Involvement of β-carbonic anhydrase (β-CA) genes in bacterial genomic islands and horizontal transfer to protists

Reza Zolfaghari Emameh,[a#] Harlan R. Barker,[b] Vesa P. Hytönen,[b,c] Seppo Parkkila,[b,c]

[a]Department of Energy and Environmental Biotechnology, Division of Industrial & Environmental Biotechnology, National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran

[b]Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland

[c]Fimlab Laboratories Ltd and Tampere University Hospital, FI-33520 Tampere, Finland

Running Head: β-carbonic anhydrase genes in genomic islands

#Address correspondence to Reza Zolfaghari Emameh, zolfaghari@nigeb.ac.ir

E-mail addresses:

Reza Zolfaghari Emameh: zolfaghari@nigeb.ac.ir

Harlan R. Barker: harlan.barker@uta.fi

Vesa P Hytönen: vesa.hytonen@uta.fi

Seppo Parkkila: seppo.parkkila@staff.uta.fi

ABSTRACT

Genomic islands (GIs) are a type of mobile genetic element (MGE) that are present in bacterial chromosomes. They consist of a cluster of genes which produce proteins that contribute to a variety of functions, including, but not limited to, regulation of cell metabolism, anti-microbial resistance, pathogenicity, virulence, and resistance to heavy metals. The genes carried in MGEs can be used as a trait reservoir in times of adversity. Transfer of genes using MGEs, occurring outside of reproduction, is called horizontal gene transfer (HGT). Previous literature has shown that numerous HGT events have occurred through endosymbiosis between prokaryotes and eukaryotes.

Beta carbonic anhydrase (β-CA) enzymes play a critical role in the biochemical pathways of many prokaryotes and eukaryotes. We have previously suggested horizontal transfer of *β-CA* genes from plasmids of some prokaryotic endosymbionts to their protozoan hosts. In this study, we set out to identify β-CA genes that might have transferred between prokaryotic and protist species through HGT in GIs. Therefore, we investigated prokaryotic chromosomes containing β-CA-encoding GIs and utilized multiple bioinformatics tools to reveal the distinct movements of *β-CA* genes among a wide variety of organisms. Our results identify the presence of *β-CA* genes in GIs of several medically and industrially relevant bacterial species, and phylogenetic analyses reveal multiple cases of likely horizontal transfer of *β-CA* genes from GIs of ancestral prokaryotes to protists.

IMPORTANCE

The evolutionary process is mediated by mobile genetic elements (MGEs), such as genomic islands (GIs). A gene or set of genes in the GIs are exchanged between and within various species through horizontal gene transfer (HGT). Based on the crucial role that GIs can play in bacterial survival and proliferation, they were introduced as the environmental- and pathogen-associated factors. Carbonic anhydrases (CAs) are involved in many critical biochemical pathways, such as regulation of pH homeostasis and electrolyte transfer. Among the six evolutionary families of CAs, *β-CA* gene sequences are present in many bacterial species, which can be horizontally transferred to protists during evolution. This study shows for the first time the

involvement of bacterial *β-CA* gene sequences in the GIs, and suggests their horizontal transfer to protists during evolution.

KEYWORDS

β-carbonic anhydrase; Evolution; Genomic island; Mobile genetic element

Horizontal Gene Transfer (HGT) is an evolutionary phenomenon by which a gene, or set of genes, are exchanged between and within various species. This makes HGT unique compared with other evolutionary processes, such as gene duplication, mutation, and sexual reproduction. While a heritable HGT in eukaryotes entails entrance of a foreign gene to the nucleus of the germ cell and successful insertion to chromatin packed DNA, HGT has multiple pathways in prokaryotic species. This evolutionary process is mediated by mobile DNA or mobile genetic elements (MGEs), which can include: genomic islands (GIs), plasmids, transposons, retrotransposons, and prophages (1-6). During HGT, selfish "parasitic" elements are often associated with toxin resistance genes, metabolic genes, virulence factors, and a wide range of secreted factors. The acquisition of a useful gene repertoire could offset the cost of maintaining and transferring a large selfish element, such as a conjugal plasmid (7). Transformation, conjugation, and transduction are each distinct methods of HGT in prokaryotes.

Varieties of important genes are transferred between prokaryotes, or from prokaryotes to eukaryotes, through HGT (8), including those for virulence factors, antibiotic resistance, and toxins (1-3). In 1990, some clusters of virulence genes, which transfer through HGT, were identified in *Escherichia coli* and described as pathogenicity islands (PAIs) (9). Later GIs were defined as any cluster of genes (10–200 kb) that has been acquired by HGT (10). GIs represent a part of a cell's chromosome, recognized as discrete DNA segments, and can differ between closely related strains. Different GI families have been recognized on the basis of sequence and functional homologies by GI prediction tools (11). The nucleotide sequence length of GIs is >10 Kb, while it is <10 Kb for smaller genomic islets (12).

Interest in GIs has increased commensurately with developing knowledge of their role in bacterial survival and proliferation. The common environmental- and pathogen-associated virulence factors found disproportionately in GIs tend to serve functions. For example, the pathogenicity role of β-CA has been approved in *Pseudomonas aeruginosa* (13) and the critical role of β-CA in detoxification of cyanate by providing bicarbonate for cyanase enzyme has been shown in *Pseudomonas pseudoalcaligenes* (14).

However, clustered regularly interspaced short palindromic repeats (CRISPRs), used by bacteria in defense against insertion of phage DNA, are also found overrepresented in GIs (15).

The presence of GIs have been studied using various computational biology methods. There are two main methods for prediction of GIs, including: (1) evaluation of sequence compositions, using tools such as SIGI-HMM (16), IslandPath-DIMOB (17), PAI-IDA (18), and Centroid (19), and (2) application of comparative genomics, such as BLAST homology search and whole-genome sequence alignment. Among the sequence analysis methods, SIGI-HMM and IslandPath-DIMOB have shown the highest overall accuracy (15). Two computational methods for prediction of GIs based on comparative genomics, include IslandPick (20) and MobilomeFINDER (21). The latter method focuses on identification of the islands associated with *tRNA* genes. However, not all GIs use *tRNA* genes as insertion sites, which thus limits the usage of MobilomeFINDER compared to the IslandPick method (22). Prediction based on IslandPick is provided at the IslandViewer 4 database [http://www.pathogenomics.sfu.ca/islandviewer/] (23). The IslandViewer server combines the three most accurate GI prediction methods into a single analysis: IslandPath-DIMOB, SIGI-HMM, and IslandPick.

