MARTTI TOLVANEN

Pseudogenes and Gene Duplications Tell a Story of Evolutionary Fates of Animal Alpha Carbonic Anhydrases

ACADEMIC DISSERTATION
To be presented, with the permission of the Board of
the Institute of Biomedical Technology of the University of Tampere,
for public discussion in the Auditorium of
School of Health Sciences, Medisiinarinkatu 3,
Tampere, on September 20th, 2013, at 12 o'clock.

UNIVERSITY OF TAMPERE

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere, Institute of Biomedical Technology
Tampere Graduate Program in Biomedicine and Biotechnology (TGPBB)
Finland

Supervised by
Professor Mauno Vihinen
Lund University
Sweden
Affiliated research group
Institute of Biomedical Technology
University of Tampere
Finland

Reviewed by
Professor Joachim Deitmer
University of Kaiserslautern
Germany
Docent Tommi Nyrönen
University of Helsinki
Finland

Copyright ©2013 Tampere University Press and the author

Cover design by
Mikko Reinikka

Acta Universitatis Tamperensis 1847
ISBN 978-951-44-9197-9 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1327
ISBN 978-951-44-9198-6 (pdf)
ISSN 1456-954X
http://tampub.uta.fi

Suomen Yliopistopaino Oy – Juvenes Print

Table of Contents

List of original communications

This thesis is based on the following original communications, which are referred to in the text by their Roman numerals (I-III).

I. Hilvo, M.*, **Tolvanen, M.***, Clark, A., Shen, B., Shah, G.N., Waheed, A., Halmi, P., Hänninen, M., Hämäläinen, J.M., Vihinen, M., Sly, W.S. & Parkkila, S. 2005, "Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase", *The Biochemical Journal*, 392: 83-92.
*: equally contributed

II. **Tolvanen, M.E.E.**, Ortutay, C., Barker, H.R., Aspatwar, A., Patrikainen, M. & Parkkila, S. 2013, "Analysis of evolution of carbonic anhydrases IV and XV reveals a rich history of gene duplications and a new group of isozymes", *Bioorganic & Medicinal Chemistry*, 21:1503-10

III. Ortutay, C., Nair, P.S., Aspatwar, A., Parkkila, S., Vihinen, M. & **Tolvanen, M.E.E.** 2013, "Computational inference of functional relationships from multiple genomes - Drosophila carbonic anhydrases" [Manuscript]

Abbreviations

AE anion exchanger (protein)

BLAST Basic Local Alignment Search Tool

BLAT Blast-Like Alignment Tool

CA carbonic anhydrase

CAH carbonic anhydrase (in Drosophila)

CAHZ CA II-like cytoplasmic CA in fishes

CARP carbonic anhydrase related protein

cDNA complementary DNA

chr chromosome

CP cytoplasmic (domain)

DEAE diethylaminoethyl

DGCR di George critical region

EST expressed sequence tag

GPI glycosyl phosphoinositol

HMM hidden Markov model

MCT monocarboxylate transporter

NCBI National Center for Biotechnology Information

NE non-existent

PAUP* Phylogenetic Analysis Using Parsimony (* and Other Methods)

PCR polymerase chain reaction

PDB Protein Data Bank

PG proteoglycan (domain)

PI-PLC phosphatidylinositol-specific phospholipase C

PTPR protein tyrosine phosphatase receptor

PTPRG protein tyrosine phosphatase receptor gamma

PTPRZ protein tyrosine phosphatase receptor zeta

PTX pentraxin domain

RBC red blood cell

RT-PCR reverse transcription polymerase chain reaction

SDS-PAGE sodium dodecyl sulfate polyacrylamide gel electrophoresis

TM transmembrane (domain)

UCSC University of California, Santa Cruz

UCSF University of California, San Francisco

Tiivistelmä

Tässä väitöskirjassa olen tutkinut kahta eläinten α-hiilihappoanhydraasien (CA:iden) osaperhettä, nimittäin GPI-sidottuja (glykosyylifosfoinositolisidottuja) α-CA:ita selkärankaisissa ja kaikkia α-CA:ita kahdentoista mahlakärpäsen genomeissa.

GPI-sidottujen CA:iden osaperheessä teimme ensimmäiset biokemialliset tutkimukset uudelle isoentsyymille, hiiren CA XV:lle, joilla osoitimme sen olevan CA IV:n tavoin glykosyloitu, GPI-ankkuroitu kalvoproteiini. *CA15*-geenin ilmentyminen havaittiin hiiren munuaisissa, kiveksissä ja aivoissa, kuten ihmisen *CA4*:llä.

Tutkimuksemme osoittavat *CA15*:n olevan pseudogeeni kädellisissä, ja siinä nähdään monia eri geenivirheitä eri lajeissa. Niiden analyysistä ilmenee, että yhden aktiivisen keskuksen histidiinin mutaatio tai insertio aktiivisen keskuksen lähellä oli luultavasti kädellisillä lajeilla alkusyy *CA15*:n inaktivoitumiselle ja pseudogeeniksi muuttumiselle.

Kaikkien selkärankaisten *CA4*:n ja *CA15*:n kaltaisten geenien fylogeneettisellä analyysillä on selvitetty kahdentumien todennäköinen historia tässä CA-alaperheessä, mikä johti uuden CA-isoentsyymin, CA XVII:n, löytöön nisäkkäiden ulkopuolisissa selkärankaisissa. Seeprakalan kahdeksan *CA4*- ja *CA17*-geenin ilmentymistietojen analyysi viittaa samojen tehtävien työnjakoon, jotka ihmisellä ovat *CA4*-geenituotteen vastuulla. Kaloilla näitä isoentsyymien geenejä on useita kopioita, minkä arvelen heijastavan eroja kalojen ja maalla elävien selkärankaisten välillä hengityksen ja hapon erityksen fysiologiassa.

α-CA:iden joukko Drosophila-mahlakärpästen genomeissa on yhtä suuri kuin nisäkkäillä, mutta varsin erilainen koostumukseltaan ja evoluutiohistorialtaan. Käytimme evoluutioon perustuvaa korrelaatiomenetelmää neljän eri α-CA-ryhmän tunnistamiseksi, joilla on kullakin erityiset ominaispiirteensä.

GPI-liitetyn osaperheen vertailu eri lajeissa sekä selkärankaisten ja Drosophilan α-CA:iden vertailu osoittavat CA-entsyymijärjestelmän muovautuvuuden, miten toimintoja voidaan jakaa eri tavoin entsyymiperheen jäsenten välillä.

Abstract

In this study I have examined two subfamilies within metazoan α-CAs, namely the GPI-linked α-CAs in vertebrates and all α-CAs in 12 fruit fly genomes.

Within the GPI-linked subfamily, we characterized biochemically the novel isozyme, murine CA XV, to show that it is a glycosylated, GPI-anchored, membrane protein, similar to CA IV. The expression of gene *CA15* was seen in mouse kidney, testis, and brain, similar to human *CA4*.

Our studies show that *CA15* is a pseudogene in primates, with multiple defects seen in different species. Their analysis suggests that the inactivation and 'pseudogenization' of *CA15* in primates was probably initiated by a mutation of one of the active-site histidines or an insertion near the active site.

Phylogenetic analysis of all vertebrate *CA4*/*CA15*-like genes has elucidated the likely history of duplications within this CA subfamily and led to the discovery of a novel CA isozyme, CA XVII, in non-mammalian vertebrates. Analysis of expression data for the eight *CA4* and *CA17* CA genes in zebrafish hints at a division of tasks which are the responsibility of the human *CA4* gene product. I postulate that the high multiplicity of the genes for these isozymes in fish reflects the differences in the physiology of breathing and acid excretion between fish and land-living vertebrates.

The set of α-CAs in Drosophila genomes is as large as in mammals, but quite different in its composition and evolution. We used an evolution-based method of correlations to identify four distinct α-CA sets with specific characteristics.

The comparisons of the GPI-linked subfamily in various species on one hand, and between the α-CAs of vertebrates and Drosophila on the other hand, demonstrate the plasticity of the CA enzyme system, how functionalities can be redistributed between the members of an enzyme family.

1. Introduction

Carbonic anhydrases (CAs) are enzymes which catalyze the reversible reaction of carbon dioxide with water to form hydrogen ion and bicarbonate:

$CO_2 + H_2O \leftrightarrow H^+ + HCO_3^-$

Enzymatically active CAs contain a metal cofactor, most frequently zinc. The metal centre is required for the production of a metal-bound hydroxide ion (OH^-), which is the active species in the reaction (Lindskog, Coleman 1973).

CAs are ubiquitous enzymes which are found in all kingdoms of life (Hewett-Emmett 2000). There are three major gene families for this catalytic activity (EC 4.2.1.1), namely α-, β-, and γ-CAs, which bear no resemblance to each other, neither by sequence nor by structure, suggesting that the same function has emerged three times in evolution (Liljas, Laurberg 2000). We know the α-CAs best, because vertebrate CAs are exclusively α-CAs, and human is the most intensively studied model organism. Invertebrate and fungal genomes contain additionally β-CAs, whereas plants possess CAs of all three major groups (Syrjänen *et al.* 2010, Elleuche, Poggeler 2009, Zimmerman, Ferry 2008). Bacteria contain nearly universally at least one β-CA gene per genome and less frequently α- and γ-CAs, whereas Archeae would seem to be devoid of α-CAs (Zimmerman, Ferry 2008). Two minor families of CAs, the δ- and ζ-CAs, have been found only in marine phytoplankton, such as diatoms (McGinn, Morel 2008).

CA activity is crucial in maintaining the homeostasis of pH and availability of CO_2 and/or HCO_3^-. To cater for these needs all multicellular eukaryotes maintain multiple CA isozymes that differ in their subcellular localization, tissue distribution, protein-protein interactions, and kinetic properties. For example, the mammalian CA set comprises thirteen enzymatically active members: CA I, II, III, IV, VA, VB, VI, VII, IX, XII, XIII, XIV, and XV [reviewed in (Supuran 2008, Hilvo *et al.* 2008)], plus additionally three acatalytic α-CAs, or CA-related proteins (CARPs): CARP VIII, X, and XI [reviewed in (Tashian *et al.* 2000, Aspatwar, Tolvanen & Parkkila 2010)].

CA XV was characterized as a part of this doctoral study, and the discovery that the corresponding *CA15* gene is inactive in human and chimpanzee genomes was the first demonstration that different mammalian species have different complements of CA genes.

In fishes some parts of the CA isozyme system are slightly different from that of mammals. Instead of the gene cluster coding for cytoplasmic isozymes CA I, II, III, and XIII in mammals, fishes have just one or two other genes which have descended from the common ancestor of these CA isozymes (Peterson, Tu & Linser 1997, Esbaugh, Tufts 2006, Lin *et al.* 2008). In contrast, the fish set of CAs in the group which contains the glycosyl phosphoinositol-anchored (GPI-anchored) CA IV and CA XV is somewhat more complex than in mammals (Lin *et al.* 2008, Gilmour, Perry 2009). When we expanded the study of CA IV-related isozymes into all available vertebrate genomes, we discovered phylogenetic relationships not noticed in the previous studies (Lin *et al.* 2008), most notably a novel CA isozyme class. The full analysis of GPI-anchored CA isozymes became the main focus in the second article in this study.

Further examples of complex division of labour between CAs can be found in plants and phytoplankton. In Arabidopsis thaliana, there are 3 γ-CAs and 2 γ-CA-like proteins, 6 β-CAs, and at least 3 α-CAs [(Parisi *et al.* 2004, Sunderhaus *et al.* 2006); Tolvanen MEE, unpublished database search results]. The genome of the diatom *Thalassiosira pseudonana* contains potential genes for 3 α-CAs, 5 β-CAs, 4 δ-CAs, and one ζ-CA (Tachibana *et al.* 2011).

There are various lineage-specific CA gene duplications and deletions in insects and other arthropods, similar to evolution within vertebrates. For example, the genome of *Drosophila melanogaster* contains 15 genes for α-CAs and CARPs, 8 of which are Drosophila-specific paralogues, and one gene for a β-CA (Ortutay *et al.* 2010, Syrjänen *et al.* 2010). If we compare the above examples from mammalian, plant, and insect genomes, it is obvious that there are many different ways to form a set of multiple CAs to work in all required locations and circumstances. However, there is little published information on functional roles of Drosophila CAs. This took us to the third study of this thesis, an analysis of sequence conservation and functional grouping in all α-CAs in the genomes of 12 Drosophila species.

2. Review of the literature

2.1 General functions of carbonic anhydrases

Carbon dioxide is a gas which is moderately soluble in water and easily diffusible through biological membranes, whereas the bicarbonate ion is highly soluble and impermeant through membranes, requiring specific anion transport proteins. Carbon dioxide and bicarbonate have a central role in the pH homeostasis in all cells, being one of the three major buffer systems, in addition to phosphate and ammonia. The unhindered diffusion of CO_2 and its ultimate disposal into gas phase (in air-breathing animals) make this system especially versatile in pH regulation. The reaction $CO_2 + H_2O \leftrightarrow H^+ + HCO_3^-$ occurs spontaneously even when uncatalyzed, but at a slow rate. The ubiquitous presence of CAs in all tissues guarantees that the above reaction is in rapid equilibrium, so that diffusion and transport of CO_2 and bicarbonate have a high efficacy in the regulation of pH.

In a global perspective, the most essential function of CA is in assisting CO_2 fixation. Plants, phytoplankton, and photosynthetic cyanobacteria have an ability to concentrate and fix inorganic carbon from CO_2 in photosynthesis. CA activity is implicated in cellular/chloroplastic retention of CO_2 as well as in providing a saturating supply of CO_2 to the key enzyme, ribulose 1,5-bisphosphate carboxylase/oxygenase in chloroplasts or carboxysomes (Dudoladova *et al.* 2007, Giordano, Beardall & Raven 2005, Reinfelder 2011, Yamano, Fukuzawa 2009). CAs are also suggested to play a role in the retention and recycling of CO_2 released from plant mitochondria (Zabaleta, Martin & Braun 2012).

Bicarbonate participates in numerous metabolic processes, such as fatty acid biosynthesis and gluconeogenesis (and other pathways requiring biotin carboxylation), formation of urea, etc. (a full list can be found in the Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/dbget-bin/www_bget?cpd:C00288). CO_2 is a required substrate in photosynthesis and a terminally oxidized product released in many essentially irreversible reactions, such as those catalyzed by isocitrate dehydrogenase and α-ketoglutarate dehydrogenase in the citric acid cycle (see full list in http://www.genome.jp/dbget-bin/www_bget?cpd:C00011). CAs can obviously be useful in providing a speedy supply of the appropriate substrates.

The involvement of CAs in the physiology of tetrapods is most prominent in the respiratory system and kidneys, which collaborate in reacting to acid-base changes by means of respiratory compensation (elimination of CO_2) and/or metabolic compensation (excretion of acid in the urine), respectively. Fish have less chance for respiratory compensation due to low O_2 supply in water and a low CO_2 partial pressure in the blood, so the acid-base regulation in the majority of fish is based on adjusting plasma HCO_3^- levels through the independent regulation of H^+ and HCO_3^- effluxes across the gill. Fish kidneys collaborate by adjustments in bicarbonate reabsorption (to a variable extent in different fish genera). Again, CAs are required for fast conversions of H^+, HCO_3^-, and CO_2, and as one would expect, high CA activity is found in gill and kidney (Lin *et al.* 2008, Gilmour, Perry 2009).

Other physiological processes in which CAs are implicated include bone resorption (Väänänen, Parvinen 1983) and ion transport, which will be discussed in more detail in section 2.4 on p. 18.

2.2 Carbonic anhydrase families, the wheel reinvented?

2.2.1 Three major families: α, β, and γ

The majority of all CAs can be assigned to one of the three main classes by their amino acid sequences. These classes, designated α, β, and γ, were also seen to be completely distinct when corresponding structures were elucidated by x-ray crystallography. The obvious conclusion was that different classes had originated and evolved independently for the same enzymatic activity (Liljas, Laurberg 2000).

Figure 1 shows example structures of the major CA families. Active sites are highlighted by sphere representation of the metal atom and stick representation of the metal-binding residues.

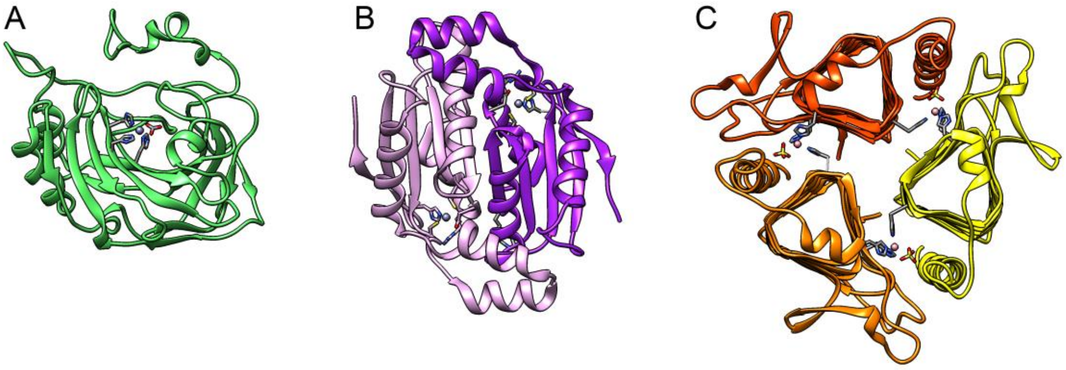

Figure 1 Representatives of the major CA classes in the same scale. 3D structures retrieved from PDB, visualization with Chimera by M. Tolvanen. Catalytic metal atoms and metal-binding amino acid side chains are shown in atomic detail. **A**, α-CA: human CA I, PDB 1hcb (Kumar, Kannan 1994); **B**, β-CA: “cab-type” dimer from the archaeon *Methanobacterium thermoautotrophicum*, chains A and B of PDB 1g5c (Strop *et al.* 2001); and **C**, γ-CA from the archaeon *Methanosarcina thermophila*, chain A of PDB 1qrf (Iverson *et al.* 2000), tripled by crystallographic symmetry operations to give the biological trimer.

The α-CA fold is characterized by an extensive, antiparallel β-sheet (violet colour in Fig. 1 A). The three zinc-binding histidines are on the β-sheet, not far from the centre of the protein, at the bottom of a deep tunnel.

2.2.2 The minor families: δ and ζ

There are two additional CA enzyme classes with very limited occurrence, with designations δ and ζ.

