

doi:10.1371/journal.pone.0062272.g006
The Widespread CLCA Metalloprotease Family

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 estimation 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 online 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, Reproylin: 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.

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

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.

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

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.

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

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, Gehler AK, Kudrjal A, Carlyon 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. FEBS Lett 581(15): 2761–2767.
9. Pawlowski K, Muszewska A, Lenart A, Szczepinska T, Godzik A, et al. (2010) A novel protein kinase-like domain in a selenoprotein, widespread in the tree of life. BMC Genomics 11: 590.
10. Tallant C, Marrero A, Gomis-Ruth FX (2010) Matrix metalloproteinases: fold and function of their catalytic domains. Biochim Biophys Acta 1803(1): 20–28.
The Widespread CLCA Metalloprotease Family

23. Nakajig S, Nishio DA, Wipat A, Bolam DN, Hirr RP (2012) A novel extracellular metalloproteinase 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. Phys Rev 85(3): 1093–1092.

25. Winpenny JP, Marsey LL, Sexton DW (2009) The CLCA gene family: putative therapeutic target for respiratory diseases. Inflamm Allergy Drug Targets 8(2): 146–154.

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 (unknown, unloved) Proc Am Thorac Soc 1(1): 22–27.

28. Pawlowski K, Lepisto M, Meinander N, Sirvas 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): 424–430.

29. Patel AC, Brett TJ, Holtzman MJ (2009) The role of CLCA proteins in inflammatory airway disease. Annu Rev Physiol 71: 423–449.

30. Botte MK, Mundhenk L, Kaup M, Weise C, Gruber AD (2011) The marine goblet cell protein MCLCA3 is a zinc-dependent metalloprotease with autoprotolytic activity. Mol Cells 32(6): 535–541.

31. Yurtsizer Z, Sala-Rabanal M, Randolph DT, Scheaffer SM, Rosswt WT, et al. (2012) Self-cleavage of human CLCA1 Protein by a Novel Internal Metalloprotease Domain Controls Calcium-activated Chloride Conductance in Mouse Airways. Am J Respir Cell Mol Biol 47(1): 87–93.

32. Botte MK, Mundhenk L, Beck CL, Kaup M, Gruber AD (2012) Impaired autoprotolytic 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.

33. Keeling PJ (2009) Functional and ecological impacts of horizontal gene transfer in eukaryotes. Curr Opin Genet Dev 19(6): 613–619.

34. Poole AM (2009) Horizontal gene transfer and the earliest stages of the evolution of life. Res Microbiol 160(7): 473–480.

35. Anderson MT, Seifert HS (2011) Opportunity and means: horizontal gene transfer in archaeal and bacterial genomes, identified using the DarkHorse Bioinformatics 9 Suppl 1: S22.

36. 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.

37. 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.

38. Bateman A, Rawlings ND (2003) The CHAP domain: a large family of amidases and metallopeptidases (BACOVA_00663) from Bacteroides ovatus at 1.93 A resolution. J Biol Chem 278:5437–5440.

39. 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(20): 12800–12805.

40. Brown CK, Madaus K, Lián 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(10): 5317–5327.

41. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) Clustering of protein families database. Nucleic Acids Res 33(Database issue): D290–301.

42. Jaroszewski L, Cai XH, Weber C, Godzik A (2011) FFAS server: novel transitive profile search using HMM-HMM comparison. Nucleic Acids Res 39 Suppl 2: W384–288.

43. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database in 2011. Nucleic Acids Res 38(Database issue): D211–222.

44. Gough TL, Arolas JL, Gomis-Ruth FX (2011) Structure, function and latency of Wolbachia endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A 98(10): 5317–5327.

45. Toledano MM, Diano S, Gregoire BM, Ziegler S, Groot CH, et al. (2006) CLCA2 as a p53-inducible senescence mediator. Neoplasia 14(2): 141–149.

46. Cai W, Pei J, Grishin NV (2004) Reconstruction of ancestral protein sequences using transitive profile search using HMM-HMM comparison. Nucleic Acids Res 32(Web Server issue): W374–378.

47. Castresana J (2000) Selection of conserved blocks from multiple alignments for phylogenetic analysis. Mol Biol Evol 17(4): 540–552.

48. Letunic I, Bork P (2011) Interactive Tree Of Life v2: online annotation and logo generator. Genome Res 14(6): 1188–1190.

49. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2000) Clustering of protein families database. Nucleic Acids Res 28(Database issue): D211–222.

50. Calvo KL, Shabtai O, Grunebaum I, Netherlands ADAM 2007: a comprehensive database that includes free amino acid and protein sequences. Bioinformatics 23(22): 3057–3058.

51. Castresana J (2001) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17(4): 540–552.

52. 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.

53. Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T (2002) Evidence of preferential cleavage near the C-terminus and product inhibition from kinetic analysis of peptide hydrolysis. Biochem J 368(Pt 1): 143–150.

54. 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.

55. Leitie 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.

56. Coatney G, Hon G, Chandraon DJ, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(10): 1188–1190.

57. Geer LY, Kapp E, Heger A, et al. (2005) Infragene: an integrated database of eukaryotic genes. Database 2005(2): 202–206.