Carbonic anhydrases (CAs) are ubiquitous metalloenzymes, which are categorized into seven gene families, including α, β, γ, δ, ζ, η, and θ (24-27). CAs are involved in many important biochemical pathways including pH homeostasis, electrolyte transfer, transport of CO_2 and bicarbonate between metabolizing tissues, and some biosynthetic processes (28-31). Many ancient putative β-CAs have been discovered in protozoans, rotifers, sea louses, molluscs, starlet sea anemones, purple sea urchins, arthropods, nematodes, and trematodes (32-34), as well as in prokaryotes and some eukaryotes, such as fungi, algae, and plants (35). Notably, *β-CA* gene sequences are present in the genomes of most living organisms except vertebrates (33, 34). β-CAs are considered to be crucial metabolic enzymes (32, 36, 37). They act as virulence factors for various bacterial, fungal and parasitic species, such as *Pseudomonas aeruginosa* (13) , *Cryptococcus neoformans* (38), and *Toxoplasma gondii* (39), so β-CAs develop a cascade leading to the production of infectious spores in *C. neoformans*, prepare the adaptation of *P. aeruginosa* to low CO_2 condition through different organization of three *β-CA* genes, and play the role in rhoptry biogenesis and formation of parasitophorous vacuole in *T. gondii*. β-CA is a vital enzyme for fertility of female insects (*Drosophila melanogaster*) (36) and therefore, β-CAs are attractive targets for inhibition studies in insects and pests. There is active ongoing research in this field focusing on inhibition of β-CAs in important organisms; for example, application of sulfonamide and sulfamate for inhibition of β-CA from *Helicobacter pylori* (40), aromatic carboxylates for inhibition of β-CA from *Candida albicans* (41), sulfonamides for inhibition of β-CA from *Ascaris lumbricoides* (42), and sulfonamides for inhibition of β-CA from malaria mosquito *Anopheles gambiae* (43). CA inhibition studies have been mainly performed *in vitro*, and only a few *in vivo* studies have been carried out on parasitic infectious diseases (44-46).

Here we have studied the importance of HGT and *β-CA* gene exchange between bacterial GIs and protists genomes. We propose that GIs play a crucial role in horizontal transfer of *β-CA* genes from prokaryotes to protists.

RESULTS

***β-CA* genes are located in many genomic islands**

Our comparative analysis of the GI annotations presented in the IslandViewer 4 database and NCBI genome annotations, allowed us to identify a total of 272 instances of *β-CA* genes in bacterial GIs (Table S1). In study of all strains, nucleotides in *β-CA* genes are 3.81x more likely to occur in GIs than is expected by chance.

Identification of β-CAs from prokaryotes and protists

A multiple sequence alignment (MSA) was created for 86 amino acid residues of 25 prokaryote and protist β-CA protein sequences. The alignment revealed that all β-CA protein sequences contain the first (CXDXR; C: Cysteine, D: Aspartic acid, R: Arginine, and X: any amino acid) and second (HXXC; H: Histidine, C. Cysteine, and X: any amino acid) highly conserved motifs, which are characteristic of a β-CA protein (Fig. 1).

Using the data from the IslandViewer 4 IslandPath-DIMOB webserver and corresponding NCBI genome annotations, 272 *β-CA* genes were identified inside of prokaryote GIs (Table S1) (e.g. β-CA-encoding GI from *Methylibium petroleiphilum* (strain PM1)).

Table S1. Prokaryotic GIs containing *β-CA* genes.

Phylogenetic analysis

The result of the phylogenetic analysis is presented as a circular tree and divisions of interest are shown in three different clades: A, B, and C (Fig. 2). Partitioning delineates regions where *β-CA* genes appear to have a common ancestor in prokaryotes and protists.

In clade A, the *β-CA* gene of *Trichomonas vaginalis* and *Paulinella chromatophora* has a common ancestor with *β-CA* genes from prokaryotic GIs. In Clade B, there is a *β-CA* gene from a prokaryotic GI and two *β-CA* genes from the protist *Acanthamoeba castellanii.* Clade C includes *β-CA* genes from 27 bacterial GIs and 14 protists including *A. castellanii, Capsaspora owczarzaki, Dictyostelium discoideum, D. fasciculatum, D. purpureum, Leishmania donovani, L. panamensis, Leptomonas pyrrhocoris, Phaeodactylum tricornutum, Phytophthora infestans, Polysphondylium pallidum, Saprolegnia diclina, Tetrahymena thermophila,* and *Trypanosoma grayi*. We did not identify any definite bacterial common ancestor with *β-CA* genes from protists *Entamoeba invadens, E. nuttalli,* and *Galdieria sulphuraria*.

Sequence conservation analysis for HGT

In order to evaluate the hypothesis of HGT between prokaryotes and eukaryotes within Clade C, the sequence conservation among the Clade C proteins were compared to the rest of the phylogenetic tree. First, Clade C protein sequences were aligned using Clustal Omega and the residues fully conserved within Clade C were identified (14 residues). Then, all proteins within the phylogenetic tree outside Clade C, except protist sequences, were aligned using Clustal Omega (Fig. S1). The resulting MSA was analyzed using program Consurf for sequence conservation (Fig. 3). We then inspected the conservation of those 14 fully-conserved residues within Clade C for their conservation in the large group of CA sequences. Among those, 7 residues were highly conserved (conservation score 9) and 4 were well conserved (conservation score 7-

8). However, 3 of the residues showed an average of low conservation (conservation score 3-5). Therefore, conservation of those 3 residues could be considered as a possible result of HGT. These three conserved residues (Leu21, Gly71 and Gly117) of the homology modeled of β-CA from *T. vaginalis* (A2ENQ8) were shown in Fig. 4.

FIG S1 Multiple sequence alignment (MSA) for all proteins within the phylogenetic tree outside Clade C (Fig. 2), except protist sequences, were aligned using Clustal Omega.

Exon count for *β-CA* genes from protists

Single exon structure of a protist gene would provide some additional support for the hypothesis that this particular gene could be of prokaryotic origin. The exon count analysis revealed that the *β-CA* genes of certain protists have indeed a single exon, while many other *β-CA* genes have multiple exons. The single-exonic *β-CA* genes of protists have shown in Table 1.

DISCUSSION

Our previous phylogenetic analysis has suggested that *β-CA* genes have crossed species boundaries on multiple occasions (8). The present identification of *β-CA* genes within bacterial GIs also strongly suggests that prokaryotic *β-CA* genes have been horizontally transferred between and within different species (47). Specifically, we see what appears to be a very clear case of HGT of a β-CA from prokaryotic GIs to protists *T. vaginalis* and *P. chromatophora* (clade A), *A. castellanii* (clade B), and *Dictyostelium* sp., *P. pallidum, T. thermophila* (NCBI IDs: XP_001009612.1, XP_001013978.2, XP_001022390.2), *Leishmania* sp., *L. pyrrhocoris*, and *T. grayi* (clade C). Also, our phylogenetic analysis reveals that *β-CA* genes from GIs of *Bacillus thuringiensis* and *Psychrosinus fermentans* have common ancestors with *T. vaginalis* (clade A) and *A. castellanii* (clade B), respectively. In addition, a *β-CA* gene from a GI of *Rahnella aquatilis* shows a common ancestor with *Dictyostelium* sp. and *P. pallidum*. In several cases we observe clustering of multiple *β-CA* genes from the same protist species. When these cluster immediately together we believe the most likely explanation is gene duplication after HGT. In the case where we observe two distinct clusters of paralogs from the same protist, such as with *T. thermophila* in Clade A, it is possible that there have been multiple duplication events after HGT, or separate cases of HGT.