The first δ-CA was discovered in the diatom *Thalassiosira weissflogii* (Roberts, Lane & Morel 1997), and later found to be present in a broader range of eukaryotic phytoplankton including haptophytes, prasinophytes and dinoflagellates (McGinn, Morel 2008). There are no structures available of any δ-CA, but X-ray absorption spectroscopy has revealed that the catalytic Zn atom has three histidine ligands, similar to α and γ classes (Cox *et al.* 2000). Because δ-CAs do not show any detectable sequence similarity to other CA classes, they may represent yet another case of convergent evolution.

The carboxysomal CA of the chemolithoautotrophic bacterium *Halothiobacillus neapolitanus* was originally reported as an example of a novel CA class, ε-CA (So *et al.* 2004). However, when the structure was solved (Sawaya *et al.* 2006), it was seen to be a special case of the β-CAs, as shown in Fig. 2E. This enzyme contains only one catalytic site in two β-CA domains plus additional domains (shown in grey), presumably responsible for interactions with other proteins on the carboxysomal shell. Figure 2 displays the structural similarity between β-, ε-, and ζ-CAs. In each structure there is a 180-degree rotation pseudosymmetry axis near the midpoint, towards the viewer (not shown), which relates the red/yellow domain or subunit to the cyan/purple one. In panels A, B, and D the structures are dimers, whereas panels C, E, and F show single-chain proteins which have a similar two-domain organization. The active sites are indicated by sphere representation of the metal ion (Zn in A to E, Cd in F) and stick representation of the metal-binding residues (one His and two Cys) plus the conserved Asp and Glu residues, which contribute to the reaction rate increase (Smith, Ingram-Smith & Ferry 2002). Grey regions are specific to each structure (A to E) or not alignable to others due to sequence permutation (F).

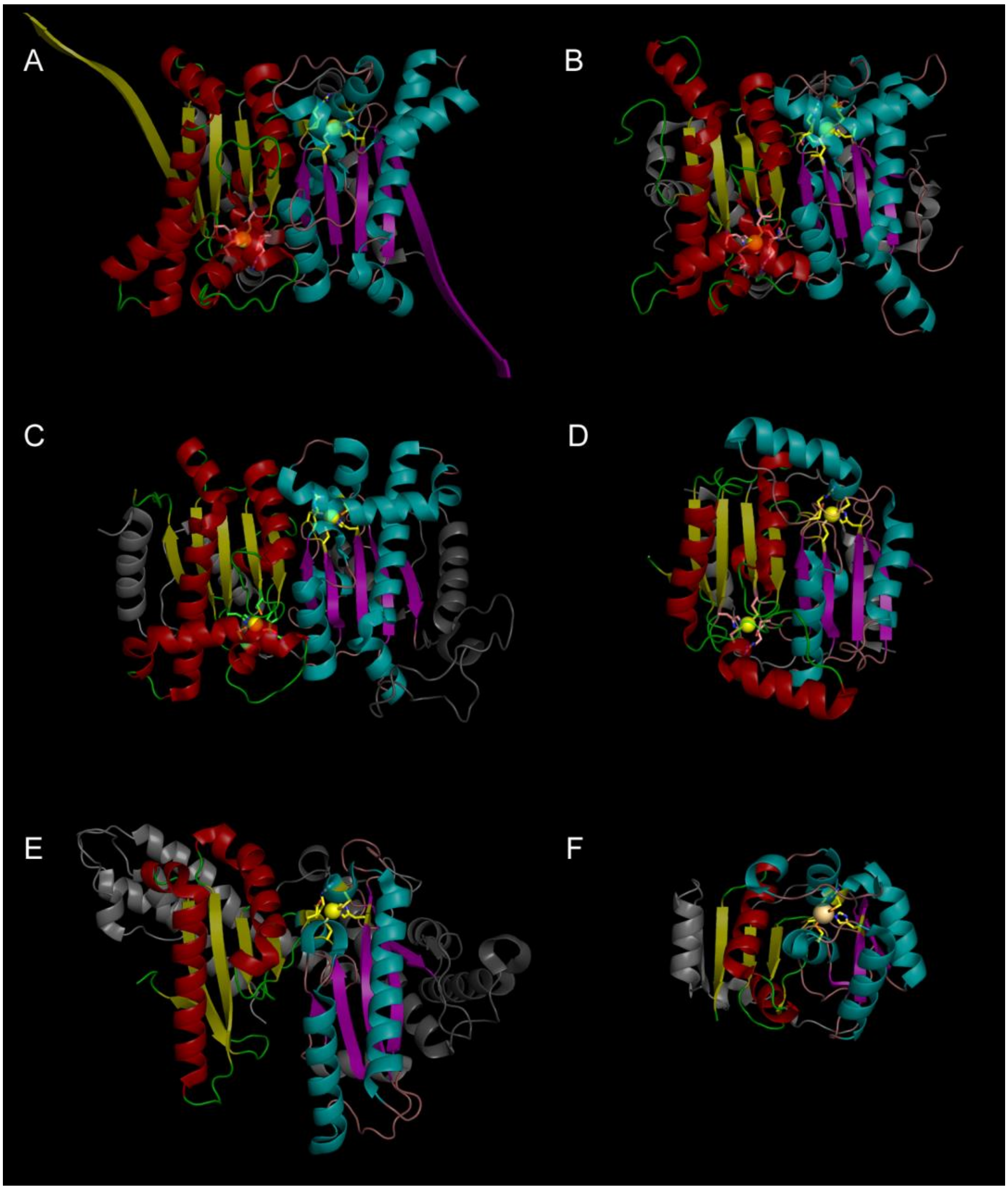

Figure 2 Structures of β-CA-like proteins, visualized with PyMol by M. Tolvanen. All structures A to F are in same scale. Yellow and purple indicate conserved β-strands, red and cyan are conserved α-helices, whereas grey parts are structural elements not found in all β-CAs, or ones which come from an unexpected part of the sequence. **A,** β-CA dimer from pea, *Pisum sativum*, chains A and B of the octamer, PDB 1ekj (Kimber, Pai 2000); **B,** Can2 β-CA dimer from the fungus *Cryptococcus neoformans*, chains B and C of PDB 2w3n (Schlicker *et al.* 2009); **C,** dual-domain β-CA from the red alga *Porphyridium purpureum*, chain A of PDB 1ddz (Mitsuhashi *et al.* 2000); **D,** cab-type β-CA Rv1284 dimer from *Mycobacterium tuberculosis*, chains A and B of PDB 1ylk (Suarez Covarrubias *et al.* 2005); **E,** carboxysomal β-CA (previously "ε-CA") from *Halothiobacillus neapolitanus*, chain A of PDB 2fgy (Sawaya *et al.* 2006); and **F,** ζ-CA from marine diatom *Thalassiosira weissflogii*, PDB 3BOB (Xu *et al.* 2008).

2.3 Vertebrate α-CA family

Vertebrate CAs occur as 10+ isoforms. In human there are 12 enzymatically active CA isozymes and three acatalytic “CA-related proteins” (CARPs). They are found in various subcellular localizations as indicated in Fig. 3. The group traditionally named as “cytoplasmic” CAs (I, II, III, VII, and XIII) also includes peripheral membrane proteins, as there is evidence for direct binding to ion transporter proteins on the cytoplasmic face of the plasma membrane at least in the case of CA II (Vince, Reithmeier 2000).

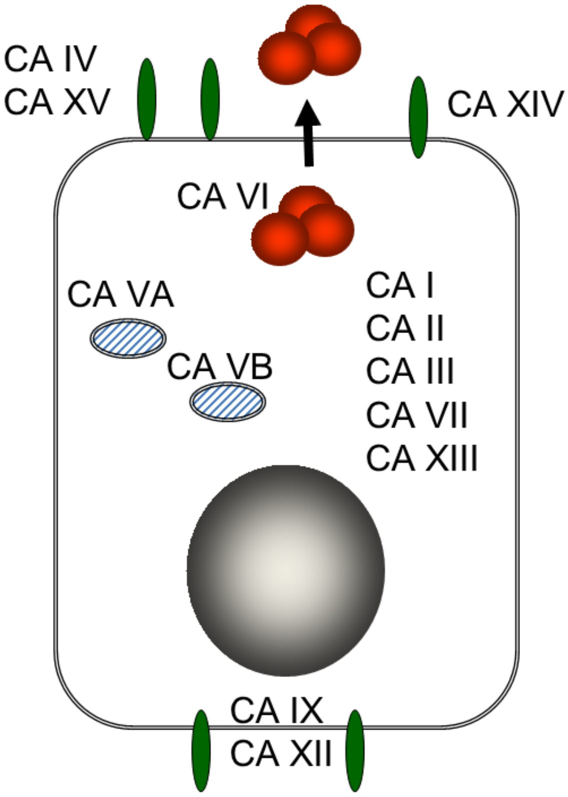

Figure 3 Localizations of CA isozymes in mammals in a graphic mammalian cell model. Top and bottom represent the apical and basolateral surfaces, respectively, in polarized cells. Diagonally shaded areas represent mitochondria. Image courtesy of S. Parkkila.

The exact composition of the CA set varies slightly in different groups of vertebrates, most notably between fish and tetrapods (Esbaugh, Tufts 2006). This study contributes new information in this aspect, which will be summarized in the results section.

2.4 Carbonic anhydrases and ion transport

Whenever an ion transporter or channel protein transports bicarbonate or H^+, it is obvious that the availability of CA activity on one or the other side of the channel will help by either forming more of the species to be transported, or by converting the already transported species to CO_2 and/or H_2O. In the case of CA IV in the kidneys this leads to facilitated acid excretion, as reviewed by Breton (Breton 2001). In addition to this essential role in physiological pH homeostasis, CA IV is involved in assistance of ion transport in other contexts, such as catalytic assistance to anion exchanger 1 (AE1) in erythrocytes (Sterling, Alvarez & Casey 2002), to monocarboxylate transporter 1 (MCT1) in skeletal muscle (Wetzel *et al.* 2001), to AE3 in neurons (Svichar *et al.* 2009), and to sodium-bicarbonate cotransporter 1 (NBC1) in the choriocapillaris (Yang *et al.* 2005). In addition, CA IV can also provide non-catalytic assistance to lactate transport of MCT2 (Klier *et al.* 2011). Furthermore, CA IV also facilitates MCT-mediated acid/base flux in the astrocytes of rat brain (Svichar, Chesler 2003, Svichar *et al.* 2006). In rat hippocampal neurons it has been demonstrated that both CA XIV and CA IV enhance the chloride/bicarbonate exchange at AE3 (Svichar *et al.* 2009).

The physiology of oxygen transport and breathing is largely involved with the quick loading and unloading of bicarbonate in the red blood cells (RBC). The major player on the RBC surface is the chloride/bicarbonate exchanger AE1, which is assisted both externally by CA IV (Sterling, Alvarez & Casey 2002) and internally by CA II (Vince, Carlsson & Reithmeier 2000, Vince, Reithmeier 2000).

Even the acatalytic CARP VIII is involved in regulation of ion transport, modulating the activity of the inositol trisphosphate receptor, a calcium channel on the endoplasmic reticulum (Hirota *et al.* 2003). We would expect that many more such interactions between ion transporters and CAs or CARPs will be discovered.

2.5 Whole-genome duplications in vertebrates

The idea that whole-genome duplications have been a major driving force in evolution was presented nearly 50 years ago. Solid evidence to confirm this mechanism was collected easily in plants and fungi (Wolfe, Shields 1997), but in case of vertebrates the amount of sequence and gene map data was not sufficient for rigorous proof until 2002 (McLysaght, Hokamp & Wolfe 2002). The original study confirmed at least one duplication event in early vertebrates, before the split of fish and tetrapod lineages, the great radiation of jawed vertebrates (gnathostomes).

A few years later, another study gave “unmistakable evidence” that there have indeed been two rounds of whole-genome duplication in early vertebrates, and that four-way paralogous regions still cover a large part of the human genome (Dehal, Boore 2005). The authors focused on the global organization of paralogous genes which were duplicated before the split of fish and tetrapod lineages, using the sea squirt *Ciona intestinalis* as their non-gnathostomal reference, so the question of timing of the duplication events was not resolved.

Another study from 2002 investigated the *en bloc* duplication data focusing on the genes of the major histocompatibility complex (Abi-Rached *et al.* 2002). Their non-gnathostomal reference organism was a cephalochordate, amphioxus (*Branchiostoma*), which is considered similar to the archetypal vertebrate form. Their timing estimate for the duplications is after the divergence of cephalochordates and vertebrates but before the Gnathostomata radiation.

The jawless fishes (Agnathostomata) hold a critical phylogenetic position as a group that diverged from vertebrates before the radiation of jawed vertebrates. Extant species in this class include the round-mouthed (Cyclostomata) lampreys and hagfishes. Whether the two-round genome duplication occurred before or after the divergence of Agnathostomata has been studied by Kuraku and coworkers. Their conclusion, documented in several publications (Matsuura *et al.* 2008, Kuraku 2008, Kuraku, Meyer & Kuratani 2009, Kuraku 2010), is that Agnathostomata diverged after the two rounds of duplications.

3. Aims of the study

This study was initiated in early 2004 by Mika Hilvo's course work on a previously unreported pseudogene for CA14. A full search of CA-like sequences in the human genome also resulted in the discovery of three human orthologues for *CA15*, apparently pseudogenes. At the time, mouse *CA15* existed in sequence databases but there was no biochemical protein data. I extended the sequence search for *CA15* to all available genomes. In the process, I extracted numerous previously unreported CA protein and DNA sequences and observed unknown or insufficiently characterized evolutionary relationships within α-CAs in animals. It was obvious that bioinformatic and phylogenetic studies of CAs could lead to a significant number of scientific findings and articles.

With this, I embarked on a mission to complete and publish as much work as possible on the evolutionary origins, later evolutionary fates, and characterization of carbonic anhydrases in animals. The thesis at hand consists of three publications from this process.

Specific aims and questions posed for the research in this thesis:

1) Recombinant production and biochemical characterization of mouse CA XV, and expression pattern of the mouse *CA15* gene. (I)
2) Confirmation of the pseudogene status of human *CA15* orthologues. (I)
3) Where and how was CA XV inactivated in the primate lineage? (I, II)
4) How did the complex set of CA IV-like enzymes evolve in fishes? (II)
5) Can we combine bioinformatic analyses and data mining to gain understanding of the α-CA family in multiple, highly similar genomes, such as those of fruit flies? (III)

4. Materials and methods

4.1 Bioinformatic methods

4.1.1 Sequence search and reconstruction (I, II, III)

In publication I, the retrieval of sequences for CA XV required many manual steps due to missing annotation in the genomes. The procedure was started with the search and retrieval of known *CA15* sequences (*Mus musculus*). Then BLAT (Kent 2002) searches were performed in all selected genomes using sequences found in the previous step as query sequences. The genomic sequences were obtained, and these were translated in three frames. The translations were aligned with known sequences, and the exon locations were visually identified. The gene models were confirmed and fine-tuned using GeneWise (Birney, Clamp & Durbin 2004). Finally, the final best transcripts and protein sequences were assembled manually.

The following sequences were obtained from Ensembl Genome Browser (http://www.ensembl.org) (Flicek *et al.* 2012): *M. musculus* (ENSMUSP00000012152) and *Rattus norvegicus* (ENSRNOP00000000312). The UCSC Genome Browser (http://genome.ucsc.edu) (Kent *et al.* 2002, Meyer *et al.* 2013) showed two mRNAs for the *CA15* of *Gallus gallus*: of these two the accession number BX929589 was selected and translated into protein because it was in closest accordance with other species. For *Danio rerio* there was an EST sequence (CO960501) representing *CA15* that was mainly of high quality. Because the latter part of the EST sequence was of lower quality, it was constructed manually from the genome. The sequence for *Tetraodon nigroviridis* was also obtained from UCSC Genome Browser, Genscan Gene Prediction (GSCT00001777001). The *CA15* of *Fugu rubripes* was constructed manually using the information of Ensembl Genome Browser (SINFRUP00000165581 and SINFRUP00000175429). The UCSC Genome Browser also showed some EST sequences for this gene. The gene for the *CA15* of *Canis familiaris* could not be found in the biological databases and it was constructed with GeneWise based on a BLAT hit in dog chromosome 26. The CA XV amino acid sequence of *Xenopus tropicalis* was constructed manually from two EST sequences.

For humans (*Homo sapiens*) or chimpanzees (*Pan troglodytes*) there were no mRNAs or EST sequences representing *CA15* in the Ensembl Genome Browser. However, the human genome was shown to have three copies of *CA15* located in chromosome 22. In chimpanzee two gene candidates were observed in chromosome 23. The chimpanzee gene candidate that would be syntenic with the human gene candidate 2 falls in a sequence gap of the chimpanzee chromosome 23 sequence. All of these gene candidates were observed to be pseudogenic, because they contained several frameshifts and point mutations.

In publication II the procedure for obtaining sequences was much more straight-forward. All available CA IV-like and CA XV protein sequences were retrieved from Ensembl starting from the Gene Tree for human *CA4* (ENSG00000167434). We selected a parent node at the root of a subtree containing all vertebrate *CA4* and *CA15* orthologues and a group of four Ciona homologues, a total of 123 genes. An alignment of the corresponding proteins was created, and low-quality sequences were eliminated in several steps. First, all sequences with X characters to indicate missing regions, from supercontigs, were removed. Next, sequences with gaps of more than 20 residues within mainly gapless regions were removed. Then, sequences with insertions of more than 20 residues between conserved blocks were removed. After these steps, the alignment was remade with ClustalW and, finally, sequences with gaps longer than 15 residues within conserved blocks were removed. Of the original 123 proteins this final selection contained 75 sequences. Four additional protein sequences from phylogenetically interesting and less represented groups were retrieved from the NCBI protein database.

In publication III Drosophila CA protein sequences were retrieved from FlyBase by EC number search (for 4.2.1.1). Additionally, CG10899-PB was deemed to be an incorrect gene prediction, and therefore discarded, but CG10899-PA was retained. GC1402 contained regions which aligned poorly with its orthologues, and was replaced by ABC86467.1 from GenBank, which aligns well over its entire length.

These collected Drosophila sequences were used as queries to locate orthologues from other Drosophila and nematode proteomes using NCBI BLAST against assembled genomes. Altogether 12 proteomes from the genus Drosophila were used to avoid missing orthologues and paralogues. We acknowledge that we may still have overlooked some CA genes due to missing or incorrect annotation or incomplete coverage in the genomes.

Multiple sequence alignments were made to detect incomplete sequences or poorly aligning sequence regions. Further searches were made in genomes at Ensembl to verify gene predictions, and in some cases, to retrieve more complete protein sequences. Short fragments and sequences with excessive length were rejected as likely to be improperly annotated. The final set of sequences consisted of 140 alpha CA and 23 CARP sequences.