A previous study has shown that *β-CA* genes in protists exist as single or multiple exon chromosomal genes, while in metazoans these genes exist only as multiple exon chromosomal genes (48). Single exon *β-CA* genes can be found in most of our candidate HGT species, *Entamoeba* sp., *Leishmania* sp., *L. pyrrhocoris*, *P. chromatophora, Phytophthora infestans* (XP_002909250.1), *T. thermophila* (XP_001009111.1), *T. vaginalis* (NCBI ID: XP_001317907.1), and *T. grayi*; while multiple exon *β-CA* genes can be found in the other protists. The single exon structure of some *β-CA* genes of protists thus suggests that they are closely associated to prokaryotic β-CAs. Therefore, it seems that *β-CA* genes with prokaryotic GIs origins have integrated into stable chromosomal loci in genomes of protists without association between GIs and *β-CA* genes of protists. Due to the large and complexity of eukaryotic genomes and heterogeneous chromosomes leading to high rates of false-positive results, the horizontal transfer of GIs to the eukaryotic genomes is in a halo of ambiguity and largely unexplored (49). Currently, identification studies of horizontally transferred genes to eukaryotes are performed through comparative analyses than experimental methods to show GIs in the genome of the eukaryotes. Therefore, further studies are needed to design the databases for identification of eukaryotic GIs.

Our studies have revealed that β-CA protein has, in some cases, potentially evolved into a virulence factor for some pathogenic bacteria, such as *B. pseudomallei* (50). We have identified a significant number of *β-CA* genes which reside in GIs in bacterial species, many of which are known to be pathogenic. These *β-CA* genes occur inside of GIs at a significantly higher rate than expected by chance, implying some function. Known virulence factors can be influenced by attenuators or RNA-based regulatory strategies, which lead to premature termination of transcription (51). Based on the lack of β-CAs in vertebrates, these proteins can be considered potential targets for anti-parasitic drugs. On the other hand, due to the presence of β-CA in *M. petroleiphilum*, it is suggested that this enzyme plays a major role in a CO_2-concentrating-mechanism (CCM) in carboxysomes through use of methyl tert-butyl ether (MTBE) as the sole source of carbon. The β-

CAs from other extremophilic bacteria (Table S1), such as *Halothermothrix orenii* (52), *Acidithiobacillus caldus* (53), *Thioalkalivibrio nitratireducens* (54), and *Thiomicrospira crunogena* (55) may play critical metabolic roles through carboxysome or non-carboxysome-associated mechanisms.

The GIs in some bacterial species, such as *Pseudomonas aeruginosa*, can lead to emergence of strains resistant to various antibiotics (56). It was demonstrated that *P. aeruginosa* (isolate ST235) contains Tn6162 and Tn6163 in GI1 and GI2, respectively, which function together as multiple antibiotic-resistant cassettes. An environmental study showed that *Thiomonas* sp. is able to withstand the extreme conditions of acid mine drainage (57). The comparison between the genomes of *T. arsenitoxydans* (strain 3As), *T. intermedia* (strain K12), and *Thiomonas* sp. (strain CB2) identified over 20 GIs occurring through various rearrangements containing arsenite resistance and oxidation genes, leading to divergent resistance to arsenic-rich environments.

Conclusions. A GI is a continuous genomic region which arises through HGT and can contain tens to hundreds of genes. We have previously identified cases of horizontal transfer of *β-CA* genes from plasmids of some prokaryotic endosymbionts to their protozoan hosts (48). The present results support the idea that *β-CA* genes in protists and modern eukaryotes originated by HGT from ancestral prokaryotic GIs, along with other facilitators, such as transposase and integrase. Using phylogenetics and homology modeling, we suggest that the close sequence similarity of *CA* genes in hosts and endosymbionts was due to HGT and not convergent evolution (48).

Further studies will be needed to identify the origin of *β-CA* genes in ancestral and metazoan species. Even though our results suggest that *β-CA* genes are overrepresented in GIs compared to the rest of the genome, no studies have yet been reported on whether *β-CA* genes are overrepresented there compared to other metabolic genes.

MATERIALS AND METHODS

Identification of CA proteins in genomic islands

A total of 110,913 genomic island annotations for 6,348 complete bacterial and archaeal strains were retrieved from the IslandViewer 4 database. RefSeq assembly annotations were available for 6,238 strains, which were downloaded in bulk from the NCBI assembly server (https://www.ncbi.nlm.nih.gov/assembly); annotations only available GenBank and updated annotation versions not corresponding to the current IslandViewer release were not retrieved. Using custom Python scripts, assembly IDs were retrieved for all IslandViewer genome accession IDs, and gene annotations compared with the GI locations. All *CA* genes which occur within the IslandViewer defined GI locations were kept for further analysis. To determine average overrepresentation, for each examined genome the number of *CA* gene nucleotides (nt) overlapping any GI was compared to the number expected by chance alone. Expected overlap was defined as the sum of lengths of all CAs in the genome multiplied by the sum of lengths of all GIs in the genome divided by the length of the genome.

Identification of β-CAs from prokaryotes and protists

After detection of GIs-containing *β-CA* genes in the IslandViewer version 4 database, we then collected prokaryotic β-CAs locating in GIs, and β-CA from *Klebsiella pneumoniae* subsp. pneumoniae (NCBI protein ID: WP_019705531.1), five β-CA protein sequences equally from both gram-negative and gram-positive bacteria (Table 2), and protist β-CAs (Table 3) to perform a multiple sequence alignment (MSA) analysis. We used the β-CA protein sequence from *Klebsiella pneumoniae* subsp. pneumoniae (NCBI protein ID: WP_019705531.1) as a query from prokaryotic species for the MSA analysis. All β-CA protein sequences from protists used in the analysis are described in Table 3. Also, we used the β-CA protein sequence from *A. castellanii* (XP_004344666.1) as the query from protists in the Basic Local Alignment Search Tool for proteins (blastp) from NCBI database (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) through running different phyla of protists including Stramenopiles, Alveolata, Rhizaria, Excavata, Amoebozoa, Hacrobia, Apusozoa, and Opisthokonta in the choosing search set panel. Some protists contain more than one β-CA protein sequence, in which case we used only one as a representative sequence in the

MSA analysis. In total 25 β-CA protein sequences of prokaryotes and protists (86 amino acid residues, starting three amino acid residues prior to first highly conserved motif; CXDXR) were used to compute an MSA using the Clustal Omega. The results were visualized in JalView [http://www.jalview.org/] (58).

Identification of β-CA gene sequences located in prokaryotic GIs was performed using the IslandViewer version 4 database. This webtool provides the ability to draw main circular chromosomes of defined prokaryotes containing GIs, as well as search for *β-CA* gene sequences.