4.1.2 Sequence analyses (I, II, III)

The prediction of N-glycosylation sites for mouse CA XV were either performed using the NetNGlyc 1.0 Server with default parameters (http://www.cbs.dtu.dk/services/NetNGlyc/) (I, II), or by finding the simple sequence pattern Asn-not Pro-Ser/Thr (II).

For prediction of subcellular localization, we used the following programs in the Internet. SignalP 3.0 for secretory signal peptides in eukaryotes (Emanuelsson *et al.* 2007), with the HMM algorithm; TargetP 1.1 for secretory signal peptides and mitochondrial targeting peptides in eukaryotes (Emanuelsson *et al.* 2007), using the thresholds for 95% specificity; and PredGPI for glycosyl phosphatidylinositol (GPI) anchor prediction (Pierleoni, Martelli & Casadio 2008) using the general model. For decisions within groups of orthologues in III, we checked the agreement between SignalP and TargetP so that ‘cytoplasmic’ was assigned when SignalP prediction of non-secretory protein coincided with the prediction ‘other’ (non-secreted, non-mitochondrial) in TargetP; ‘agreed signal peptide’ and ‘agreed signal anchor’ were assigned when TargetP prediction was secreted, and SignalP prediction was ‘signal peptide’ or ‘signal anchor’, respectively. In cases where the threshold scores in TargetP were not met, the prediction of signal peptide vs. signal anchor vs. cytoplasmic was left blank in Fig. 1. GPI anchor prediction was assigned when the PredGPI specificity score was at least 99% (corresponding to all predictions ranked as positive by the program).

4.1.3 Sequence alignment (I, II, III)

Multiple sequence alignments for phylogenetic analyses, figures and various intermediate inspection steps were performed with ClustalW (Larkin *et al.* 2007) (I, III), T-Coffee version 2.11 (Notredame *et al.* 2000) (I), or Clustal Omega (Sievers *et al.* 2011) (II).

4.1.4 Phylogenetic analyses (I, II, III)

In article I, phylogenetic analysis of protein alignments was carried out with PAUP* 4.0 [Swofford, DL (2002) PAUP*, Phylogenetic Analysis Using Parsimony (* and Other Methods), version 4. Sinauer Associates, Sunderland, MA, USA]. The majority rule consensus tree was obtained by three bootstrap runs with different random seeds. The dataset was bootstrapped 1000 times for each run.

In article II we used the PAL2NAL web server (Suyama, Torrents & Bork 2006) to produce codon-aligned cDNA sequences from protein alignments. This alignment was used in MrBayes v 3.2 (Ronquist *et al.* 2012) to estimate the phylogeny of the sequences by Bayesian inference. A 50% majority rule consensus tree was created and visualized using the APE R package (Paradis, Claude & Strimmer 2004).

In article III, the number of sequences was too large for a full phylogenetic analysis of all members of all CA orthologue groups. Our previous experiences with these genes indicated that the orthologue groups are well separated as we have never seen sequences from different orthologue groups to be mixed on phylogenetic trees. We resolved the relationship of the orthologue groups using the *D. melanogaster* sequences from all 16 orthologue groups. We produced a bootstrapped minimum evolution tree (1000 replicates). The CARP sequences were used as an outgroup to root the tree. Then, orthologue groups were used pairwise in a phylogenetic analysis using the same method to identify the fine structure of the individual orthologue groups. These trees were rooted between the used two groups. Finally, the trees were merged using the Archaeopteryx program package.

4.1.5 Protein models and visualization (I, II)

Visualizations of three-dimensional protein structures were made with the help of InsightII (Accelrys Inc., San Diego, CA) (I); and with PyMol Molecular Graphics System, Version 1.51, Schrödinger, LLC (II).

The structural prediction for mouse CA XV (I) was based on human CA IV at 2.8 Å resolution, PDB 1ZNC (Stams *et al.* 1996). In structural modeling amino acids 25-305 from the mouse CA XV were included. The model was constructed with software InsightII (Accelrys Inc., San Diego, CA). Amino acid substitutions were built using a side chain rotamer library. The initial model was refined with DiscoverTM in a stepwise manner by energy minimization using the AmberTM force field. The newly built loops were refined with 500 steps of minimization with a fixed and a free backbone, respectively. After that, all side chains with a constrained backbone were minimized for 500 steps, followed by another 1000 steps of minimization for the whole model.

4.1.6 Glycosylation footprinting (II)

An analysis of glycosylated and non-glycosylated regions was performed in all good-quality protein sequences of GPI-linked CAs. The 75 Ensembl protein sequences were aligned, and glycosylation sites, except sites with a sterically hindering Proline in the second position, were identified in the alignment and mapped to residue numbers in human CA IV structure, chain B of PDB 3FW3 (Vernier *et al.* 2010). In case of alignment matches to a gap in human CA IV, we mapped the site to the closest residue on the edge of the gap. The total number of sequences with a predicted glycosylation was tabulated for each position. A 3FW3-based glycosylation model was generated for CA IV, XV, and XVII protein groups (as defined by our phylogenetics results) individually, as well as for a composite of all three. Positions in CA IV matching these sites were coloured by the frequency of glycosylation sites in the corresponding position in each collection of sequences. Side chain surface exposure and secondary structure environment of each position were evaluated.

4.2 Laboratory methods

4.2.1 Polymerase chain reaction (PCR) method, sequencing of the PCR products, and *in situ* hybridization (I)

Reverse transcription polymerase chain reaction (RT-PCR) method was used to reveal murine and human tissues that would express *CA15* mRNA with commercial cDNA kits. The mouse MTC™ panel I contained first strand cDNA preparations produced from total poly A RNAs isolated from 12 different murine tissues. In addition, mRNA was isolated from 6 mouse tissues absent in the panel (stomach, duodenum, jejunum, ileum, colon and blood). Reverse transcription was performed with Mo-MuLV reverse transcriptase and random primers (500 µg/ml). The procedures were conducted according to the principles of the Declaration of Helsinki and approved by the institutional animal care committee (University of Tampere). The human MTC™ Panels I and II were used to study the expression of *CA15* mRNA in 15 human tissues.

To study mRNA in mouse, sequence specific primers for murine *CA15* were designed using the information in GenBank (NM_030558). The forward primer (MF1) was 5'-TACCTGGTGCTACGACTC-3' (nucleotides 148-165) and the reverse primer (MR1) 5'-TATCGGTAGTACCGCAAG-3' (nucleotides 739-756). The resulting amplification product size was 609 bp.

Sequence analyses revealed that the human genome contains three copies of *CA15* genes which have most likely become pseudogenes. In order to get experimental data whether any of these candidate genes are expressed, primers were designed for each of them, and additionally, one primer pair was designed to recognize all of them. Sequence information for human primers was derived from Ensembl Genome Browser. The primers used here and the corresponding PCR conditions were as described in Supplementary Data of article I at http://www.biochemj.org/bj/392/bj3920083add.pdf.

In situ hybridization was performed for mouse tissues as described previously (Heikinheimo, Scandrett & Wilson 1994).

4.2.2 Expression of CA XV in COS-7 cells and carbonic anhydrase assay (I)

The full-length mouse *CA15* and, as a control, mouse *CA4* cDNA were used for transient transfection of COS-7 cells using DEAE-dextran procedure. The CA activity from cell lysates or membrane suspension was determined in duplicate by the procedure of Maren as described (Sundaram *et al.* 1986). The CA activity was expressed as U/mg cell protein or membrane protein.

4.2.3 Expression and Purification of Recombinant CA XV produced in *E. coli* (I)

The fragment of *CA15* was amplified by PCR from a full-length *CA15*, and NdeI and BamHI sites were introduced. The amplified product encoding amino acids 22-292 of the full length native sequence was cloned into a vector and sequenced. The DNA fragment, isolated by restriction enzyme digestion, was cloned into the bacterial expression vector. *E. coli* host cells were transformed with the vector and grown with induction of expression of mouse CA XV. The protein was isolated and refolded by glutathione as described for CA IV (Waheed *et al.* 1997b). CA XV was purified with a p-aminomethylbenzene sulfonamide Affigel column and further purified over Sephacryl S-300 sizing column.

4.2.4 Post-translational modifications (I)

Mouse CA XV was deglycosylated to study the amount of sugar chains in it, and treated with PI-PLC to verify GPI anchoring. Both experiments were performed in Bill Sly's laboratory in St. Louis, MO, USA.

COS-7 cell membranes were treated with Endoglycosidase H as described (Waheed *et al.* 1997a), or treated with PI-PLC (Waheed *et al.* 1992). For control, the membranes were treated with buffer alone. Both deglycosylation as well as GPI-release products were analysed by SDS-PAGE followed by Western blot.

5. Results

5.1 Primates have pseudogenes for *CA15* (I, II)

We discovered several copies of a sequence which resembles *CA15*: three in the human genome, and two in the chimpanzee genome. All of these copies contain six serious defects, any of which would be sufficient for inactivation of the gene product, so the inactivation must have preceded the duplications (I). Figure 4 (from I) shows the locations of these defects.

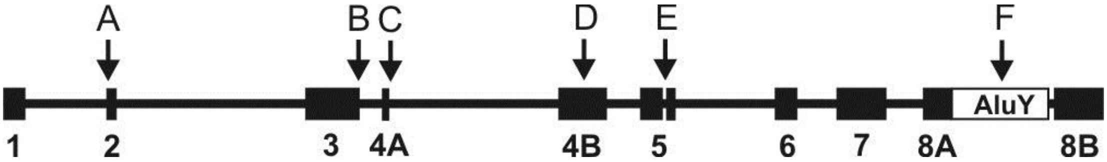

Figure 4 **Organization of human and chimpanzee CA XV pseudogenes.** Numbered boxes show reconstructed exons. Exon and intron lengths are in scale. Arrows point out the defects that are common to all five pseudogene copies. A and B: frameshifts. C: The beginning of the intron after exon 4A has GA instead of the conserved splicing dinucleotide GT. D: A 9-bp insertion predicted to disrupt the active site. E: A 4-bp insertion in exon 5, leading to a frameshifted sequence in a region which is highly conserved in all CAs. F: insertion of an AluY repeat sequence which splits exon 8, duplicating 17 bp of the exon sequence (duplicated part seen as a gap after AluY).

With all this damage accumulated in the genes, there is no chance that human or chimpanzee would have an active CA XV protein. In order to investigate whether the pseudogene sequences would be transcriptionally active, we constructed primer pairs for all three human pseudogenes. Their expression was tested by RT-PCR in 15 different tissues, with negative results in each case. In addition, there were no human transcripts in GenBank for these regions in 2005 (I), and there were none in May 2013 either (Tolvanen MEE, unpublished observation).

In the genomes of gorilla and orangutan we can find single copies of *CA15*-like sequences. In the gorilla genome there was no annotation of a gene, but we localized a potential coding region in chromosome 22. The predicted coding sequence would code for a 263-residue fragment, missing the signal peptide and an estimated 40 residues in the C terminus. The orangutan protein ENSPPYP00000009564 (Ensembl) is likewise devoid of the signal peptide, but with a full C terminus. Both of these predicted proteins have regions which match poorly with mouse CA XV, but match well with the reconstructed human pseudogene sequences (of I) and most importantly, both have a substitution of H to N in the second zinc-binding histidine position of the active site, as in seen in human and chimpanzee *CA15* pseudogenes. So the orthologues of *CA15* in gorilla and orangutan cannot code for an enzymatically active CA either. In comparison to the pseudogenes in human or chimpanzee, these sequences are considerably less disrupted, in that there are no frameshifts or intervening stop codons in the coding sequences, and no Alu element inserted. The H-to-N variation was likewise seen in the fragmentary *CA15* matches we found in the gibbon genome.

In marmoset (*Callithrix jacchus*) we found an unannotated and fragmentary *CA15* sequence in chromosome 1 (183,443,677-183,445,044). It contained sequences corresponding to the second exon (possibly with a remnant of signal sequence), third exon, more than half of the fourth exon, and start of the fifth exon. The sequence of the fourth exon still contains the first two His residues, without the H-to-N variation in the second histidine that we have seen in all other primates. The presence of the 9-bp insertion in exon 4 could not be ascertained because the sequence matching *CA15* is truncated just at that site. The sequence lacks substantial parts, so if we assume that the genome assembly is correct, marmoset *CA15* is a pseudogene as well.

5.2 Genomics and phylogenetics of *CA4*-like genes

5.2.1 Origin of GPI-anchored CAs goes back to fungi

GPI anchors are as ancient as the earliest eukaryotic cells (Orlean, Menon 2007). In case of CAs, GPI linkage has not been discussed in literature for anything more primitive than insects (Seron, Hill & Linser 2004, Ortutay *et al.* 2010). However, our studies have found very clearly predicted GPI attachment sites in alpha CAs of nematodes (*Caenorhabditis elegans* cah-5, NP_509186) and even in fungi (*Paracoccidioides brasiliensis* alpha CA, ACA28690.1). Thus far, we have not discovered any plant alpha CAs which would be predicted to be GPI-anchored, so it is reasonable to believe that the first GPI-linked carbonic anhydrases were formed early in the Fungi-Metazoa lineage.

5.2.2 Phylogenetic study reveals a novel group of isozymes

Our phylogenetic analysis, as presented in the tree of Fig. 1, indicates that the gene duplication which led into *CA4* and *CA15* groups of genes happened in very early vertebrates, before the radiation of jawed vertebrates. The *CA4*-like genes of two Agnathostomata species, lamprey (*Petromyzon marinus*) and hagfish (*Eptatretus stoutii*), are not associated with any of the branches in the Gnathostomata part of the tree. Slightly later, but still before the split of the tetrapod and fish lineages, *CA4* was duplicated to give rise to an additional isozyme gene, which we call *CA17*. Either one of these duplication events may have been related to whole-genome duplications in early vertebrate evolution. The evolutionary history, reflected in the branching pattern of the tree, shown schematically in Fig. 5, and the fact that both isozymes coexist in many species in distinct chromosomal locations are two pieces of evidence that clearly show that *CA17* genes code for a novel isozyme, a paralogue of *CA4*.

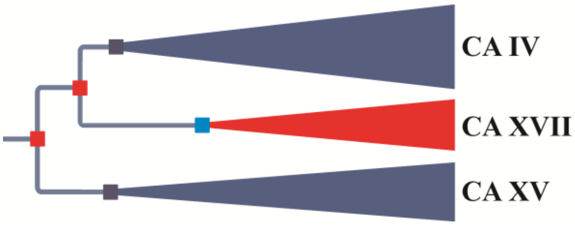

Figure 5 Schematic presentation of the relationships of extant GPI-linked CAs in vertebrates

Descendants of both *CA4* and *CA17* are found in fishes, in non-mammalian tetrapods and in the coelacanth, Latimeria, whereas therian mammals lack *CA17*. The most plausible hypothesis is that *CA17* was lost in mammals but retained in fish and non-mammalian tetrapods. During the evolution of ray-finned fishes, both *CA4* and *CA17* have undergone multiple further duplications. *CA4* is seen as tripled in fish genomes.

To visualize the duplications within the fish *CA17* group we performed a separate phylogenetic analysis with all available sequences, including the lower-quality ones. In the resulting tree (Fig. 2 in II) we observe an early duplication at the root and a series of recent duplications in the upper branch. We gave the name *CA17*A to the genes in the lower branch, which have remained single copies in all known fish genomes. In the zfin online database (http://zfin.org) there is no gene assigned to the corresponding region. The upper branch demonstrates parallel but independent multiplication events in several fish genera, resulting in up to six *CA17* genes per fish species. The genes in the upper branch are labeled starting from *CA17*B in each species. The order of the lettering is essentially random, but based on the order seen in earlier trees we made. Because only one of the earlier *CA17* copies has been multiplied further, it would be interesting to study if there are specific genome elements which would have facilitated the duplications. It is also of significance to note that the multiple copies of *CA17* are seen clustered in single chromosomal regions, and the clusters include *CA17*A as well as a variable number of the genes shown in the other branch of the tree which show genus-specific duplications. These relatively recent events suggest that the CA gene family is still evolving rapidly in fishes.

All functional *CA15* orthologues are found as single copies in each genome, as shown in Fig. 1 and in the Ensembl GeneTree. They should have the name *CA15* in each species (with the exception of historical rodent naming, *car15*). This includes the zebrafish gene labeled as *car15* in Ensembl (ENSDARG00000060829), LOC568143 in NCBI (protein NP_001038604.1), and *ca16a* in zfin.org and in a previous publication (Lin *et al.* 2008). This "*ca16a*" is undoubtedly a true one-to-one orthologue to all other known *CA15* genes, including the original mouse *car15*. The assignment as *CA15* is supported by the Ensembl orthologue tables and also by the syntenic location in a chromosome region which contains the genes DGCR2 and DGCR14 in the proximity of *CA15*. This sequence of zebrafish genes (located on chromosome 10: 43,394,645-43,404,294) is syntenic with the neighbourhood of *CA15* in all mammals. Therefore, the gene name *ca16a* should be rejected and changed into *ca15*.

There are two other "ca16" genes in zebrafish gene annotation, *ca16b* and *ca16c*, and these are clearly incorrect names as well. The gene products are in fact protein tyrosine phosphatase receptors (PTPRs) which have a CA-like domain as a minor part in their sequence. The gene *ca16b* is an orthologue to PTPRG, and *ca16c* is an orthologue to PTPRZ. The CA-related protein domains in PTPRs are not catalytically active CAs, and therefore these genes should not be called carbonic anhydrases, but instead protein tyrosine phosphatase receptors. In zfin.org *ptprgb* refers to D2U7Y2_DANRE in UniProt, a 270-residue fragment which is a perfect match with 248 consecutive residues, of 1382 amino acid residues, of the “ca16b” gene product F1QWY5_DANRE. The protein D2U7Y2_DANRE is also mapped to the *ca16b* gene locus in Ensembl. Therefore, *ptprgb* would be appropriate instead of *ca16b*. For *ca16c* we would suggest the name *ptprz*, because no other protein tyrosine phosphatase receptor zeta is yet assigned in zfin.org.

However, because these two PTPR proteins contain CA-related protein domains, and we would like to disrupt previous nomenclature as little as possible, we suggest retention of CARP XVI as a descriptor for CA-related domains found in PTPRs. We have chosen the next available number *CA17*/CA XVII for the newly described group of isozymes.

In the case of *CA4* isozymes in zebrafish, our trees retain the identities of ca4a, ca4b, and ca4c as defined in zfin.org, and these would be the names we suggest for their orthologues in other shown fish species too (except for *T. rubripes,* in which there are two copies, labeled *CA4*A1 and *CA4*A2 here).

The following Tables 1 and 2 summarize the observations of publications I and II on existence of CA isozymes in different vertebrate taxons, and they also contain unpublished information I have collected from databases.

Table 1. CA isoforms in tetrapods. NE = non-existent, PTX = pentraxin domain, PG = proteoglycan domain, PTPR = protein tyrosine phosphatase receptor, chr = chromosome. The bottom row indicates numbers of PTPRG and PTPRZ orthologs separated by a plus sign. Empty squares indicate existence of a single orthologue.

	Tetrapods						
	Mammals				**Birds**	**Reptiles**	**Amphibians**
Isoform	**Primates**	**Other placental**	**Marsupials**	**Monotremes**			
I				bad prediction/ pseudogene?	Lost	Lost?	
II			2x separate chr				
III					2x same chr	1x	2x same chr, independent of birds?
XIII							
CAHZ	NE	NE	NE	NE	NE	NE	NE
VII						?	
V	2x	2x	2x	2x			
CARP VIII							
VI				+PTX (?)	+PTX	+PTX	+PTX
IX	+PG	+PG	+PG				
XII							
XIV							
IV				?			
XV	Pseudogene						
XVII	Lost	Lost	Lost	?			
CARP X							
CARP XI							
CARP XVI domain in PTPRs	1+1	1+1	1+1	1+1	1+1	1+1	1+1

Table 2. CA isoforms in fishes and the subcellular localizations of CA isoforms in vertebrates. NE = non-existent, CAHZ = CA II-like cytoplasmic CA in fishes, PTX = pentraxin domain, TM = transmembrane, GPI = glycosylphosphatidylinositol, PTPR = protein tyrosine phosphatase recector, PTPRG = protein tyrosine phosphatase recector gamma, PTPR = protein tyrosine phosphatase recector zeta. The bottom row indicates numbers of PTPRG and PTPRZ orthologs separated by a plus sign. Empty squares indicate existence of a single orthologue.

	Fishes				
	Lobe-finned	**Ray-finned**	**Cartilaginous**	**Jawless**	
Isoform					**Location**
I	only one?	NE	*Galeocerdo*, most similar to CAHZs	2x of cytoplasmic type	**Cytoplasmic**
II		NE			**Cytoplasmic**
III		NE			**Cytoplasmic**
XIII		NE			**Cytoplasmic**
CAHZ	not seen	2x			**Cytoplasmic**
VII			?		**Cytoplasmic**
V	1x ?		?	?	**Mitochondrial**
CARP VIII			?		**Cytoplasmic**
VI	+PTX	+PTX	?	?	**Secreted/lectin-anchored?**
IX		Missing in most	?	?	**TM**
XII			?	?	**TM**
XIV			?	?	**TM**
IV		3x to 4x	*Squalus*	CA-IV-like 1x to 2x, before IV-XV split	**GPI-anchored**
XV			?		**GPI-anchored**
XVII		2x to 6x	?		**GPI-anchored**
CARP X		2x	?	3x, before X-XI split	**Secreted**
CARP XI		Missing	?		**Secreted**
CARP XVI domain in PTPRs	1+1	2+2?	?	At least 1, before PTPRG-PTPRZ split	**In PTPRG and PTPRZ, TM**

5.3 Glycosylation footprinting reveals potential interaction surface regions

We mapped the N-glycosylation sites of all available CA IV, CA XV, and CA XVII isozyme sequences separately and jointly to get four molecular models of glycosylation site conservation on protein surface, projected on a CA IV structure as seen in Figure 6.

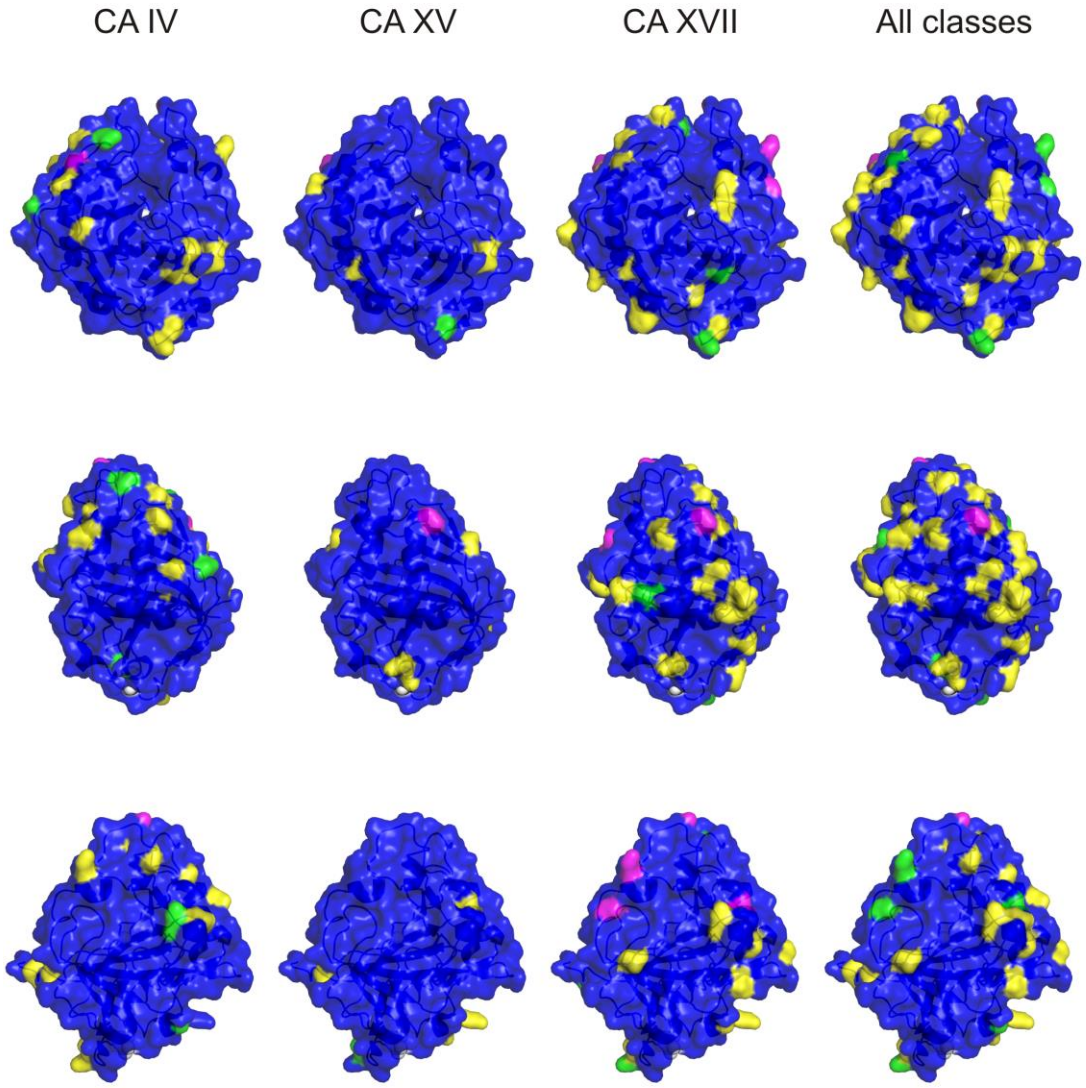

Figure 6 N-glycosylation sites were predicted in 75 protein sequences. Positions of sites were mapped onto a structure of human CA IV (chain B of PDB 3FW3) for the three GPI-linked CA classes (CA IV, CA XV, and CA XVII) and their composite. Within each CA class the total number of sequences with glycosylation predicted at each position was tabulated. For each CA class the range of scores was divided into thirds and corresponding residues in the 3FW3 model were coloured accordingly: top third, magenta; middle third, green; and lower third, yellow. The Zn^{2+} in the active site and the C-terminal residue are indicated in white spheres. Three views are shown for each coloured model. The C-terminus, representing the GPI anchor site, is oriented downwards in each view. Top row, a direct view into the active site; middle row, first view rotated 120 degrees vertically to the right; and bottom row, first view rotated 120 degrees vertically to the left. Visualized with PyMol.

Figure 6 shows these four models in three orientations. The main finding is that the glycosylation sites seem to be enriched on the side shown in the view of the middle row of images, indicating that the protein surface on this side is not needed for protein-protein interactions. In contrast, the region near the active site opening, seen in the top row of images, mainly shows absence of glycosylation. This is to be expected, because sugar chains near the opening might block access to the active site, but the glycosylation-free region extends slightly farther than just the active-site cavity. Another area which lacks glycosylation sites can be observed in the bottom row of images, in the lower (membrane-proximal) region, corresponding to loops centred at residues Thr-36 and Asp-110 (in CA II numbering used in 3FW3).

5.4 Alpha CA and CARP sequences in 12 Drosophila genomes (III)

Based on the orthologue groups established in previous work (Ortutay *et al.* 2010) we investigated whether all alpha carbonic anhydrase genes are present in all 12 available Drosophila genomes. Figure 7 shows the species from which full genomes have been sequenced and their phylogenetic relationships.

The 15 orthologue groups analysed occur throughout the phylogenetic tree of the investigated genomes. The CAH2, 7, 8, and 9 orthologues appear in all the genomes, while all others, with the exception of CAH16, are missing in up to three genomes. We could not detect any apparent trend or pattern regarding which gene is missing in any branch, except the CAH16 group, which is exclusive to the melanogaster subgroup (refer to Figure 7).

Only the *D. melanogaster* genome was found to contain all 15 orthologues, while the majority of other genomes lack one to four CA or CARP orthologues. The exception here is the genome of *D. willistoni*, in which CAH4, 6, 13, 14, 15, and CAH16 orthologues are missing. In the *D. grimshawii* genome duplications have resulted in seven tandem copies of CAH5 paralogues, and *D. mojavensis* has a partially duplicated copy of CAH5.

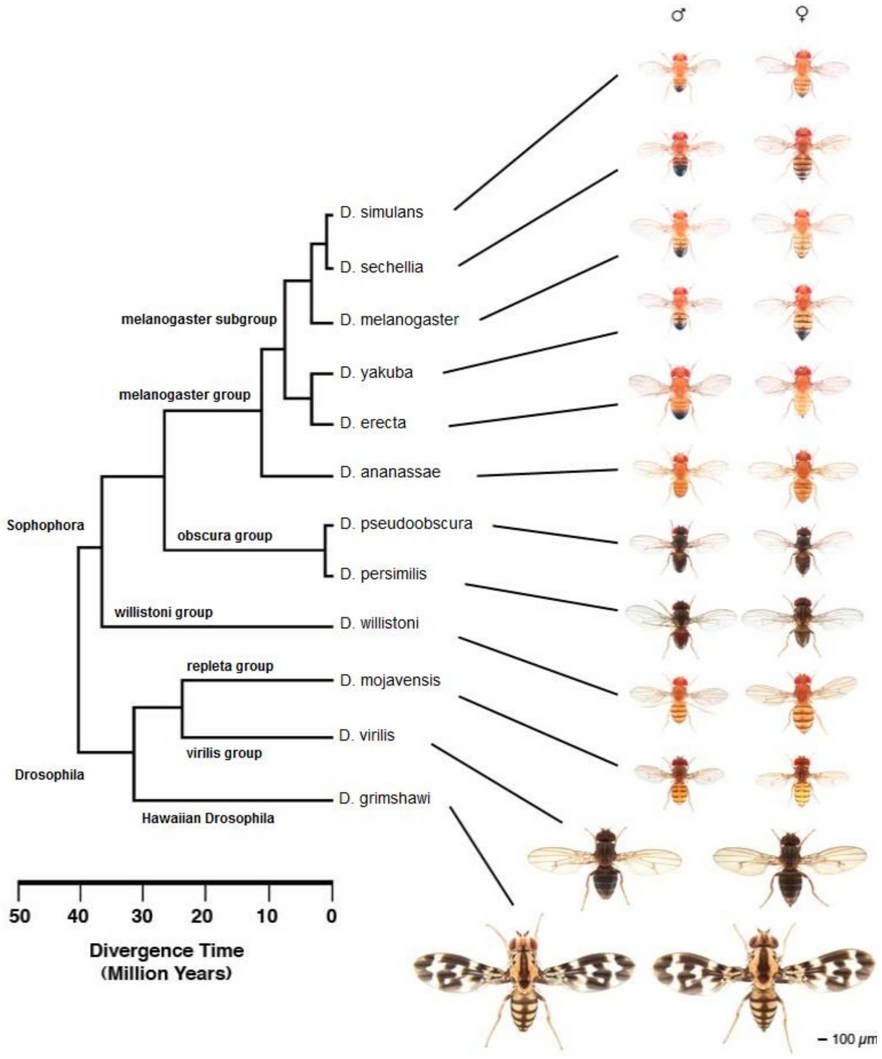

Figure 7 Phylogenetic relationships between 12 Drosophila species of the multi-species genome project. Tree image from flybase.org, fly images provided by Nicolas Gumpel.

5.5 Phylogenetics of Drosophila CAs

Orthologues with similar functional features co-cluster in the phylogenetic tree. The CAH and CARP sequences are separate from each other in all analyses, regardless of the exact sequence selection. Among the CAH groups, the seemingly acatalytic CAH13, 14, and 15 groups of CAs (without orthologues in vertebrates), are also always separate.

5.6 Comparative analysis of functional data

The functional data correlate well with the phylogeny of the groups. The genes are expressed at different time points during development and with variation in the tissue expression patterns. CAH1 and CAH3 are predicted to be intracellular. They are expressed in all larval and adult tissues except for the reproductive organs and S2 cells (during the larval, metamorphosis, and adult stages).

Based on sequence analysis, CAH2 and CAH7 would be targeted to the secretory pathway and contain a GPI anchor attachment recognition sequence for retention on the plasma membrane, similar to mammalian CA IV and CA XV [(Waheed *et al.* 1992), I] and non-mammalian vertebrate CA XVII (II). The extracellular localization of CAH2 and CAH7 has been confirmed experimentally in a large-scale mass spectroscopy study of glycopeptides derived from proteins present in *D. melanogaster* head (Baycin-Hizal *et al.* 2011). Glycosylated peptides derived from CAH2 and CAH7 were observed in this study, which proves that these proteins have been in contact with the secretory machinery, as the predicted signal peptide and GPI modification target sequences also indicate. These two proteins are expressed in a broad variety of tissues. The only difference between the two is that CAH7 is expressed in S2 cell line whereas CAH2 is absent. During development CAH2 is most highly expressed during the late embryonic stage, while CAH7 is expressed at relatively highest levels in the larval and metamorphosis stages.

It was not possible to obtain a consistent subcellular localization prediction for the CAH4 family members, and developmental gene expression data was missing for CAH4. It has a low expression level in most tissues, while somewhat elevated only in male reproductive organs. CAH4 orthologues were not found in the genome sequences of *D. grimshawi*, *D. persimilis*, and *D. willistoni*.

5.7 Analysis of substitution rates

Synonymous variations per synonymous sites (Ks) and non-synonymous variations per non-synonymous sites (Ka) were calculated together with the Z scores and Ka/Ks ratios for all possible species pairs in all orthologue groups. Altogether the calculations were done for 743 pairs (see Additional file 1).

The mean Ka/Ks ratio was highest in the CAH14 orthologue group (0.474, sd=0.231), and lowest in the CAH3 group (0.0191, sd=0.0115). The highest single Ka/Ks ratios were observed between the CAH8 orthologues of *D. persimilis* and *D. yakuba* (1.369) and between the CAH14 orthologues of *D. sechellia* and *D. virilis* (1.057).

Four correlations were calculated for all 91 gene pairs, those of: 1) synonymous substitution rates (Ks), 2) non-synonymous substitution rates (Ka), 3) the Ka/Ks ratios, and 4) Z scores. The CAH16 orthologue group was excluded from this analysis because it is represented in only three species and thus provided too few data points for a reliable correlation analysis.

Although different orthologue group pairs showed the best correlations within the four parameters above, some overall tendencies can be observed. For the following orthologue group pairs the correlation of synonymous substitution rates (Ks) is higher than 0.8: CAH14-CAH15, CAH1-CAH2, CAH1-CAH4, CAH2-CAH7, CAH1-CAH9, CAH1-CAH7, CAH4-CAH15, CAH7-CAH9, CAH13-CAH14, CAH4-CAH9, and CAH4-CARP-A. Correlation of non-synonymous substitution rates (Ka) is generally stronger than that of the Ks rates. Very strong positive correlation (over 0.9) appears for the pairs of groups: CAH2-CAH4, CAH6-CAH14, CAH2-CAH14, CAH2-CAH15, CAH9-CAH14, CAH13-CAH15, CAH4-CAH13, CAH4-CAH15, and CAH4-CAH5. The correlations of the Ka/Ks ratios were almost negligible. Only three pairs of orthologue groups have a correlation coefficient higher than 0.7: CAH7-CAH9, CAH13-CAH14, and CAH8-CARP-B. Correlation of the Z scores is higher than 0.7 for the following pairs: CAH2-CAH6, CAH3-CAH12, CAH1-CAH7, CAH1-CAH9, CAH7-CAH9, and CAH13-CAH14.

The most striking feature of the comparison of calculated results is the strong correlation of the CAH13-CAH14 orthologue gene pair. The CAH7-CAH9 pair has strong and significant correlation coefficients as well. By analysing the lists of strongly correlating orthologue pairs we formed three sets of orthologues, where the members within each set show strong correlation in one or more of the parameters.

The first set is made of the CAH1, CAH2, CAH7, and CAH9 orthologues. These genes have low tissue expression specificity, and they are expressed in all larval and adult tissues, except for certain specific organs like the Malphigian tubule and reproductive organs. The second set contains CAH4, CAH5, CAH6, CAH13, CAH14, and CAH15. These genes are specifically expressed in the larval fat body and male reproductive organs (including testis and male accessory glands). The third set is made of CAH8 and CARP-B, which are expressed in virgin and mated spermathecae.

Two orthologues show unique gene expression patterns. CAH3 is expressed in almost all larval and adult tissues, except for male reproductive organs. Because the tissue specificity is very similar to the first set, the CAH3 group could belong there, although our statistical analysis cannot confirm this. CARP-A is specific to the larval and adult central nervous system of the fruit fly. Since this is the only orthologue group which was found in central nervous system tissues, we think that it evolved its function early, and therefore does not show significant correlation with any of the other groups.

6. Discussion

6.1 *CA15* pseudogenes in primates (I, II)

There are only two obvious gene defects which are shared in the nearly full-length *CA15* pseudogenes of gorilla and orangutan and in all five copies of human and chimpanzee pseudogenes, namely the mutation of the second active-site histidine into asparagine and an insertion of 3 codons between the codons of the second and third active-site histidines, which we predict to disrupt the active site. The His-to-Asn variation is found in all Old World primate genomes, including human, but not in the genome of marmoset, a New World monkey, which has a histidine in this position. We cannot determine if marmoset has the 9-bp insertion because the exon sequence is cut short before that position, but by way of exclusion we could hypothesize that the 9-bp insertion within the active-site sequences may have been the original defect that inactivated *CA15* before the separation of Old World and New World monkey lineages and led to its conversion into a pseudogene. The pseudogenes kept accumulating further defects in later species, and the His-to-Asn mutation is probably the second one in ancestry.