Phylogenetic analysis

The β-CAs identified to reside in bacterial GIs, using annotations from the IslandViewer database, were clustered to 90% similarity centroids with the "cluster_fast" algorithm of the search tool (59) in order to reduce the number of sequences for phylogenetic analysis. Similarly, a set of 35 protist β-CAs were clustered to 90% similarity. The resulting reduced set of 122 prokaryote β-CAs found within GIs and 35 protist β-CAs were aligned using Clustal Omega. Model testing was performed to identify the best evolutionary model for analysis of the target sequences using ModelFinder (60). A maximum likelihood phylogenetic analysis was performed using the IQTree software (61, 62), with parameters set to "-alrt 100000 -bb 100000 -nt AUTO -m LG+R7" and all other options run as default. A consensus tree was generated from the 100,000 bootstrap replicates, with a final log-likelihood value of -37626.11. The tree was then visualized using the ETE Toolkit Python library (63).

Sequence conservation analysis for HGT

In order to analyze sequence conservation among the β-CA proteins, their sequences were aligned using the Clustal Omega and the resulting MSA was then analyzed using the ConSurf Server [http://consurftest.tau.ac.il/] (64).

Exon count for *β-CA* genes from protists

In order to count the exons of *β-CA* genes from protists, we used the NCBI gene server (https://www.ncbi.nlm.nih.gov/gene/) (65). In this feature, a summary of a specific gene including gene type, symbol and description, locus tag, RNA name, RefSeq status, organism lineage, and genomic context (exon count) are presented.

ACKNOWLEDGMENTS

To perform this study RZE received a scholarship support from the Ministry of Science, Research and Technology, and National Institute of Genetic Engineering and Biotechnology of Islamic Republic of Iran. This work was supported by research grants from the Academy of Finland to SP, Finnish Cultural Foundation (Pirkanmaa Regional Fund to RZE and Maili Autio Fund to HRB), Sigrid Juselius Foundation to SP, Tampere Tuberculosis Foundation to SP, and Competitive Research Funding of the Tampere University Hospital to SP. The authors declare no conflict of interest. The funding organizations had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Authors' contributions: All authors participated in the design of the study. RZE carried out the search and collection of relevant prokaryotic and eukaryotic species, identification of β-CAs, and conservation analysis. RZE created the MSAs. HRB identified the *β-CA* genes in GIs, made protein sequence corrections and predictions, and performed the phylogenetic analysis. VPH performed sequence conservation analysis and protein modeling. RZE drafted the first version of the manuscript. All authors participated in writing, read and approved the final manuscript.

REFERENCES

1. Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709-42.
2. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711-21.
3. Aminov RI. 2011. Horizontal gene exchange in environmental microbiota. Front Microbiol 2:158.
4. Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T. 2016. Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. Genome Biol Evol 8:1785-801.
5. Kidwell MG. 1993. Lateral transfer in natural populations of eukaryotes. Annu Rev Genet 27:235-56.
6. Springael D, Top EM. 2004. Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. Trends Microbiol 12:53-8.
7. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472-82.

8. Zolfaghari Emameh R, Barker HR, Tolvanen ME, Parkkila S, Hytonen VP. 2016. Horizontal transfer of beta-carbonic anhydrase genes from prokaryotes to protozoans, insects, and nematodes. Parasit Vectors 9:152.
9. Hacker J, Bender L, Ott M, Wingender J, Lund B, Marre R, Goebel W. 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. Microb Pathog 8:213-25.
10. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23:1089-97.
11. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 33:376-93.
12. Hacker J, Carniel E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep 2:376-81.
13. Lotlikar SR, Hnatusko S, Dickenson NE, Choudhari SP, Picking WL, Patrauchan MA. 2013. Three functional beta-carbonic anhydrases in Pseudomonas aeruginosa PAO1: role in survival in ambient air. Microbiology 159:1748-59.
14. Luque-Almagro VM, Huertas MJ, Saez LP, Luque-Romero MM, Moreno-Vivian C, Castillo F, Roldan MD, Blasco R. 2008. Characterization of the Pseudomonas pseudoalcaligenes CECT5344 Cyanase, an enzyme that is not essential for cyanide assimilation. Appl Environ Microbiol 74:6280-8.
15. Ho Sui SJ, Fedynak A, Hsiao WW, Langille MG, Brinkman FS. 2009. The association of virulence factors with genomic islands. PLoS One 4:e8094.
16. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics 7:142.
17. Hsiao W, Wan I, Jones SJ, Brinkman FS. 2003. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 19:418-20.
18. Tu Q, Ding D. 2003. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS Microbiol Lett 221:269-75.
19. Rajan I, Aravamuthan S, Mande SS. 2007. Identification of compositionally distinct regions in genomes using the centroid method. Bioinformatics 23:2672-7.
20. Langille MG, Hsiao WW, Brinkman FS. 2008. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics 9:329.
21. Ou HY, He X, Harrison EM, Kulasekara BR, Thani AB, Kadioglu A, Lory S, Hinton JC, Barer MR, Deng Z, Rajakumar K. 2007. MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. Nucleic Acids Res 35:W97-W104.
22. Langille MG, Hsiao WW, Brinkman FS. 2010. Detecting genomic islands using bioinformatics approaches. Nat Rev Microbiol 8:373-82.
23. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, Pereira SK, Waglechner N, McArthur AG, Langille MG, Brinkman FS. 2015. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res 43:W104-8.
24. Elleuche S, Poggeler S. 2010. Carbonic anhydrases in fungi. Microbiology 156:23-9.
25. Del Prete S, Vullo D, Fisher GM, Andrews KT, Poulsen SA, Capasso C, Supuran CT. 2014. Discovery of a new family of carbonic anhydrases in the malaria pathogen Plasmodium falciparum--the eta-carbonic anhydrases. Bioorg Med Chem Lett 24:4389-96.
26. Kikutani S, Nakajima K, Nagasato C, Tsuji Y, Miyatake A, Matsuda Y. 2016. Thylakoid luminal theta-carbonic anhydrase critical for growth and photosynthesis in the marine diatom Phaeodactylum tricornutum. Proc Natl Acad Sci U S A 113:9828-33.
27. Capasso C, Supuran CT. 2015. An overview of the alpha-, beta- and gamma-carbonic anhydrases from Bacteria: can bacterial carbonic anhydrases shed new light on evolution of bacteria? J Enzyme Inhib Med Chem 30:325-32.

28. Alterio V, Vitale RM, Monti SM, Pedone C, Scozzafava A, Cecchi A, De Simone G, Supuran CT. 2006. Carbonic anhydrase inhibitors: X-ray and molecular modeling study for the interaction of a fluorescent antitumor sulfonamide with isozyme II and IX. J Am Chem Soc 128:8329-35.

29. Nishimori I, Minakuchi T, Onishi S, Vullo D, Scozzafava A, Supuran CT. 2007. Carbonic anhydrase inhibitors. DNA cloning, characterization, and inhibition studies of the human secretory isoform VI, a new target for sulfonamide and sulfamate inhibitors. J Med Chem 50:381-8.