An alternative explanation would be that either the His-to-Asn mutation or the 9-bp insertion inactivated *CA15* in Old World monkeys, and the truncation of the *CA15* gene is an independently acquired defect in New World monkeys.

6.2 GPI-linked CA isozyme subfamily (II)

Phylogenetic analysis of all vertebrate *CA4*/*CA15*-like genes revealed an intricate pattern of duplications starting from a single *CA4*-like gene in early jawed vertebrates.

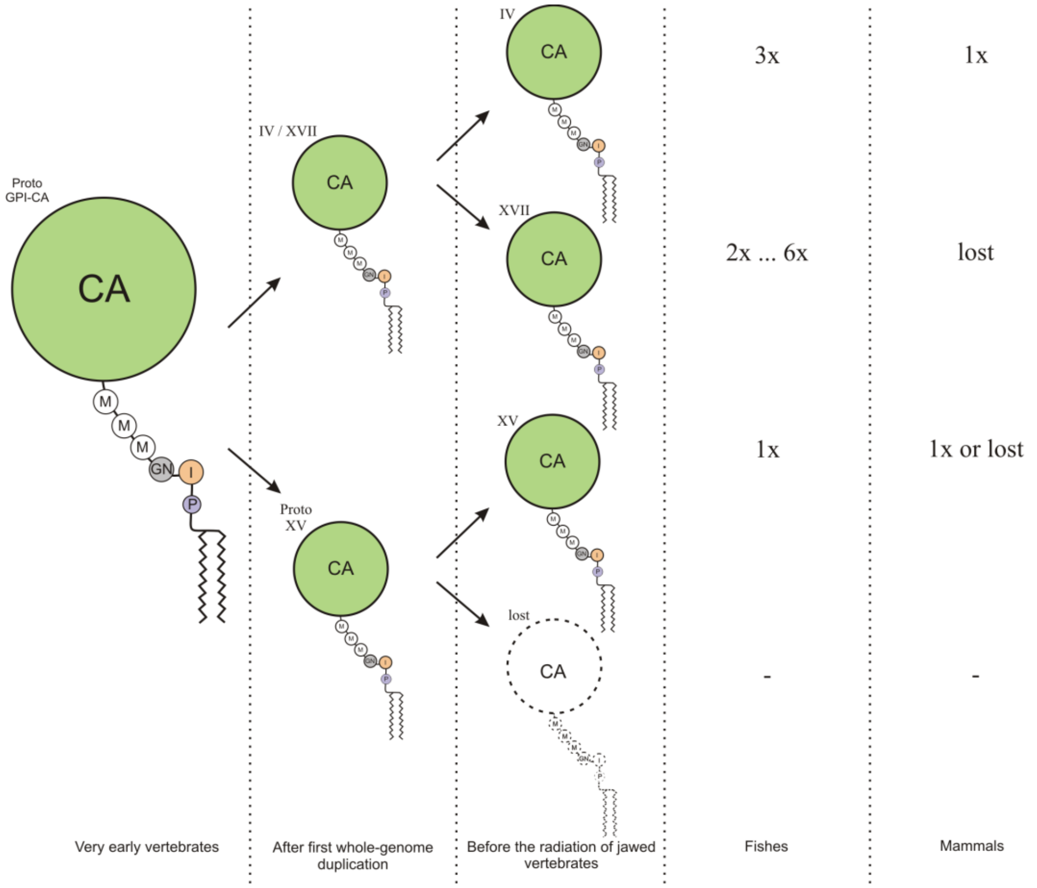

Figure 8 Presumed history of GPI-linked CAs. CA = CA catalytic domain. The structure underneath CA is the GPI anchor, with components M = mannose, GN = glucosamine, I = inositol, and P = phosphate. The first three columns represent evolution in early jawed vertebrates, before the divergence of fish and tetrapod lineages, and the last two columns show the gene copy numbers in each CA class in present-day fish and mammalian genomes.

The first early duplication produced the ancestor of *CA15* and another gene, ancestor of all other *CA4*-like sequences. Our phylogenetic analysis showed proof of yet another duplication before the separation of fish and tetrapod lineages within jawed vertebrates, which produced *CA4* and another cluster of isozymes which we named *CA17*. There have been extensive duplications of *CA17* in the lineage of ray-finned fishes, forming up to six *CA17* paralogues per species, and there are three or four *CA4* paralogues in each fish genome as well, so that the total number of descendants of the original *CA4*/*CA17* gene amounts to four to nine paralogues per each fish genome, possibly even more if unsequenced and/or unannotated paralogues exist. In contrast, no mammalian genome has more than a single *CA4*.

Presumably another copy was produced in the second whole-genome duplication of early vertebrates in the *CA15* branch of this tree, but we have no evidence of another *CA15*-like gene in any of the sequenced present-day genomes, so I assume this copy to have been lost, as indicated in Figure 8.

The branching pattern of the phylogenetic tree suggests an early duplication to *CA4* and *CA17*. However, *CA17* is missing from some lineages, including mammals and birds. With only three bird genomes, we cannot be sure if *CA17* is really lost in this lineage or just not observed because of incomplete sequencing and/or annotation. The absence of *CA17* from more than 30 therian mammalian genomes is a certain indication of the gene loss in this lineage, but there is uncertainty in the platypus genome, which shares many characteristics of bird and reptile genomes (Warren *et al.* 2008). There is a single *CA4*-like gene reported in the platypus genome, which in some analyses grouped with *CA17*, but it is a partial sequence which was excluded from the final analysis, so its assignment as *CA4* or *CA17* remains uncertain (Tolvanen MEE, unpublished observation).

The pattern of early gene duplications in the GPI-linked subfamily of CAs is similar to that of transmembrane-linked CAs. In the later evolutionary history we do not see further duplication as in the case of GPI-linked CAs, but instead, there are losses and additions of domains (Tolvanen MEE, unpublished data). Figure 9 summarizes the most likely sequence of events leading into present-day CA VI, CA IX, CA XII, and CA XIV based on our unpublished results.

The whole-genome duplications in early vertebrates have led to the expansion of the α-CA family in two of the subfamilies, as shown in Figure 8 and Figure 9. In contrast, only one copy of the acatalytic and highly conserved *CA8* has survived in the present-day genomes (Aspatwar *et al.* 2010), of the four assumed present in the last common ancestor of tetrapods and fishes. In case of mitochondrial and cytoplasmic isozymes conclusions of early evolutionary history have not been reached yet.

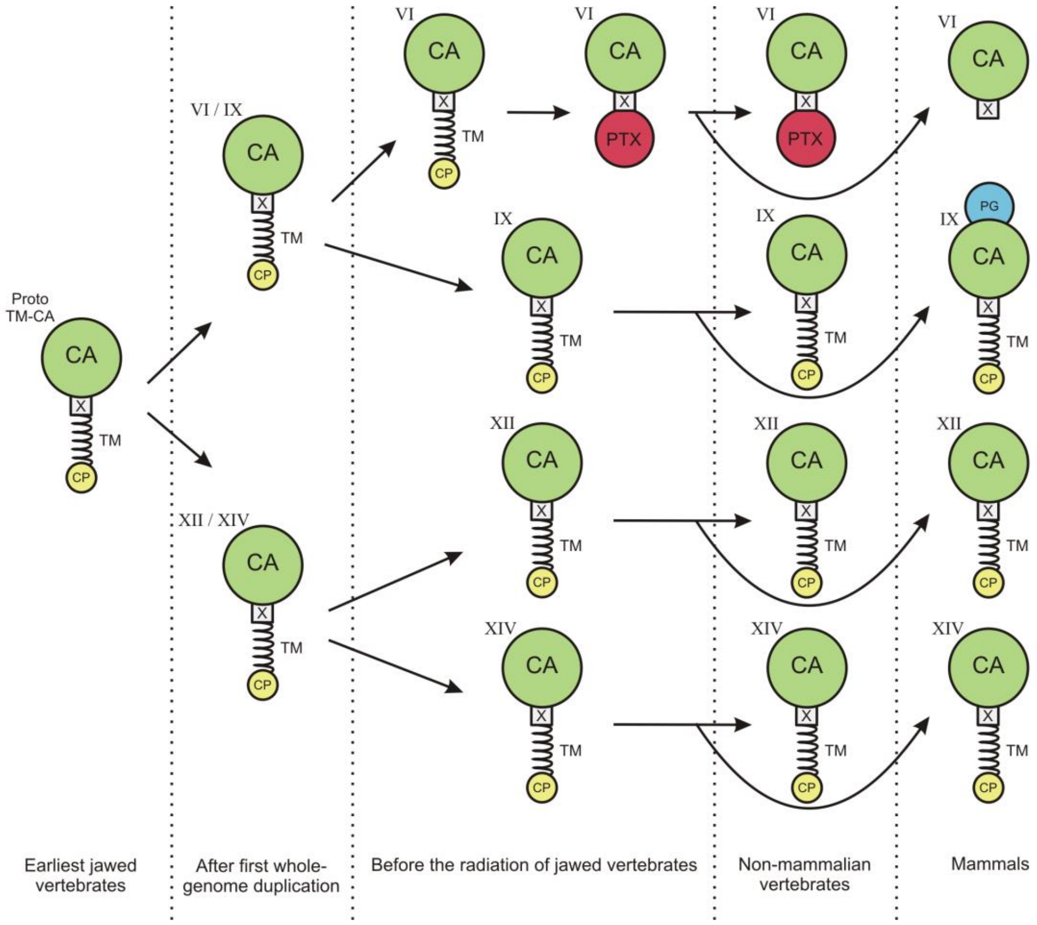
Proto
TM-CA
CA
X
TM
CP
VI / IX
XII / XIV
VI
IX
XII
XIV
PTX
PG
Earliest jawed vertebrates
After first whole-genome duplication
Before the radiation of jawed vertebrates
Non-mammalian vertebrates
Mammals

Figure 9 Presumed evolutionary history of transmembrane CAs. CA = catalytic CA domain, TM = transmembrane (domain), X = short spacer between the CA and TM domains, CP = cytoplasmic domain, PTX = pentraxin domain, PG = proteoglycan domain

6.3 Glycosylation footprinting (II)

Our experimental work on murine CA XV revealed that all three glycosylation sites are exposed on the surface of the protein model, and that all three are N-glycosylated (I). Inspired by this result, I decided to map the N-glycosylation sites of all available GPI-anchored CA protein sequences on 3D structures, with two goals in mind:

1) Functionally important glycosylation sites would show invariant positions.
2) Functionally important protein sites would never be covered by glycans.

We mapped the N-glycosylation sites of all available CA IV, CA XV, and CA XVII isozyme sequences separately and jointly to get four views of glycosylation site conservation on protein surface (Figure 6). In case of goal 1), none of the glycosylation sites proved to be strongly conserved in any of the isozymes, and even less so in the combined set of all. In contrast, goal 2) was achieved. We saw two potential functional regions marked by the absence of glycosylation in any of the species and in any of the GPI-linked isozymes. As expected, the active-site cavity and most of the region surrounding it on the protein surface are devoid of glycosylation. Another region that was always left without glycosylation was in two adjacent loops near positions Thr-36 and Asp-110 in the membrane-proximal half of the protein (bottom row of Figure 6). This area might be kept clear of glycosylation to provide an essential protein-protein interaction surface. This novel method of glycosylation 'footprinting' brings forth a working hypothesis that this region or part of the active-site-proximal region could contribute to the interactions between CA IV and ion transporters, as discussed in the review of the literature, and possibly similar interactions in CA XV and CA XVII isozymes.

6.4 Presence of CAH and CARP orthologues in the Drosophila genomes (III)

The majority of the 13 CAH, and two CARP, gene groups are present in all 12 investigated Drosophila genomes. Only the CAH16 group is specific to the melanogaster subgroup, whereas the other genes are found in all the genomes with few exceptions. In addition to truly missing genes, these exceptions may be due to breaks in genomic contigs or to missing annotations.

We investigated more closely the missed instances of CAH16, CAH5, and CAH6 genes. These three genes are adjacent in *D. melanogaster* on chromosome 3R, forward strand, between 27,083,000 and 27,090,000. The genes CAH5 and CAH6 form an adjacent pair outside the melanogaster subgroup. However, annotations are missing for CAH5 in *D. simulans*, and CAH6 in *D. erecta* and *D. willistoni*. Alignment of *D. melanogaster* CAH16, CAH5, and CAH6 protein sequences in the UCSC Genome Browser (Kent *et al.* 2002) by BLAT (Kent 2002) shows that CAH5 and CAH6 orthologues exist in 11 of the investigated Drosophila genomes, and additionally CAH16 orthologues in the five genomes of the melanogaster subgroup. For *D. willistoni*, the comparative genomics tools of EnsemblMetazoa (Kersey *et al.*

2010) confirm that there are orthologues of both CAH5 and CAH6 adjacent to one another. We conclude that CAH5 and CAH6 orthologues are missing as a result of incomplete annotation. Because there was no pattern in the missing orthologues, except for CAH16 which is limited to the melanogaster subgroup, we assume that most of the missing entries in the other CAH and CARP groups would be due to missing annotation.

Duplications of alpha carbonic anhydrases have been reported within various lineages, such as vertebrates (Aspatwar, Tolvanen & Parkkila 2010), Daphnia (Weber, Pirow 2009), and also in Drosophila and other insects (Ortutay *et al.* 2010). The current study also led to a minor new finding that the CAH5 gene is present as six complete and two partial copies in the genome of *D. grimshawii* as a mix of complete-seeming open reading frames and obvious pseudogenes which have in-frame stop codons. These CAH5 copies are located in tandem at the CAH5 locus, and they seem fairly recent. The support for the recent origin comes from the low copy number in the genomes of the closest sequenced species, *D. mojavensis* and *D. virilis*. In *D. mojavensis*, there are two adjacent copies of the CAH5 gene, but the region coding for first 70 residues is lost in the copy distal relative to CAH6, and therefore we predict it to be a non-functional pseudogene. *D. virilis* has only single copies of CAH5 and CAH6.

6.5 Correlation of substitution rates and gene functions

The method of phylogenetically independent contrasts is based on statistics inferred from phylogenetics (Felsenstein 1985). These statistics, called contrasts, are derived from changes of traits, which allow testing to see if two traits have been connected during evolution. The calculated statistical measures can be tested to see if their distributions on the phylogenetic tree are correlated. Such correlations are frequently associated with a functional connection of the investigated traits. We designed our statistical analysis based on this concept, using the novel idea that synonymous and non-synonymous substitution rate parameters could be used as trait data. Accordingly, we investigated whether any correlations would be found in four substitution rate-related parameters. The answer was yes, we found several strong correlations, and investigated further if the correlating gene groups have functional relationships. We were able to divide the orthologues into groups by networks of correlation relationships and discovered shared functional properties within the groups based on available expression data.

Since the correlation distributions are different for the four parameters, the thresholds for each correlation parameter were set as high as possible to observe only cases of high statistical significance.

The final result of this exercise was four lists of orthologue gene group pairs (Table 3), in which our analyses indicate that the evolutionary processes to show similar, well-correlating rates, and allow a tentative assignment of functional relationship. Some of the pairs appeared in more than one list. For example, the pair CAH13-CAH14 showed a strong correlation of their Ks, Ka/Ks, and Z score parameters. The gene pairs in the different lists often, but not always, showed good correlations for other parameters too, but just below the thresholds.

Table 3. List of orthologue group pairs showing strong and significant correlation of the investigated parameters

K_a	K_s	K_a/K_s	Z score
CAH2-CAH4 CAH6-CAH14 CAH2-CAH14 CAH2-CAH15 CAH9-CAH14 CAH13-CAH15 CAH4-CAH13 CAH4-CAH15 CAH4-CAH5 CAH4-CAH6	CAH14-CAH15 CAH1-CAH2 CAH1-CAH4 CAH2-CAH7 CAH1-CAH9 CAH1-CAH7 CAH4-CAH15 CAH7-CAH9 CAH13-CAH14 CAH4-CAH9 CAH4-CAH11	CAH7-CAH9 CAH13-CAH14 CAH8-CARP-B	CAH2-CAH6 CAH3-CAH12 CAH1-CAH7 CAH1-CAH9 CAH7-CAH9 CAH13-CAH14

The genes can be clustered into three distinct groups. We found common features within these sets from the functional annotation of individual orthologues, which we derived from expression data of *D. melanogaster*. CAH1, CAH2, CAH7, and CAH9 genes have low tissue-specificity; they are expressed in all larval and adult tissues, except for Malpighian tubule and reproductive organs. CAH4, CAH5, CAH6, CAH13, CAH14, and CAH15 are expressed in the larval fat body and male reproductive organs. CAH8 and CARP-B are expressed in virgin and mated spermathecae.

Only two orthologue groups are left out from the three sets. The first one is CARP-A, which has a unique tissue expression in the central nervous system in both larva and adult, very distinct from all other orthologue groups. The other one is CAH3, which probably belongs to the low tissue-specificity set. Several lines of evidence suggest that CAH3 would belong to the CAH1 group, including: phylogenetic relationship and common subcellular localization with CAH1, and similar gene expression patterns (both in tissue distribution and developmental stages), but this functional similarity did not emerge from the contrast analysis.

The CAH16 isozyme was excluded from this analysis, because it is present only in the melanogaster subgroup. Despite this absence, we argue that the CAH16 group would be a member of the second set based on its expression, which is restricted to male reproductive organs. CAH16 is phylogenetically a close relative of the CAH6 group, as is shown in our phylogenetic tree (Fig. 2 in III). Both genes are highly expressed during a brief period in the larval stage. Furthermore, both are expressed predominantly in the male accessory glands. These three independent lines of evidence suggest their relatedness. CAH5, CAH6 and CAH16 genes are adjacent in their chromosomal locations in all available Drosophila genomes, suggesting that they have originated from recent duplication events, which provides an obvious explanation for their evolutionary and functional relatedness.

Our results show that the set of alpha CA and CARP paralogues in Drosophila is as functionally diverse and complex as the vertebrate system of alpha CAs. There are clear patterns of “division of labour” both in temporal and tissue expression in Drosophila, but these patterns and the whole set of isoforms are distinct from those of vertebrates. The only functional similarities between Drosophila and mammals coincide with the few observed clear phylogenetic relationships, namely the nearly ubiquitous, cytoplasmic expression of Drosophila CAH1 [and Anopheles AgCA9 (Smith, Vanekeris & Linser 2007)] and mammalian CA II, and the neural tissue expression of Drosophila CARP-B and vertebrate CARP X and XI.