30. Vullo D, Franchi M, Gallori E, Antel J, Scozzafava A, Supuran CT. 2004. Carbonic anhydrase inhibitors. Inhibition of mitochondrial isozyme V with aromatic and heterocyclic sulfonamides. J Med Chem 47:1272-9.

31. Vullo D, Innocenti A, Nishimori I, Pastorek J, Scozzafava A, Pastorekova S, Supuran CT. 2005. Carbonic anhydrase inhibitors. Inhibition of the transmembrane isozyme XII with sulfonamides-a new target for the design of antitumor and antiglaucoma drugs? Bioorg Med Chem Lett 15:963-9.

32. Fasseas MK, Tsikou D, Flemetakis E, Katinakis P. 2010. Molecular and biochemical analysis of the beta class carbonic anhydrases in Caenorhabditis elegans. Mol Biol Rep 37:2941-50.

33. Syrjanen L, Tolvanen M, Hilvo M, Olatubosun A, Innocenti A, Scozzafava A, Leppiniemi J, Niederhauser B, Hytonen VP, Gorr TA, Parkkila S, Supuran CT. 2010. Characterization of the first beta-class carbonic anhydrase from an arthropod (Drosophila melanogaster) and phylogenetic analysis of beta-class carbonic anhydrases in invertebrates. BMC Biochem 11:28.

34. Zolfaghari Emameh R, Barker H, Tolvanen ME, Ortutay C, Parkkila S. 2014. Bioinformatic analysis of beta carbonic anhydrase sequences from protozoans and metazoans. Parasit Vectors 7:38.

35. Smith KS, Jakubzick C, Whittam TS, Ferry JG. 1999. Carbonic anhydrase is an ancient enzyme widespread in prokaryotes. Proc Natl Acad Sci U S A 96:15184-9.

36. Syrjanen L, Valanne S, Kuuslahti M, Tuomela T, Sriram A, Sanz A, Jacobs HT, Ramet M, Parkkila S. 2015. beta carbonic anhydrase is required for female fertility in Drosophila melanogaster. Front Zool 12:19.

37. Ali MY, Pavasovic A, Mather PB, Prentis PJ. 2015. Analysis, characterisation and expression of gill-expressed carbonic anhydrase genes in the freshwater crayfish Cherax quadricarinatus. Gene 564:176-87.

38. Bahn YS, Cox GM, Perfect JR, Heitman J. 2005. Carbonic anhydrase and CO2 sensing during Cryptococcus neoformans growth, differentiation, and virulence. Curr Biol 15:2013-20.

39. Chasen NM, Asady B, Lemgruber L, Vommaro RC, Kissinger JC, Coppens I, Moreno SNJ. 2017. A Glycosylphosphatidylinositol-Anchored Carbonic Anhydrase-Related Protein of Toxoplasma gondii Is Important for Rhoptry Biogenesis and Virulence. mSphere 2.

40. Nishimori I, Minakuchi T, Kohsaki T, Onishi S, Takeuchi H, Vullo D, Scozzafava A, Supuran CT. 2007. Carbonic anhydrase inhibitors: the beta-carbonic anhydrase from Helicobacter pylori is a new target for sulfonamide and sulfamate inhibitors. Bioorg Med Chem Lett 17:3585-94.

41. Innocenti A, Hall RA, Schlicker C, Muhlschlegel FA, Supuran CT. 2009. Carbonic anhydrase inhibitors. Inhibition of the beta-class enzymes from the fungal pathogens Candida albicans and Cryptococcus neoformans with aliphatic and aromatic carboxylates. Bioorg Med Chem 17:2654-7.

42. Zolfaghari Emameh R, Kuuslahti M, Vullo D, Barker HR, Supuran CT, Parkkila S. 2015. Ascaris lumbricoides beta carbonic anhydrase: a potential target enzyme for treatment of ascariasis. Parasit Vectors 8:479.

43. Syrjanen L, Kuuslahti M, Tolvanen M, Vullo D, Parkkila S, Supuran CT. 2015. The beta-carbonic anhydrase from the malaria mosquito Anopheles gambiae is highly inhibited by sulfonamides. Bioorg Med Chem 23:2303-9.

44. Rodrigues GC, Feijo DF, Bozza MT, Pan P, Vullo D, Parkkila S, Supuran CT, Capasso C, Aguiar AP, Vermelho AB. 2014. Design, synthesis, and evaluation of hydroxamic acid derivatives as promising agents for the management of Chagas disease. J Med Chem 57:298-308.

45. Syrjanen L, Vermelho AB, Rodrigues Ide A, Corte-Real S, Salonen T, Pan P, Vullo D, Parkkila S, Capasso C, Supuran CT. 2013. Cloning, characterization, and inhibition studies of a beta-carbonic

anhydrase from Leishmania donovani chagasi, the protozoan parasite responsible for leishmaniasis. J Med Chem 56:7372-81.

46. Vermelho AB, Capaci GR, Rodrigues IA, Cardoso VS, Mazotto AM, Supuran CT. 2017. Carbonic anhydrases from Trypanosoma and Leishmania as anti-protozoan drug targets. Bioorg Med Chem 25:1543-1555.

47. Supuran CT, Capasso C. 2016. New light on bacterial carbonic anhydrases phylogeny based on the analysis of signal peptide sequences. J Enzyme Inhib Med Chem 31:1254-60.

48. Zolfaghari Emameh R, Barker HR, Tolvanen ME, Parkkila S, Hytonen VP. 2016. Horizontal transfer of beta-carbonic anhydrase genes from prokaryotes to protozoans, insects, and nematodes. Parasit Vectors 9:152.

49. Clasen FJ, Pierneef RE, Slippers B, Reva O. 2018. EuGI: a novel resource for studying genomic islands to facilitate horizontal gene transfer detection in eukaryotes. BMC Genomics 19:323.

50. Del Prete S, Vullo D, di Fonzo P, Carginale V, Supuran CT, Capasso C. 2017. Comparison of the anion inhibition profiles of the beta- and gamma-carbonic anhydrases from the pathogenic bacterium Burkholderia pseudomallei. Bioorg Med Chem 25:2010-2015.

51. Naville M, Gautheret D. 2009. Transcription attenuation in bacteria: theme and variations. Brief Funct Genomic Proteomic 8:482-92.

52. Yousuf B, Sanadhya P, Keshri J, Jha B. 2012. Comparative molecular analysis of chemolithoautotrophic bacterial diversity and community structure from coastal saline soils, Gujarat, India. BMC Microbiol 12:150.

53. Acuna LG, Cardenas JP, Covarrubias PC, Haristoy JJ, Flores R, Nunez H, Riadi G, Shmaryahu A, Valdes J, Dopson M, Rawlings DE, Banfield JF, Holmes DS, Quatrini R. 2013. Architecture and gene repertoire of the flexible genome of the extreme acidophile Acidithiobacillus caldus. PLoS One 8:e78237.

54. Berben T, Overmars L, Sorokin DY, Muyzer G. 2017. Comparative Genome Analysis of Three Thiocyanate Oxidizing Thioalkalivibrio Species Isolated from Soda Lakes. Front Microbiol 8:254.

55. Dobrinski KP, Boller AJ, Scott KM. 2010. Expression and function of four carbonic anhydrase homologs in the deep-sea chemolithoautotroph Thiomicrospira crunogena. Appl Environ Microbiol 76:3561-7.

56. Roy Chowdhury P, Scott MJ, Djordjevic SP. 2017. Genomic islands 1 and 2 carry multiple antibiotic resistance genes in Pseudomonas aeruginosa ST235, ST253, ST111 and ST175 and are globally dispersed. J Antimicrob Chemother 72:620-622.

57. Freel KC, Krueger MC, Farasin J, Brochier-Armanet C, Barbe V, Andres J, Cholley PE, Dillies MA, Jagla B, Koechler S, Leva Y, Magdelenat G, Plewniak F, Proux C, Coppee JY, Bertin PN, Heipieper HJ, Arsene-Ploetze F. 2015. Adaptation in Toxic Environments: Arsenic Genomic Islands in the Bacterial Genus Thiomonas. PLoS One 10:e0139011.

58. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189-91.

59. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-1.

60. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589.

61. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-74.

62. Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188-95.

63. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33:1635-8.

64. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N. 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 44:W344-50.

65. Maglott D, Ostell J, Pruitt KD, Tatusova T. 2011. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39:D52-7.

TABLE 1 The single exon *β-CA* genes of protists.

No.	Protist species	NCBI IDs	Gene name
1	*Entamoeba* sp.	XP_004183626.1	EIN_065450
		XP_008860421.1	ENU1_204230
2	*Leishmania* sp.	XP_003858369.1	LDBPK_060630
		XP_010703940.1	LPMP_060590
3	*Leptomonas pyrrhocoris*	XP_015662104.1	ABB37_01925
		XP_015662099.1	ABB37_01923
4	*Paulinella chromatophora*	YP_002049530.1	PCC_0911
5	*Phytophthora infestans*	XP_002909256.1	PITG_00682
		XP_002909250.1	PITG_00674
6	*Tetrahymena thermophila*	XP_001009111.1	TTHERM_00263620
7	*Trichomonas vaginalis*	XP_001317907.1	TVAG_005270
8	*Trypanosoma grayi*	XP_009310034.1	DQ04_02331000

TABLE 2 β-CA protein sequences from bacterial species.

No.	Gram-staining	Bacterial species	NCBI IDs
1	Gram-negative	*Klebsiella pneumoniae*	WP_019705531.1
2		*Brucella abortus*	WP_002965854.1
3		*Yersinia enterocolitica*	WP_005165125.1
4		*Bordetella parapertussis*	YP_006895229.1
5		*Pseudomonas stutzeri*	WP_011914306.1
6	Gram-positive	*Streptomyces* sp.	WP_015579823.1
7		*Bifidobacterium angulatum*	WP_003825226.1
8		*Pseudonocardia* sp.	WP_060712833.1
9		*Desulfocapsa sulfexigens*	WP_015403686.1
10		*Arthrobacter alpinus*	WP_062006860.1

TABLE 3 β-CA protein sequences from protists.

No.	Protist species	NCBI IDs
1	*Acanthamoeba castellanii*	XP_004344666.1, XP_004335990.1, XP_004337607.1
2	*Capsaspora owczarzaki*	XP_004342925.1, XP_4349240.1
3	*Dictyostelium* sp.	XP_646739.1, XP_644170.1, XP_003283430.1, XP_004361116.1
4	*Entamoeba* sp.	XP_004183626.1, XP_008860421.1
5	*Galdieria sulphuraria*	XP_005703553.1
6	*Leishmania* sp.	XP_003858369.1, XP_010703940.1
7	*Leptomonas pyrrhocoris*	XP_015662104.1, XP_015662099.1
8	*Paulinella chromatophora*	YP_002049530.1
9	*Phaeodactylum tricornutum*	XP_002176594.1
10	*Phytophthora infestans*	XP_002909256.1, XP_002909250.1, XP_002909249.1
11	*Polysphondylium pallidum*	XP_020436034.1
12	*Saprolegnia diclina*	XP_008607403.1, XP_008604330.1
13	*Tetrahymena thermophila*	XP_001009617.1, XP_001009612.1, XP_001009111.1, XP_001022390.2, XP_001009116.2, XP_001009616.1, XP_976601.1, XP_001013978.2
14	*Trichomonas vaginalis*	XP_001317907.1, XP_001579768.1
15	*Trypanosoma grayi*	XP_009310034.1

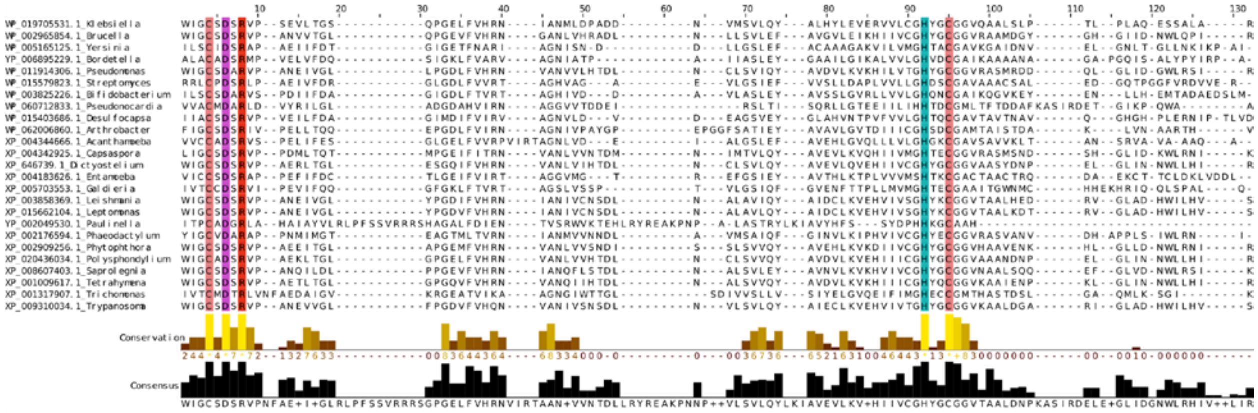
WP_019705531.1_Klebsiella
WP_002965854.1_Brucella
WP_005165125.1_Yersinia
YP_006895229.1_Bordetella
WP_011914306.1_Pseudomonas
WP_015579823.1_Streptomyces
WP_003825226.1_Bifidobacterium
WP_060712833.1_Pseudonocardia
WP_015403686.1_Desulfocapsa
WP_062006860.1_Arthrobacter
XP_004344666.1_Acanthamoeba
XP_004342925.1_Capsaspora
XP_646739.1_Dictyostelium
XP_004183626.1_Entamoeba
XP_005703553.1_Galdieria
XP_003858369.1_Leishmania
XP_015662104.1_Leptomonas
YP_002049530.1_Paulinella
XP_002176594.1_Phaeodactylum
XP_002909256.1_Phytophthora
XP_020436034.1_Polysphondylium
XP_008607403.1_Saprolegnia
XP_001009617.1_Tetrahymena
XP_001317907.1_Trichomonas
XP_009310034.1_Trypanosoma
Conservation
Consensus

FIG 1 Multiple sequence alignment (MSA) of β-CA protein sequences from prokaryotes and protists. The alignment of 25 β-CA protein sequences shows that they all contain the first (CXDXR; C: Cysteine, D: Aspartic acid, R: Arginine, and X: any residue) and second (HXXC; H: Histidine, C: Cysteine, and X: any residue) highly conserved motifs. The alignment begins three amino acid residues prior to the first highly conserved residues (CXDXR).