7. Summary and conclusions

In this study I have analysed two subfamilies within metazoan α-CAs, namely the GPI-linked α-CAs in vertebrates and α-CAs in 12 fruit fly genomes.

Within the GPI-linked subfamily, we characterized biochemically the novel isoform, murine CA XV, to show that it is a glycosylated, GPI-anchored, membrane protein, similar to CA IV. The expression of *CA15* was seen in mouse kidney, testis, and brain, similar to human *CA4*.

Our studies confirm the pseudogene status of human *CA15*. First, no mRNA expression was observed in 15 human tissues by RT-PCR using primers for each of the three human pseudogenes, and second, no mRNAs or ESTs were present in sequence databases. Third, the three pseudogene copies of human *CA15* and two in chimpanzee were shown to contain six shared defects, each of which could deactivate *CA15* on its own. This also indicates that the loss of activity of *CA15* preceded its duplications in the chimpanzee/human lineage.

Further study of other primate genomes showed single *CA15* orthologues per genome, each one inactivated or fragmented in various ways. Their analysis suggested a probable sequence of initial events that lead to inactivation and pseudogenization of *CA15* in primates.

The analysis of expression and activities of the *CA4* and *CA15* gene products in human and mouse shows that the expression of *CA4* in human covers the combined expression patterns of *CA4* and *CA15* in mouse, and human CA IV has a higher activity than mouse CA IV. It is possible that the deeper underlying reason for the pseudogenization of primate *CA15* was that *CA4* was able to fulfill the physiological roles of both isoforms, which made *CA15* dispensable.

Phylogenetic analysis of all vertebrate *CA4*/*CA15*-like genes revealed an intricate pattern of duplications starting from a single *CA4*-like gene in early jawed vertebrates. The earliest duplication led to the formation of *CA15*, which seems to occur as a single isozyme in all vertebrate genomes except for primates. Another duplication event, still before the jawed vertebrate radiation, separated *CA4* and

CA17, the novel isozyme discovered in this study. Both *CA4* and *CA17* have been duplicated into multiple paralogues in the fish lineage, whereas mammals have lost *CA17* and maintain a single copy of *CA4*. The expression data available for the eight *CA4*/*CA17*-like CA genes in zebrafish, even if scanty, indicates a division of tasks which in human are the responsibility of the product of the single *CA4* gene. Given the differences of breathing and acid excretion physiology in fish and land-living vertebrates, as discussed in section 2.1 in the Review of the Literature, it is reasonable to guess that multiple GPI-linked CA isoforms in the gills and kidneys in fish might provide some advantages, such as improved adaptability to varying external conditions (such as varying oxygen partial pressure and pH in water).

The losses and acquisitions of genes in the GPI-linked subfamily of α-CAs demonstrate the plasticity of the CA enzyme system, how functionalities can be merged and split between different isozymes.

The set of α-CAs in Drosophila genomes is as large as in mammals, but quite different in its composition and evolution. The third article of this study tries to make sense of this system, to discover functional relationships. We used a method called phylogenetically independent contrasts and tested whether the distributions of the contrasts on the phylogenetic tree correlate, which would imply functional relatedness between correlating genes. Our novel idea was to use the mutation rate values Ka and Ks and related parameters as traits for calculating the contrasts, and to analyse the values in the context of whole groups of orthologues in multiple genomes instead of single genes. In addition, using data from multiple genomes lead into more reliable predictions of subcellular localization of each CA isoform.

We observed many strong correlations between different CA orthologue groups in the fruit flies, and the network of correlations allowed us to divide the family into smaller, functionally connected groups. We identified four distinct alpha CA sets with specific characteristics. One set is nearly ubiquitous, while other sets are specific for male reproductive organs, for the spermatheca, and for the central nervous system. The study works as a proof of concept for the mutation-rate-driven analysis of phylogenetic contrasts.

Acknowledgements

I am grateful to my supervisor, prof. Mauno Vihinen, for sharing his enthusiasm for bioinformatics and research and for recruiting me as a full-time bioinformatician more than 12 years ago. This was a real turning point in my career, and without Mauno I would not have taken up serious bioinformatics research. Prof. Seppo Parkkila, the leader of the CA research group, deserves a huge thank you for accepting me as an associate member in his group and making me feel really welcome. With Seppo's never-failing optimism and good humour there has never been a doubt about the success of our projects. Our lively speculations on the meaning of bioinformatic findings, and on nearly impossible project ideas, have been the main driving force to keep my studies going.

Among my coauthors, special thanks are due to Mika Hilvo. He was the person who got me involved in CA research in the first place, and we had great fun in the discovery of human *CA15* pseudogenes, not to speak of the excitement. Ashok Aspatwar has been an essential part in in almost all of the CA studies I have been involved with. Ashok, thanks for being my outsourced project management and motivation resource for all these years and a special friend in culinary pleasures. Csaba Ortutay has been an expert collaborator in phylogenetics and a great long-time friend. Harlan Barker is the person with whom I like to test the viability of new ideas and who also constantly provides more of the same. He has been pivotal in our studies of protein structures, and I am really grateful to Harlan for the language checking of many papers and this thesis. It has been a pleasure to collaborate with Maarit Patrikainen in this thesis and in other studies. Thomas Gorr brought the interesting diversity of CAs in Drosophila to our attention, which has led to four articles by now. Much of the initial work on fruit fly CAs was done by late Ayodeji Olatubosun, and Preethy Nair wrote ingenious scripts for their analysis. I shared my office for many years in the happy company of Bairong Shen, who also contributed in the CA XV paper. Many thanks to all of you: The work of this thesis would have been much inferior or not made at all without your efforts.

Joachim Deitmer and Tommi Nyrönen deserve my gratitude for expedient expert work in reviewing this thesis manuscript. Special thanks are due to Tommi for shooting down the boring title the book first had.

Leo, Peiwen, Heini, Marianne, Aulikki, Fatemeh, Reza, and Prajwol in the CA group, it has been a great pleasure to work together with you and share your company on and off duty.

My colleagues at IBT have been invaluable for their moral support and for extensive off-topic discussions that everyone needs to keep their sanity - especially Ari, Jarkko, Riitta, and Janette. Marjatta, Erika, Henna, and Kuku, thanks for your help during my graduation process.

Finally, there is no way I can be grateful enough for my nearest and dearest. Leena, Tuuli, Ville, and Tuomas – thank you for being there.

Tampere, August 2013
Martti Tolvanen

References

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. 2002, "Evidence of en bloc duplication in vertebrate genomes", *Nature genetics,* vol. 31, no. 1, pp. 100-105.

Aspatwar, A., Tolvanen, M.E., Ortutay, C. & Parkkila, S. 2010, "Carbonic anhydrase related protein VIII and its role in neurodegeneration and cancer", *Current pharmaceutical design,* vol. 16, no. 29, pp. 3264-3276.

Aspatwar, A., Tolvanen, M.E. & Parkkila, S. 2010, "Phylogeny and expression of carbonic anhydrase-related proteins.", *BMC Molecular Biology,* vol. 11, pp. 25.

Baycin-Hizal, D., Tian, Y., Akan, I., Jacobson, E., Clark, D., Chu, J., Palter, K., Zhang, H. & Betenbaugh, M.J. 2011, "GlycoFly: a database of Drosophila N-linked glycoproteins identified using SPEG--MS techniques", *Journal of proteome research,* vol. 10, no. 6, pp. 2777-2784.

Birney, E., Clamp, M. & Durbin, R. 2004, "GeneWise and Genomewise", *Genome research,* vol. 14, no. 5, pp. 988-995.

Breton, S. 2001, "The cellular physiology of carbonic anhydrases", *JOP : Journal of the pancreas,* vol. 2, no. 4 Suppl, pp. 159-164.

Cox, E.H., McLendon, G.L., Morel, F.M., Lane, T.W., Prince, R.C., Pickering, I.J. & George, G.N. 2000, "The active site structure of Thalassiosira weissflogii carbonic anhydrase 1", *Biochemistry,* vol. 39, no. 40, pp. 12128-12130.

Dehal, P. & Boore, J.L. 2005, "Two rounds of whole genome duplication in the ancestral vertebrate", *PLoS biology,* vol. 3, no. 10, pp. e314.

Dudoladova, M.V., Kupriyanova, E.V., Markelova, A.G., Sinetova, M.P., Allakhverdiev, S.I. & Pronina, N.A. 2007, "The thylakoid carbonic anhydrase associated with photosystem II is the component of inorganic carbon accumulating system in cells of halo- and alkaliphilic cyanobacterium Rhabdoderma lineare", *Biochimica et biophysica acta,* vol. 1767, no. 6, pp. 616-623.

Elleuche, S. & Poggeler, S. 2009, "Evolution of carbonic anhydrases in fungi", *Current genetics,* vol. 55, no. 2, pp. 211-222.

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. 2007, "Locating proteins in the cell using TargetP, SignalP and related tools", *Nature protocols,* vol. 2, no. 4, pp. 953-971.

Esbaugh, A.J. & Tufts, B.L. 2006, "The structure and function of carbonic anhydrase isozymes in the respiratory system of vertebrates", *Respiratory physiology & neurobiology,* vol. 154, no. 1-2, pp. 185-198.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X.M., Harrow, J., Herrero, J., Hubbard, T.J., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A. & Searle, S.M. 2012, "Ensembl 2012", *Nucleic acids research,* vol. 40, no. Database issue, pp. D84-90.

Gilmour, K.M. & Perry, S.F. 2009, "Carbonic anhydrase and acid-base regulation in fish", *The Journal of experimental biology,* vol. 212, no. Pt 11, pp. 1647-1661.

Giordano, M., Beardall, J. & Raven, J.A. 2005, "CO2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution", *Annual review of plant biology,* vol. 56, pp. 99-131.

Heikinheimo, M., Scandrett, J.M. & Wilson, D.B. 1994, "Localization of transcription factor GATA-4 to regions of the mouse embryo involved in cardiac development", *Developmental biology,* vol. 164, no. 2, pp. 361-373.

Hewett-Emmett, D. 2000, "Evolution and distribution of the carbonic anhydrase gene families", *EXS,* vol. (90), no. 90, pp. 29-76.

Hilvo, M., Innocenti, A., Monti, S.M., De Simone, G., Supuran, C.T. & Parkkila, S. 2008, "Recent advances in research on the most novel carbonic anhydrases, CA XIII and XV", *Current pharmaceutical design,* vol. 14, no. 7, pp. 672-678.

Hirota, J., Ando, H., Hamada, K. & Mikoshiba, K. 2003, "Carbonic anhydrase-related protein is a novel binding protein for inositol 1,4,5-trisphosphate receptor type 1", *The Biochemical journal,* vol. 372, no. Pt 2, pp. 435-441.

Iverson, T.M., Alber, B.E., Kisker, C., Ferry, J.G. & Rees, D.C. 2000, "A closer look at the active site of gamma-class carbonic anhydrases: high-resolution crystallographic studies of the carbonic anhydrase from Methanosarcina thermophila", *Biochemistry,* vol. 39, no. 31, pp. 9222-9231.

Kent, W.J. 2002, "BLAT--the BLAST-like alignment tool", *Genome research,* vol. 12, no. 4, pp. 656-664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. & Haussler, D. 2002, "The human genome browser at UCSC", *Genome research,* vol. 12, no. 6, pp. 996-1006.

Kersey, P.J., Lawson, D., Birney, E., Derwent, P.S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kahari, A., Kinsella, R.J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A.J. & Yates, A. 2010, "Ensembl Genomes: extending Ensembl across the taxonomic space", *Nucleic acids research,* vol. 38, no. Database issue, pp. D563-9.

Kimber, M.S. & Pai, E.F. 2000, "The active site architecture of Pisum sativum beta-carbonic anhydrase is a mirror image of that of alpha-carbonic anhydrases", *The EMBO journal,* vol. 19, no. 7, pp. 1407-1418.

Klier, M., Schuler, C., Halestrap, A.P., Sly, W.S., Deitmer, J.W. & Becker, H.M. 2011, "Transport activity of the high-affinity monocarboxylate transporter MCT2 is enhanced by extracellular carbonic anhydrase IV but not by intracellular carbonic anhydrase II", *The Journal of biological chemistry,* vol. 286, no. 31, pp. 27781-27791.

Kumar, V. & Kannan, K.K. 1994, "Enzyme-substrate interactions. Structure of human carbonic anhydrase I complexed with bicarbonate", *Journal of Molecular Biology,* vol. 241, no. 2, pp. 226-232.

Kuraku, S. 2010, "Palaeophylogenomics of the vertebrate ancestor--impact of hidden paralogy on hagfish and lamprey gene phylogeny", *Integrative and comparative biology,* vol. 50, no. 1, pp. 124-129.

Kuraku, S. 2008, "Insights into cyclostome phylogenomics: pre-2R or post-2R", *Zoological Science,* vol. 25, no. 10, pp. 960-968.

Kuraku, S., Meyer, A. & Kuratani, S. 2009, "Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after?", *Molecular biology and evolution,* vol. 26, no. 1, pp. 47-59.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. 2007, "Clustal W and Clustal X version 2.0", *Bioinformatics (Oxford, England),* vol. 23, no. 21, pp. 2947-2948.

Liljas, A. & Laurberg, M. 2000, "A wheel invented three times. The molecular structures of the three carbonic anhydrases", *EMBO reports,* vol. 1, no. 1, pp. 16-17.

Lin, T.Y., Liao, B.K., Horng, J.L., Yan, J.J., Hsiao, C.D. & Hwang, P.P. 2008, "Carbonic anhydrase 2-like a and 15a are involved in acid-base regulation and Na+ uptake in zebrafish H+-ATPase-rich cells", *American journal of physiology.Cell physiology,* vol. 294, no. 5, pp. C1250-60.

Lindskog, S. & Coleman, J.E. 1973, "The catalytic mechanism of carbonic anhydrase", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 70, no. 9, pp. 2505-2508.

Matsuura, M., Nishihara, H., Onimaru, K., Kokubo, N., Kuraku, S., Kusakabe, R., Okada, N., Kuratani, S. & Tanaka, M. 2008, "Identification of four Engrailed genes in the Japanese lamprey, Lethenteron japonicum", *Developmental dynamics : an official publication of the American Association of Anatomists,* vol. 237, no. 6, pp. 1581-1589.

McGinn, P.J. & Morel, F.M. 2008, "Expression and regulation of carbonic anhydrases in the marine diatom Thalassiosira pseudonana and in natural phytoplankton assemblages from Great Bay, New Jersey", *Physiologia Plantarum,* vol. 133, no. 1, pp. 78-91.

McLysaght, A., Hokamp, K. & Wolfe, K.H. 2002, "Extensive genomic duplication during early chordate evolution", *Nature genetics,* vol. 31, no. 2, pp. 200-204.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B.M., Fujita, P.A., Dreszer, T.R., Diekhans, M., Cline, M.S., Clawson, H., Barber, G.P., Haussler, D. & Kent, W.J. 2013, "The UCSC Genome Browser database: extensions and updates 2013", *Nucleic acids research,* vol. 41, no. Database issue, pp. D64-9.

Mitsuhashi, S., Mizushima, T., Yamashita, E., Miyachi, S. & Tsukihara, T. 2000, "Crystallization and preliminary X-ray diffraction studies of a beta-carbonic anhydrase from the red alga Porphyridium purpureum", *Acta crystallographica.Section D, Biological crystallography,* vol. 56, no. Pt 2, pp. 210-211.

Orlean, P. & Menon, A.K. 2007, "Thematic review series: lipid posttranslational modifications. GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycophospholipids", *Journal of lipid research,* vol. 48, no. 5, pp. 993-1011.

Ortutay, C., Olatubosun, A., Parkkila, S., Vihinen, M. & Tolvanen, M. 2010, "An evolutionary analysis of insect carbonic anydrases" in *Advances in Medicine and Biology,* ed. L.E. Bernhardt, Nova Science Publishers, Inc., Hauppauge, NY, pp. 145-168.

Paradis, E., Claude, J. & Strimmer, K. 2004, "APE: Analyses of Phylogenetics and Evolution in R language", *Bioinformatics (Oxford, England),* vol. 20, no. 2, pp. 289-290.

Parisi, G., Perales, M., Fornasari, M.S., Colaneri, A., Gonzalez-Schain, N., Gomez-Casati, D., Zimmermann, S., Brennicke, A., Araya, A., Ferry, J.G., Echave, J. & Zabaleta, E. 2004, "Gamma carbonic anhydrases in plant mitochondria", *Plant Molecular Biology,* vol. 55, no. 2, pp. 193-207.

Peterson, R.E., Tu, C. & Linser, P.J. 1997, "Isolation and characterization of a carbonic anhydrase homologue from the zebrafish (Danio rerio)", *Journal of Molecular Evolution,* vol. 44, no. 4, pp. 432-439.

Pierleoni, A., Martelli, P.L. & Casadio, R. 2008, "PredGPI: a GPI-anchor predictor", *BMC bioinformatics,* vol. 9, pp. 392.

Reinfelder, J.R. 2011, "Carbon concentrating mechanisms in eukaryotic marine phytoplankton", *Annual review of marine science,* vol. 3, pp. 291-315.

Roberts, S.B., Lane, T.W. & Morel, F.M.M. 1997, "Carbonic anhydrase in the marine diatom*Thalassiosira weissflogii*(Bacillariophyceae)", *J Phycol,* vol. 33, pp. 845-850.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A. & Huelsenbeck, J.P. 2012, "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space", *Systematic Biology,* vol. 61, no. 3, pp. 539-542.

Sawaya, M.R., Cannon, G.C., Heinhorst, S., Tanaka, S., Williams, E.B., Yeates, T.O. & Kerfeld, C.A. 2006, "The structure of beta-carbonic anhydrase from the carboxysomal shell reveals a distinct subclass with one active site for the price of two", *The Journal of biological chemistry,* vol. 281, no. 11, pp. 7546-7555.

Schlicker, C., Hall, R.A., Vullo, D., Middelhaufe, S., Gertz, M., Supuran, C.T., Muhlschlegel, F.A. & Steegborn, C. 2009, "Structure and inhibition of the CO2-sensing carbonic anhydrase Can2 from the pathogenic fungus Cryptococcus neoformans", *Journal of Molecular Biology,* vol. 385, no. 4, pp. 1207-1220.

Seron, T.J., Hill, J. & Linser, P.J. 2004, "A GPI-linked carbonic anhydrase expressed in the larval mosquito midgut", *The Journal of experimental biology,* vol. 207, no. Pt 26, pp. 4559-4572.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D. & Higgins, D.G. 2011, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", *Molecular systems biology,* vol. 7, pp. 539.