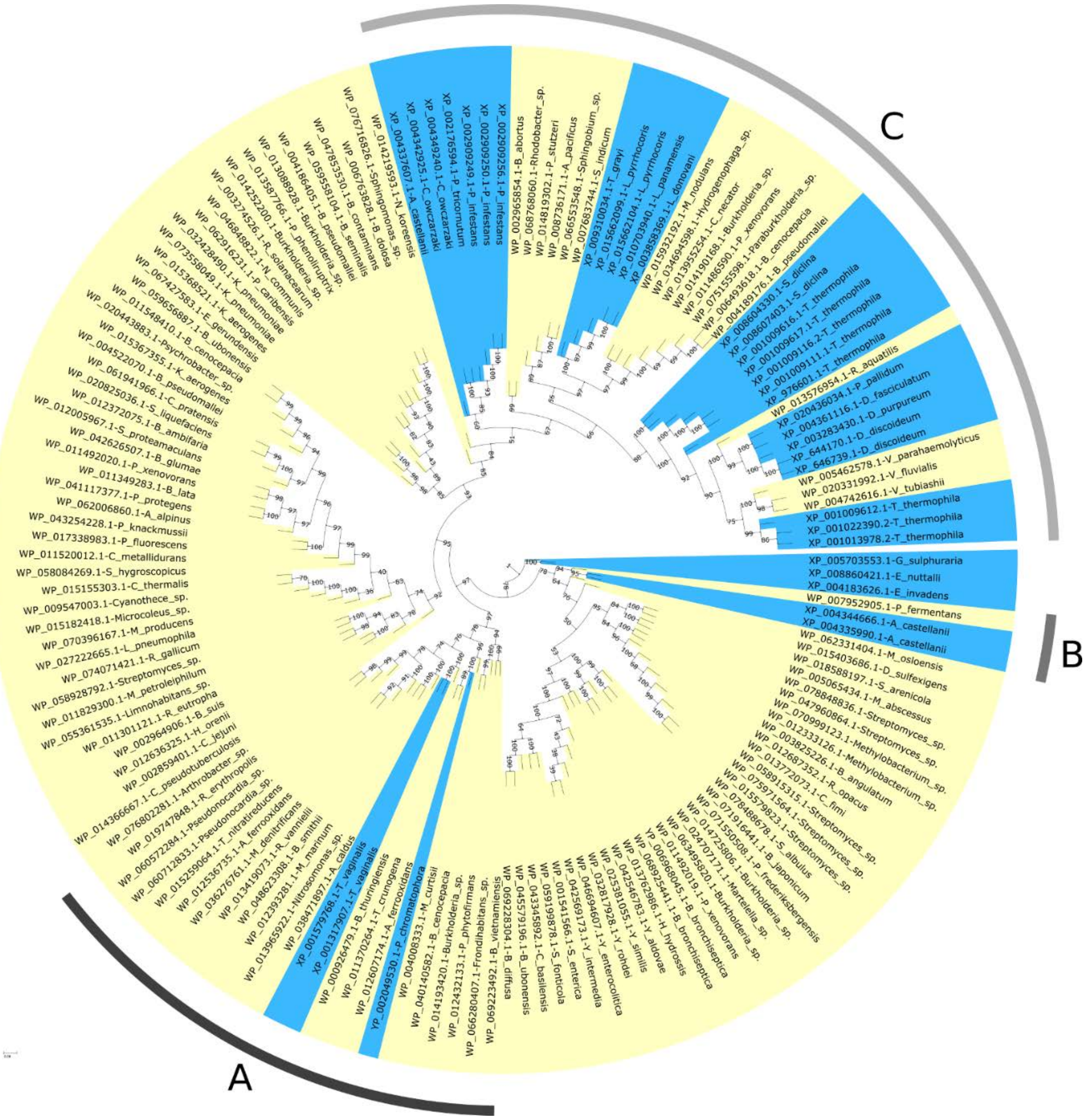

FIG 2 Phylogenetic analysis of β-CAs from prokaryotes and protists. Phylogenetic relationships were determined using the IQTree software for β-CAs from prokaryotes and protists, yellow and blue respectively. Three clades (A, B, and C) reveal regions where β-CAs from prokaryotes and protists appear to have a common ancestor.

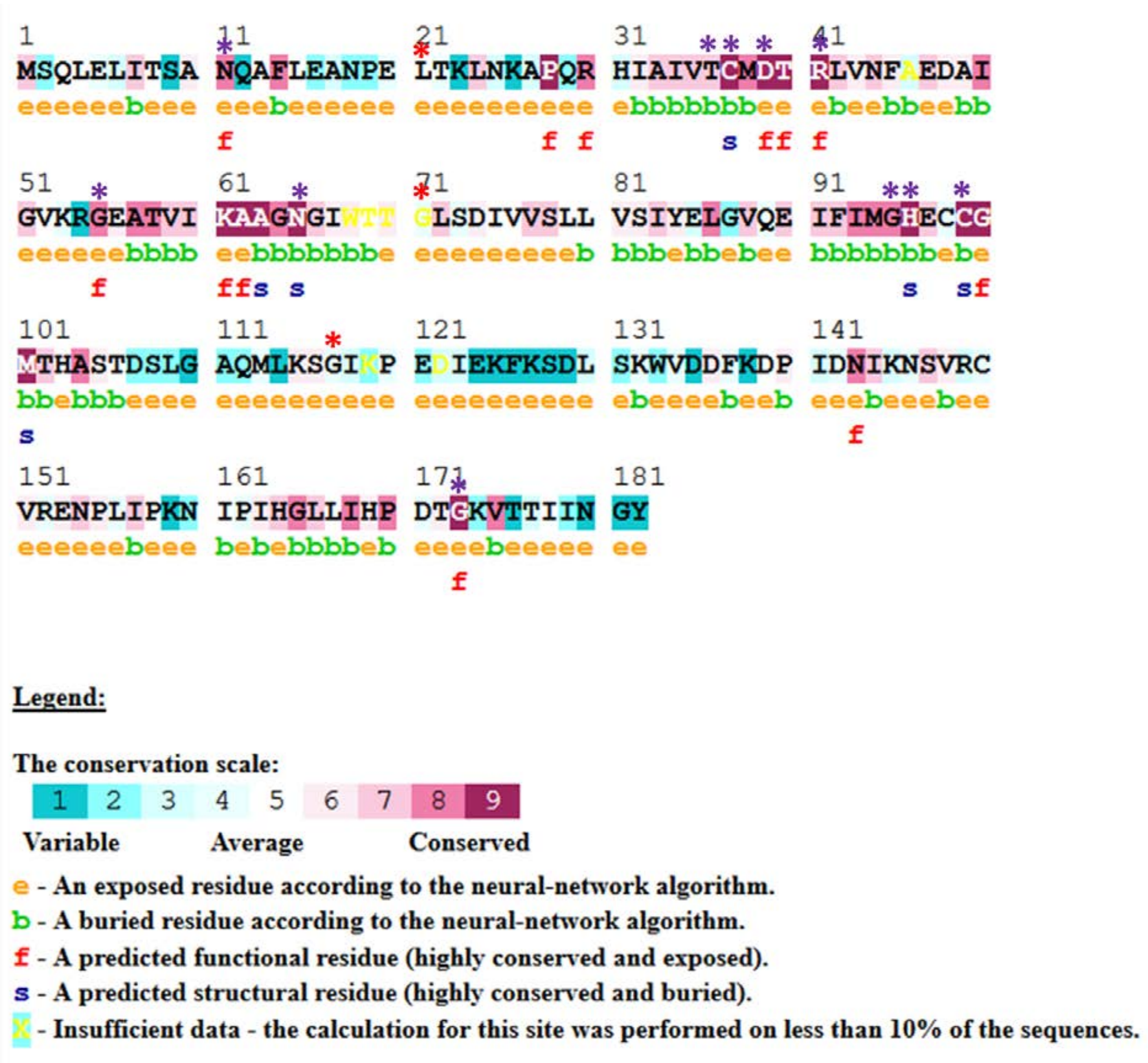
1 11 21 31 41
MSQLELITSA NQAFLEANPE LTKLNKAPQR HIAIVTCMDT RLVNFAEDAI
eeeeeebeee eeebeeeeee eeeeeeeeee ebbbbbbbee ebeebbeebb
f f f s ff f
51 61 71 81 91
GVKRGEATVI KAAGNGIWTT GLSDIVVSLL VSIYELGVQE IFIMGHECCG
eeeeeebbbb eebbbbbbbe eeeeeeeeeb bbbebbebee bbbbbbbebe
f ffs s s sf
101 111 121 131 141
MTHASTDSLG AQMLKSGIKP EDIEKFKSDL SKWVDDFKDP IDNIKNSVRC
bbebbbeeee eeeeeeeeee eeeeeeeeee ebeeeebeeb eeebeeebee
s f
151 161 171 181
VRENPLIPKN IPIHGLLIHP DTGKVTTIIN GY
eeeeeebeee bebebbbbeb eeeebeeeee ee
f
Legend:
The conservation scale:
1 2 3 4 5 6 7 8 9
Variable
Average
Conserved
e - An exposed residue according to the neural-network algorithm.
b - A buried residue according to the neural-network algorithm.
f - A predicted functional residue (highly conserved and exposed).
s - A predicted structural residue (highly conserved and buried).
X - Insufficient data - the calculation for this site was performed on less than 10% of the sequences.

FIG 3 Sequence conservation analysis. Consurf analysis performed for the β-CA sequences in the phylogenetic tree except Clade C and protists. The conservation score is projected onto the *T. vaginalis* CA sequence. The residues strictly conserved within the Clade C sequences are indicated with violet stars (red stars used for three residues with average or low concentration).

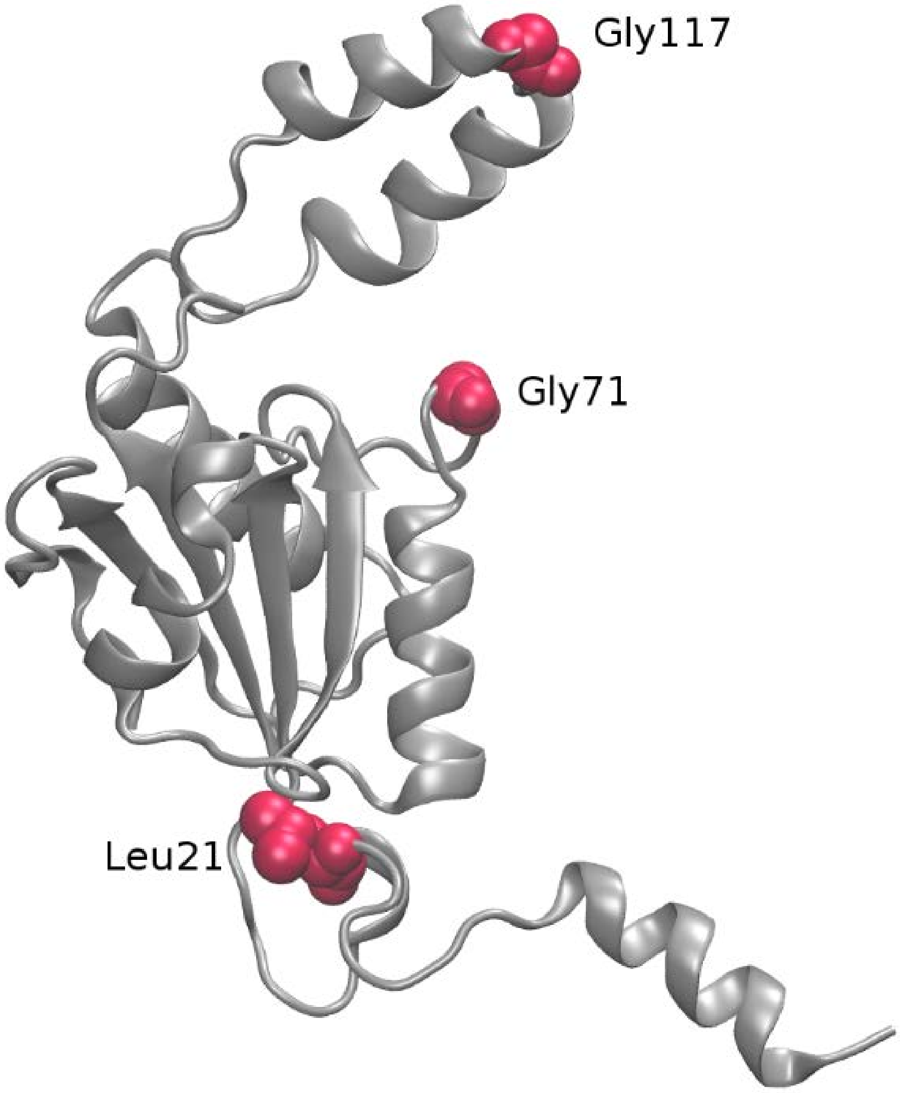

FIG 4 Evaluation of the functional importance of highly conserved residues in Clade C. A homology model of β-CA from *T. vaginalis* (A2ENQ8) (8) was used to project the conserved residues to a β-CA structure. Three conserved residues including Leu21, Gly71 and Gly117 were all located in flexible regions and are mostly exposed to solvent, indicating a non-essential structural role for these residues.