Smith, K.E., Vanekeris, L.A. & Linser, P.J. 2007, "Cloning and characterization of AgCA9, a novel alpha-carbonic anhydrase from Anopheles gambiae Giles sensu stricto (Diptera: Culicidae) larvae", *The Journal of experimental biology,* vol. 210, no. Pt 22, pp. 3919-3930.

Smith, K.S., Ingram-Smith, C. & Ferry, J.G. 2002, "Roles of the conserved aspartate and arginine in the catalytic mechanism of an archaeal beta-class carbonic anhydrase", *Journal of Bacteriology,* vol. 184, no. 15, pp. 4240-4245.

So, A.K., Espie, G.S., Williams, E.B., Shively, J.M., Heinhorst, S. & Cannon, G.C. 2004, "A novel evolutionary lineage of carbonic anhydrase (epsilon class) is a component of the carboxysome shell", *Journal of Bacteriology,* vol. 186, no. 3, pp. 623-630.

Stams, T., Nair, S.K., Okuyama, T., Waheed, A., Sly, W.S. & Christianson, D.W. 1996, "Crystal structure of the secretory form of membrane-associated human carbonic anhydrase IV at 2.8-A resolution", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 93, no. 24, pp. 13589-13594.

Sterling, D., Alvarez, B.V. & Casey, J.R. 2002, "The extracellular component of a transport metabolon. Extracellular loop 4 of the human AE1 Cl-/HCO3- exchanger binds carbonic anhydrase IV", *The Journal of biological chemistry,* vol. 277, no. 28, pp. 25239-25246.

Strop, P., Smith, K.S., Iverson, T.M., Ferry, J.G. & Rees, D.C. 2001, "Crystal structure of the "cab"-type beta class carbonic anhydrase from the archaeon Methanobacterium thermoautotrophicum", *The Journal of biological chemistry,* vol. 276, no. 13, pp. 10299-10305.

Suarez Covarrubias, A., Larsson, A.M., Hogbom, M., Lindberg, J., Bergfors, T., Bjorkelid, C., Mowbray, S.L., Unge, T. & Jones, T.A. 2005, "Structure and function of carbonic anhydrases from Mycobacterium tuberculosis", *The Journal of biological chemistry,* vol. 280, no. 19, pp. 18782-18789.

Sundaram, V., Rumbolo, P., Grubb, J., Strisciuglio, P. & Sly, W.S. 1986, "Carbonic anhydrase II deficiency: diagnosis and carrier detection using differential enzyme inhibition and inactivation", *American Journal of Human Genetics,* vol. 38, no. 2, pp. 125-136.

Sunderhaus, S., Dudkina, N.V., Jansch, L., Klodmann, J., Heinemeyer, J., Perales, M., Zabaleta, E., Boekema, E.J. & Braun, H.P. 2006, "Carbonic anhydrase subunits form a matrix-exposed domain attached to the membrane arm of mitochondrial complex I in plants", *The Journal of biological chemistry,* vol. 281, no. 10, pp. 6482-6488.

Supuran, C.T. 2008, "Carbonic anhydrases--an overview", *Current pharmaceutical design,* vol. 14, no. 7, pp. 603-614.

Suyama, M., Torrents, D. & Bork, P. 2006, "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments", *Nucleic acids research,* vol. 34, no. Web Server issue, pp. W609-12.

Svichar, N. & Chesler, M. 2003, "Surface carbonic anhydrase activity on astrocytes and neurons facilitates lactate transport", *Glia,* vol. 41, no. 4, pp. 415-419.

Svichar, N., Esquenazi, S., Waheed, A., Sly, W.S. & Chesler, M. 2006, "Functional demonstration of surface carbonic anhydrase IV activity on rat astrocytes", *Glia,* vol. 53, no. 3, pp. 241-247.

Svichar, N., Waheed, A., Sly, W.S., Hennings, J.C., Hubner, C.A. & Chesler, M. 2009, "Carbonic anhydrases CA4 and CA14 both enhance AE3-mediated Cl--HCO3- exchange in hippocampal neurons", *The Journal of neuroscience : the official journal of the Society for Neuroscience,* vol. 29, no. 10, pp. 3252-3258.

Syrjänen, L., Tolvanen, M., Hilvo, M., Olatubosun, A., Innocenti, A., Scozzafava, A., Leppiniemi, J., Niederhauser, B., Hytonen, V.P., Gorr, T.A., Parkkila, S. & Supuran, C.T. 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 Biochemistry,* vol. 11, pp. 28.

Tachibana, M., Allen, A.E., Kikutani, S., Endo, Y., Bowler, C. & Matsuda, Y. 2011, "Localization of putative carbonic anhydrases in two marine diatoms, Phaeodactylum tricornutum and Thalassiosira pseudonana", *Photosynthesis Research,* vol. 109, no. 1-3, pp. 205-221.

Tashian, R.E., Hewett-Emmett, D., Carter, N. & Bergenhem, N.C. 2000, "Carbonic anhydrase (CA)-related proteins (CA-RPs), and transmembrane proteins with CA or CA-RP domains", *EXS,* vol. (90), no. 90, pp. 105-120.

Väänänen, H.K. & Parvinen, E.K. 1983, "High active isoenzyme of carbonic anhydrase in rat calvaria osteoclasts. Immunohistochemical study", *Histochemistry,* vol. 78, no. 4, pp. 481-485.

Vernier, W., Chong, W., Rewolinski, D., Greasley, S., Pauly, T., Shaw, M., Dinh, D., Ferre, R.A., Meador, J.W.,3rd, Nukui, S., Ornelas, M., Paz, R.L. & Reyner, E. 2010, "Thioether benzenesulfonamide inhibitors of carbonic anhydrases II and IV: structure-based drug design, synthesis, and biological evaluation", *Bioorganic & medicinal chemistry,* vol. 18, no. 9, pp. 3307-3319.

Vince, J.W., Carlsson, U. & Reithmeier, R.A. 2000, "Localization of the Cl-/HCO3- anion exchanger binding site to the amino-terminal region of carbonic anhydrase II", *Biochemistry,* vol. 39, no. 44, pp. 13344-13349.

Vince, J.W. & Reithmeier, R.A. 2000, "Identification of the carbonic anhydrase II binding site in the Cl(-)/HCO(3)(-) anion exchanger AE1", *Biochemistry,* vol. 39, no. 18, pp. 5527-5533.

Waheed, A., Parkkila, S., Zhou, X.Y., Tomatsu, S., Tsuchihashi, Z., Feder, J.N., Schatzman, R.C., Britton, R.S., Bacon, B.R. & Sly, W.S. 1997a, "Hereditary hemochromatosis: effects of C282Y and H63D mutations on association with beta2-microglobulin, intracellular processing, and cell surface expression of the HFE protein in COS-7 cells", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 94, no. 23, pp. 12384-12389.

Waheed, A., Pham, T., Won, M., Okuyama, T. & Sly, W.S. 1997b, "Human carbonic anhydrase IV: in vitro activation and purification of disulfide-bonded enzyme following expression in Escherichia coli", *Protein expression and purification,* vol. 9, no. 2, pp. 279-287.

Waheed, A., Zhu, X.L., Sly, W.S., Wetzel, P. & Gros, G. 1992, "Rat skeletal muscle membrane associated carbonic anhydrase is 39-kDa, glycosylated, GPI-anchored CA IV", *Archives of Biochemistry and Biophysics,* vol. 294, no. 2, pp. 550-556.

Warren, W.C., Hillier, L.W., Marshall Graves, J.A., Birney, E., Ponting, C.P., Grutzner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A.T., Yang, S.P., Heger, A., Locke, D.P., Miethke, P., Waters, P.D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., Puente, X.S., Lopez-Otin, C., Ordonez, G.R., Eichler, E.E., Chen, L., Cheng, Z., Deakin, J.E., Alsop, A., Thompson, K., Kirby, P., Papenfuss, A.T., Wakefield, M.J., Olender, T., Lancet, D., Huttley, G.A., Smit, A.F., Pask, A., Temple-Smith, P., Batzer, M.A., Walker, J.A., Konkel, M.K., Harris, R.S., Whittington, C.M., Wong, E.S., Gemmell, N.J., Buschiazzo, E., Vargas Jentzsch, I.M., Merkel, A., Schmitz, J., Zemann, A., Churakov, G., Kriegs, J.O., Brosius, J., Murchison, E.P., Sachidanandam, R., Smith, C., Hannon, G.J., Tsend-Ayush, E., McMillan, D., Attenborough, R., Rens, W., Ferguson-Smith, M., Lefevre, C.M., Sharp, J.A., Nicholas, K.R., Ray, D.A., Kube, M., Reinhardt, R., Pringle, T.H., Taylor, J., Jones, R.C., Nixon, B., Dacheux, J.L., Niwa, H., Sekita, Y., Huang, X., Stark, A., Kheradpour, P., Kellis, M., Flicek, P., Chen, Y., Webber, C., Hardison, R., Nelson, J., Hallsworth-Pepin, K., Delehaunty, K., Markovic, C., Minx, P., Feng, Y., Kremitzki, C., Mitreva, M., Glasscock, J., Wylie, T., Wohldmann, P., Thiru, P., Nhan, M.N., Pohl, C.S., Smith, S.M., Hou, S., Nefedov, M., de Jong, P.J., Renfree, M.B., Mardis, E.R. & Wilson, R.K. 2008, "Genome analysis of the platypus reveals unique signatures of evolution", *Nature,* vol. 453, no. 7192, pp. 175-183.

Weber, A.K. & Pirow, R. 2009, "Physiological responses of Daphnia pulex to acid stress", *BMC physiology,* vol. 9, pp. 9-6793-9-9.

Wetzel, P., Hasse, A., Papadopoulos, S., Voipio, J., Kaila, K. & Gros, G. 2001, "Extracellular carbonic anhydrase activity facilitates lactic acid transport in rat skeletal muscle fibres", *The Journal of physiology,* vol. 531, no. Pt 3, pp. 743-756.

Wolfe, K.H. & Shields, D.C. 1997, "Molecular evidence for an ancient duplication of the entire yeast genome", *Nature,* vol. 387, no. 6634, pp. 708-713.

Xu, Y., Feng, L., Jeffrey, P.D., Shi, Y. & Morel, F.M. 2008, "Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms", *Nature,* vol. 452, no. 7183, pp. 56-61.

Yamano, T. & Fukuzawa, H. 2009, "Carbon-concentrating mechanism in a green alga, Chlamydomonas reinhardtii, revealed by transcriptome analyses", *Journal of Basic Microbiology,* vol. 49, no. 1, pp. 42-51.

Yang, Z., Alvarez, B.V., Chakarova, C., Jiang, L., Karan, G., Frederick, J.M., Zhao, Y., Sauve, Y., Li, X., Zrenner, E., Wissinger, B., Hollander, A.I., Katz, B., Baehr, W., Cremers, F.P., Casey, J.R., Bhattacharya, S.S. & Zhang, K. 2005, "Mutant carbonic anhydrase 4 impairs pH regulation and causes retinal photoreceptor degeneration", *Human molecular genetics,* vol. 14, no. 2, pp. 255-265.

Zabaleta, E., Martin, M.V. & Braun, H.P. 2012, "A basal carbon concentrating mechanism in plants?", *Plant science : an international journal of experimental plant biology,* vol. 187, pp. 97-104.

Zimmerman, S.A. & Ferry, J.G. 2008, "The beta and gamma classes of carbonic anhydrase", *Current pharmaceutical design,* vol. 14, no. 7, pp. 716-721.

Original communications

Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase

Mika HILVO*[1,2], Martti TOLVANEN*[1], Amy CLARK†, Bairong SHEN*, Gul N. SHAH†, Abdul WAHEED†, Piia HALMI*, Milla HÄNNINEN*, Jonna M. HÄMÄLÄINEN*, Mauno VIHINEN*, William S. SLY† and Seppo PARKKILA*‡

*Institute of Medical Technology, University of Tampere and Tampere University Hospital, Biokatu 6, 33520 Tampere, Finland, †Department of Biochemistry and Molecular Biology, Saint Louis University School of Medicine, St. Louis, MO, U.S.A., and ‡Department of Clinical Chemistry, University of Oulu, Kajaanintie 50, 90220 Oulu, Finland

The main function of CAs (carbonic anhydrases) is to participate in the regulation of acid–base balance. Although 12 active isoenzymes of this family had already been described, analyses of genomic databases suggested that there still exists another isoenzyme, CA XV. Sequence analyses were performed to identify those species that are likely to have an active form of this enzyme. Eight species had genomic sequences encoding CA XV, in which all the amino acid residues critical for CA activity are present. However, based on the sequence data, it was apparent that CA XV has become a non-processed pseudogene in humans and chimpanzees. RT-PCR (reverse transcriptase PCR) confirmed that humans do not express CA XV. In contrast, RT-PCR and *in situ* hybridization performed in mice showed positive expression in the kidney, brain and testis. A prediction of the mouse CA XV structure was performed. Phylogenetic analysis showed that mouse CA XV is related to CA IV. Therefore both of these enzymes were expressed in COS-7 cells and studied in parallel experiments. The results showed that CA XV shares several properties with CA IV, i.e. it is a glycosylated glycosylphosphatidylinositol-anchored membrane protein, and it binds CA inhibitor. The catalytic activity of CA XV is low, and the correct formation of disulphide bridges is important for the activity. Both specific and non-specific chaperones increase the production of active enzyme. The results suggest that CA XV is the first member of the α-CA gene family that is expressed in several species, but not in humans and chimpanzees.

Key words: bioinformatics, carbonic anhydrase XV, glycosylphosphatidylinositol (GPI) anchor.

INTRODUCTION

CAs (carbonic anhydrases) are zinc-containing metalloenzymes that catalyse the reversible hydration of carbon dioxide according to the following reaction: $CO_2 + H_2O \leftrightarrow HCO_3^- + H^+$ [1]. This reaction forms the basis for the regulation of acid–base balance in organisms. In addition to this main function, CAs participate in a number of other physiological processes, such as CO_2 and HCO_3^- transport, bone resorption, production of body fluids, gluconeogenesis, ureagenesis and lipogenesis [2]. During evolution, at least 12 active CA isoenzymes have emerged in both rodents and humans. The isoenzymes have differences in their tissue distribution, kinetic properties and subcellular localizations: CAs I, II, III, VII and XIII [3,4] are cytoplasmic, IV, IX, XII and XIV [5–8] are membrane-bound, VI is secreted [9] and VA (as well as VB) are located in mitochondria [10]. Three CA-RPs (CA-related proteins), VIII, X, XI, are closely similar to CAs, but they lack one or more of the critical histidine residues in their active site. Since these histidines are required to bind the zinc ion, which is essential for CO_2 hydration activity, the CA-RPs lack enzymatic activity [11].

The aim of the present study was to characterize the novel CA XV, the cDNA sequence of which was submitted to the National Center for Biotechnology Information by Hewett-Emmett and Shimmin in 2000 (GenBank® accession no. AF231122). Genome-wide sequence analyses were performed to identify the species that might express an active CA XV enzyme in their proteome. The expression of this novel isoenzyme was studied in human and mouse tissues by RT (reverse transcriptase)-PCR, and the three-dimensional structure of the murine enzyme was predicted by computer modelling. Recombinant mouse CA XV was produced in COS-7 cells and *Escherichia coli* bacteria. Biochemical characterization of CA XV revealed that it is enzymatically active and is bound to the plasma membrane through a GPI (glycosylphosphatidylinositol) anchorage, as found for CA IV. The production of catalytically active recombinant enzyme in transfected COS-7 cells was markedly increased by treatment with both specific and non-specific chaperones, as recently reported for human CA IV [12].

EXPERIMENTAL

Bioinformatics analyses

Sequence analyses were performed in order to find out which species may produce an active CA XV. The procedure was started with the search and retrieval of known *CA15* sequences [*Mus musculus* (house mouse)]. Subsequently, BLAT [13] searches were performed in all selected genomes using sequences found in the previous step as query sequences. The genomic sequences were obtained, and these were translated in three frames. The translations were aligned with known sequences, and the exon locations were visually identified. The gene models were confirmed and fine-tuned using GeneWise [14]. Finally, the final best transcripts and protein sequences were assembled manually.

The following sequences were obtained from Ensembl Genome Browser (http://www.ensembl.org): *Mus musculus* (ENSMUSP-00000012152) and *Rattus norvegicus* (brown rat) (ENSRNOP-00000000312). The UCSC Genome Browser (http://genome.ucsc.edu) showed two mRNAs for the *CA15* of *Gallus gallus*

Abbreviations used: CA, carbonic anhydrase; CA-RP, carbonic anhydrase-related protein; DTT, dithiothreitol; EndoH, endoglycosidase H; EST, expressed sequence tag; GPI, glycosylphosphatidylinositol; NP40, Nonidet P40; p-AMBS, *p*-aminomethylbenzenesulphonamide; PBA, 4-phenylbutyric acid; PI-PLC, phosphoinositide-specific phospholipase C; poly(A)$^+$, polyadenylated; RT, reverse transcriptase.

[1] These authors contributed equally to this study.

[2] To whom correspondence should be addressed (email Mika.Hilvo@uta.fi).

© 2005 Biochemical Society

(chicken): of these two, the accession number BX929589 was selected and translated into protein, because it was in closest accordance with other species. For *Danio rerio* (zebrafish), there was an EST (expressed sequence tag) sequence (CO960501) representing *CA15* that was mainly of high quality. Because the latter part of the EST sequence was of lower quality, it was constructed manually from the genome. The sequence for *Tetraodon nigroviridis* (green spotted pufferfish) was also obtained from UCSC Genome Browser, Genscan Gene Prediction (GSCT00001777001). The *CA15* of *Fugu rubripes* (fugu pufferfish) was constructed manually by using the information of Ensembl Genome Browser (SINFRUP00000165581 and SINFRUP00000175429). The UCSC Genome Browser also showed some EST sequences for this gene. The gene for the *CA15* of *Canis familiaris* (dog) could not be found in the biological databases, and it was constructed manually from UCSC Genome Browser from region chr26: 32 168 776–32 171 692. The CA XV amino acid sequence of *Xenopus tropicalis* (pipid frog) was constructed manually with the help of two EST sequences (BX734706 and AL890846). The sequence alignment including all the species was constructed with T-Coffee version 2.11 [15].

For humans (*Homo sapiens*) or chimpanzees (*Pan troglodytes*), there were no mRNAs or EST sequences representing *CA15* in the Ensembl Genome Browser. However, the human genome was shown to have three copies of *CA15* located to chromosome 22q11.21: positions 17 393 598–17 396 941, 18 860 411–18 863 739 and 20 034 808–20 038 136. For convenience, they will be referred to as human *CA15* candidate genes 1, 2, and 3 respectively. In chimpanzee (*Pan troglodytes*), two gene candidates were observed in chromosome 23, at positions 17 310 504–17 312 161 and 19 990 225–19 993 434. These are referred as gene candidates 1 and 3 respectively. The second chimpanzee gene candidate that should be syntenic with the human gene candidate 2 was in a sequence gap of the genome. All of these gene candidates were observed to be pseudogenic, because they contained several frame-shifts and point mutations. Detailed information on these pseudogenes can be found in Supplementary Figures 1, 2 and 3 (see http://www.BiochemJ.org/bj/392/bj3920083add.htm).

The prediction of N-glycosylation sites for mouse CA XV was performed by using NetNGlyc 1.0 Server with default parameters (http://www.cbs.dtu.dk/services/NetNGlyc/).

The sequence alignment of CAs for phylogenetic analysis was performed with ClustalW [17]. Phylogenetic analysis was carried out with PAUP* 4.0 [18]. The majority rule consensus tree was obtained by three bootstrap runs with different random seeds. The dataset was bootstrapped 1000 times for each run.

The structural prediction for mouse CA XV was based on human CA IV at 2.8 Å (1 Å ≡ 0.1 nm) resolution (PDB entry 1ZNC) [19]. In structural modelling, amino acids 25–304 (Figure 1) from the mouse CA XV were included. The model was constructed with the program InsightII (Acelrys Inc., San Diego, CA, U.S.A.). Amino acid substitutions were built using a side chain rotamer library. The initial model was refined with Discover™ in a stepwise manner by energy minimization using the Amber™ force field. The newly built loops were refined with 500 steps of minimization with a fixed and a free backbone respectively. Subsequently, all side chains with a constrained backbone were minimized for 500 steps, followed by a further 1000 steps of minimization for the whole model.

RT-PCR experiments, sequencing of the PCR products and *in situ* hybridization

The RT-PCR method was used to reveal those murine and human tissues that express *CA15* mRNA. The expression studies were carried out using the commercial cDNA kits purchased from BD Biosciences (Palo Alto, CA, U.S.A.). The mouse MTC™ panel I contained first-strand cDNA preparations produced from total poly(A)$^+$ (polyadenylated) RNAs isolated from 12 different murine tissues. In addition, mRNA was isolated from six mouse tissues absent in the panel (stomach, duodenum, jejunum, ileum, colon and blood) by using TRIzol® reagent (Invitrogen, Carlsbad, CA, U.S.A.). Reverse transcription was performed with Mo-MuLV reverse transcriptase (Finnzymes, Espoo, Finland) using random primers (500 μg/ml). The procedures were conducted according to the principles of the Declaration of Helsinki and approved by the institutional animal care committee (University of Tampere, Finland). The human MTC™ Panels I and II were used to study the expression of *CA15* mRNA in 15 human tissues.

To study mRNA in mouse, sequence-specific primers for murine *CA15* were designed using the information published in GenBank (accession number NM_030558). The forward primer (MF1) was 5′-TACCTGGTGCTACGACTC-3′ (nt 148–165) and the reverse primer (MR1) was 5′-TATCGGTAGTACCGCAAG-3′ (nt 739–756), and thus the resulting amplification product size was 609 bp.

Sequence analyses revealed that the human genome contains three copies of *CA15* genes which have most likely become pseudogenes. In order to obtain experimental data concerning whether any of these candidate genes are expressed, primers were designed for each of them, and additionally, one primer pair was designed to recognize all of them. More detailed information on the primers used for human studies, as well as exact PCR conditions for all reactions, can be found in the Supplementary methods section (see http://www.BiochemJ.org/bj/392/bj3920083add.htm).

The results of the PCR reactions were analysed using a 1.5 % agarose gel containing 0.1 μg/ml ethidium bromide with a DNA standard (100 bp DNA Ladder; New England Biolabs, Beverly, MA, U.S.A.). The 609 and 713 bp PCR products were sequenced in order to confirm their identity and to reveal possible unspecific binding of primers. The detailed protocol of the sequencing can be found in the Supplementary Methods section (http://www.BiochemJ.org/bj/392/bj3920083add.htm).

In situ hybridization was performed for mouse tissues as described by Heikinheimo et al. [20].

Expression of CA XV in COS-7 cells and CA assay

The full-length mouse *CA15* cDNA in pME 18S-FL3 (I.M.A.G.E Consortium, clone ID 1908347, MRC Geneservice, Cambridge, U.K.) and, as a control, mouse *CA4* cDNA in pCXN [21] was used for transient transfection of COS-7 cells using the DEAE–dextran procedure [22]. After 12 h of transfection, cells were treated with 100 mM chloroquine for 1 h [23]. After 72 h of transfection, COS-7 cells were harvested in cold PBS and the cell pellets were homogenized in 25 mM Tris/SO_4 buffer, pH 7.5, containing protease inhibitors by sonication. The membrane-associated enzymes were separated after centrifugation at 100 000 ***g*** for 30 min. The protein concentration was determined by the micro-Lowry protein assay [24].

The CA activity from cell lysates or membrane suspension was determined in duplicate by the procedure of Maren, as described previously [25]. The CA activity was expressed as units/mg of cell protein or membrane protein.

Binding of CA XV to the CA-inhibitor affinity resin

p-AMBS (*p*-aminomethylbenzenesulphonamide) affigel (Sigma–Aldrich, St Louis, MO, U.S.A.) affinity resin was equilibrated with 10 mM Hepes/NaOH, pH 7.5. Cell membrane extracts in 10 mM Hepes/NaOH, pH 7.5/1 % (v/v) NP40 (Nonidet P40) buffer were

© 2005 Biochemical Society

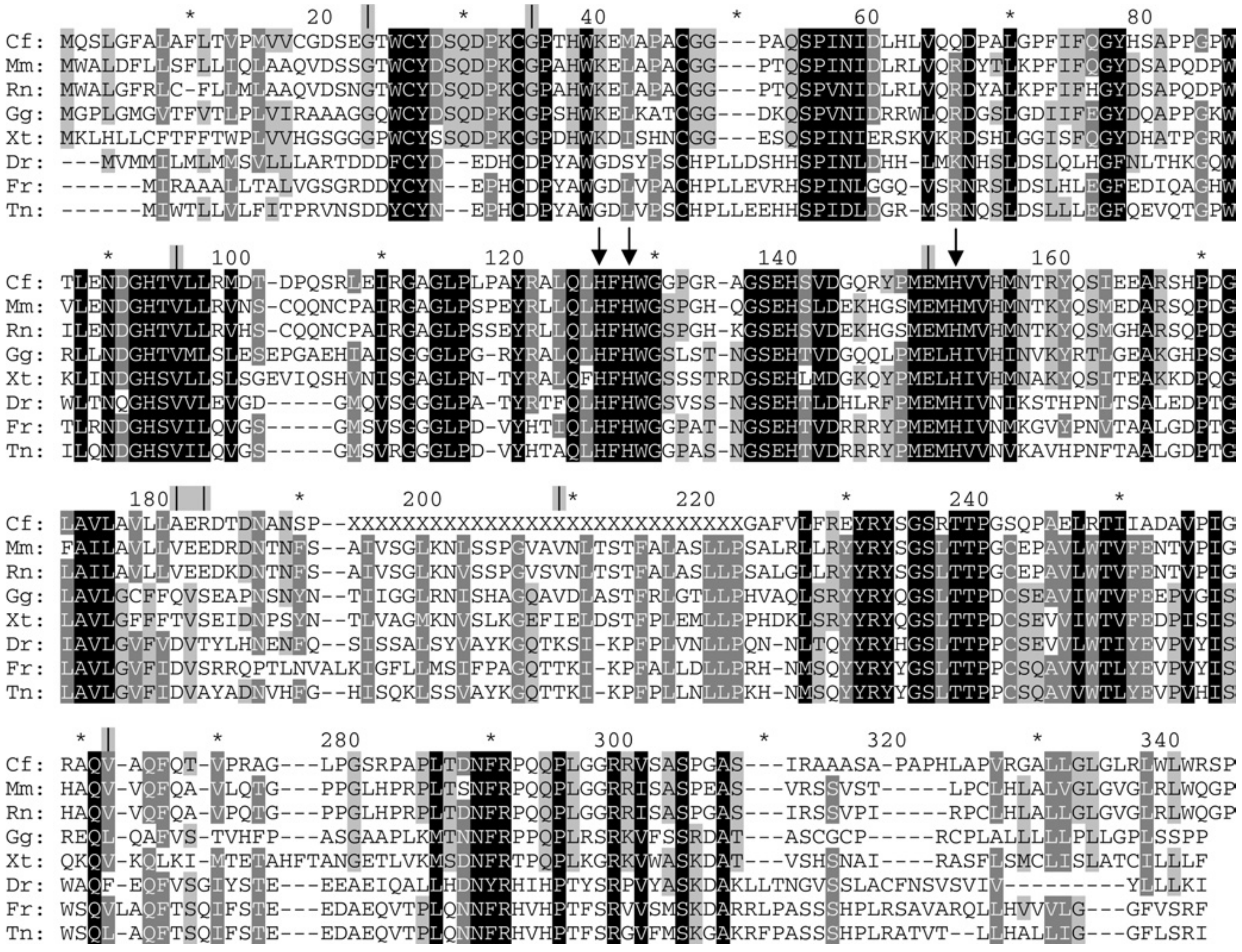
* 20 * 40 * 60 * 80
Cf: MQSLGFALAFLTVPMVVCGDSEGTWCYDSQDPKCGPTHWKEMAPACGG---PAQSPINIDLHLVQQDPALGPFIFQGYHSAPPGPW
Mm: MWALDFLLSFLLIQLAAQVDSSGTWCYDSQDPKCGPAHWKELAPACGG---PTQSPINIDLRLVQRDYTLKPFIFQGYDSAPQDPW
Rn: MWALGFRLC-FLLMLAAQVDSNGTWCYDSQDPKCGPAHWKELAPACGG---PTQSPVNIDLRLVQRDYALKPFIFHGYDSAPQDPW
Gg: MGPLGMGVTFVTLPLVIRAAAGGQWCYDSQDPKCGPSHWKELKATCGG---DKQSPVNIDRRWLQRDGSLGDIIFEGYDQAPPGKW
Xt: MKLHLLCFTFFTWPLVVHGSGGGPWCYSSQDPKCGPDHWKDISHNCGG---ESQSPINIERSKVKRDSHLGGISFQGYDHATPGRW
Dr: ---MVMMILMLMMSVLLLARTDDDFCYD--EDHCDPYAWGDSYPSCHPLLDSHHSPINLDHH-LMKNHSLDSLQLHGFNLTHKGQW
Fr: ------MIRAAALLTALVGSGRDDYCYN--EPHCDPYAWGDLVPACHPLLEVRHSPINLGGQ-VSRNRSLDSLHLEGFEDIQAGHW
Tn: ------MIWTLLVLFITPRVNSDDYCYN--EPHCDPYAWGDLVPSCHPLLEEHHSPIDLDGR-MSRNQSLDSLLLEGFQEVQTGPW
* 100 * 120 * 140 160 *
Cf: TLENDGHTVLLRMDT-DPQSRLEIRGAGLPLPAYRALQLHFHWGGPGR-AGSEHSVDGQRYPMEMHVVHMNTRYQSIEEARSHPDG
Mm: VLENDGHTVLLRVNS-CQQNCPAIRGAGLPSPEYRLLQLHFHWGSPGH-QGSEHSLDEKHGSMEMHMVHMNTKYQSMEDARSQPDG
Rn: ILENDGHTVLLRVHS-CQQNCPAIRGAGLPSSEYRLLQLHFHWGSPGH-KGSEHSVDEKHGSMEMHMVHMNTKYQSMGHARSQPDG
Gg: RLLNDGHTVMLSLESEPGAEHIAISGGGLPG-RYRALQLHFHWGSLST-NGSEHTVDGQQLPMELHIVHINVKYRTLGEAKGHPSG
Xt: KLINDGHSVLLSLSGEVIQSHVNISGAGLPN-TYRALQFHFHWGSSSTRDGSEHLMDGKQYPMELHIVHMNAKYQSITEAKKDPQG
Dr: WLTNQGHSVVLEVGD-----GMQVSGGGLPA-TYRTFQLHFHWGSVSS-NGSEHTLDHLRFPMEMHIVNIKSTHPNLTSALEDPTG
Fr: TLRNDGHSVILQVGS-----GMSVSGGGLPD-VYHTIQLHFHWGGPAT-NGSEHTVDRRRYPMEMHIVNMKGVYPNVTAALGDPTG
Tn: ILQNDGHSVILQVGS-----GMSVRGGGLPD-VYHTAQLHFHWGGPAS-NGSEHTVDRRRYPMEMHVVNVKAVHPNFTAALGDPTG
180 * 200 * 220 * 240 *
Cf: LAVLAVLLAERDTDNANSP--XXXXXXXXXXXXXXXXXXXXXXXXXXXXXGAFVLFREYRYSGSRTTPGSQPAELRTIIADAVPIG
Mm: FAILAVLLVEEDRDNTNFS--AIVSGLKNLSSPGVAVNLTSTFALASLLPSALRLLRYYRYSGSLTTPGCEPAVLWTVFENTVPIG
Rn: LAILAVLLVEEDKDNTNFS--AIVSGLKNVSSPGVSVNLTSTFALASLLPSALGLLRYYRYSGSLTTPGCEPAVLWTVFENTVPIG
Gg: LAVLGCFFQVSEAPNSNYN--TIIGGLRNISHAGQAVDLASTFRLGTLLPHVAQLSRYYRYQGSLTTPDCSEAVIWTVFEEPVGIS
Xt: LAVLGFFFTVSEIDNPSYN--TLVAGMKNVSLKGEFIELDSTFPLEMLLPPHDKLSRYYRYQGSLTTPDCSEVVIWTVFEDPISIS
Dr: IAVLGVFVDVTYLHNENFQ--SISSALSYVAYKGQTKSI-KPFPLVNLLPQN-NLTQYYRYHGSLTTPPCSEVVLWTIYEVPVYIS
Fr: LAVLGVFIDVSRRQPTLNVALKIGFLLMSIFPAGQTTKI-KPFALLDLLPRH-NMSQYYRYYGSLTTPPCSQAVVWTLYEVPVYIS
Tn: LAVLGVFIDVAYADNVHFG--HISQKLSSVAYKGQTTKI-KPFPLLNLLPKH-NMSQYYRYYGSLTTPPCSQAVVWTLYEVPVHIS
* * 280 * 300 * 320 * 340
Cf: RAQV-AQFQT-VPRAG---LPGSRPAPLTDNFRPQQPLGGRRVSASPGAS---IRAAASA-PAPHLAPVRGALLGLGLRLWLWRSP
Mm: HAQV-VQFQA-VLQTG---PPGLHPRPLTSNFRPQQPLGGRRISASPEAS---VRSSVST-----LPCLHLALVGLGVGLRLWQGP
Rn: HAQV-VQFQA-VPQTG---PPGLHPRPLTDNFRPQQPLGGRRISASPGAS---IRSSVPI-----RPCLHLALLGLGVGLRLWQGP
Gg: REQL-QAFVS-TVHFP---ASGAAPLKMTNNFRPPQPLRSRKVFSSRDAT---ASCGCP-----RCPLALLLLLPLLGPLSSPP
Xt: QKQV-KQLKI-MTETAHFTANGETLVKMSDNFRTPQPLKGRKVWASKDAT---VSHSNAI----RASFLSMCLISLATCILLLF
Dr: WAQF-EQFVSGIYSTE---EEAEIQALLHDNYRHIHPTYSRPVYASKDAKLLTNGVSSLACFNSVSVIV---------YLLLKI
Fr: WSQVLAQFTSQIFSTE---EDAEQVTPLQNNFRHVHPTFSRVVSMSKDARRLPASSSHPLRSAVARQLLHVVVLG---GFVSRF
Tn: WSQL-AQFTSQIFSTE---EDAEQVTPLQNNFRHVHPTFSRGVFMSKGAKRFPASSSHPLRATVT--LLHALLIG---GFLSRI

Figure 1 The results of sequence analyses

The alignment of CA XV in eight species, showing its conservation throughout evolution. The three histidine residues co-ordinating the zinc atom are pointed out by arrows. Exon boundaries are indicated by vertical lines above the alignment; asterisks denote every tenth residue not labelled in the Figure. The only exon boundary having interspecies differences is located at amino acid residues 181–183. Abbreviations: Cf, *Canis familiaris*; Mm, *Mus musculus*; Rn, *Rattus norvegicus*; Gg, *Gallus gallus*; Xt, *Xenopus tropicalis*; Dr, *Danio rerio*; Fr, *Fugu rubripes*; Tn, *Tetraodon nigroviridis*. X, sequencing gap in the genome.

mixed with the affinity resin at 4 °C for 30 min or on ice with intermittent mixing. Unbound proteins were removed and the resin–enzyme complex was washed with 25 mM Tris/SO_4, pH 7.5, containing protease inhibitor buffer. The bound enzyme was eluted with SDS/PAGE sample buffer and analysed by Western blot. SDS/PAGE was carried out according to Laemmli [26] under reducing/non-reducing conditions. The polypeptides were characterized by Western blotting, as described previously [7]. For the detection of CA XV, the antibody for CA IV was used, because mouse CA XV shares a distinct homology with mouse CA IV. Affinity-purified recombinant mouse CA IV enzyme was used to raise antibodies in rabbit, as described previously [8,27]. The goat anti-rabbit IgG–peroxidase conjugate was purchased from Sigma–Aldrich. The intensity of the bands was quantified and expressed as the percentage enzyme bound of the total enzyme present in the membrane extract.

Deglycosylation of mouse CA XV and PI-PLC (phosphoinositide-specific phospholipase C) treatment

COS-7 cell membranes equivalent to 50 μg of protein were treated with EndoH (endoglycosidase H) from Boehringer Mannheim (Mannheim, Germany) as described in [28]. COS-7 cell membranes were treated with PI-PLC from ICN Biomedical Research Products (Costa Mesa, CA, U.S.A.) for 2 h at room temperature [29]. For a control, the membranes were treated with buffer alone. Both deglycosylation as well as GPI anchoring were analysed by SDS/PAGE followed by Western blotting.

In vitro refolding of CA XV by oxidized glutathione and *in vivo* refolding by chemical chaperones

CA XV protein was refolded using 10 mM oxidized glutathione (GSSG; Sigma-Aldrich) as described previously [30]. COS-7 cell membrane extract in 25 mM Tris/$SO_{4,}$ pH 7.5, containing 1 % NP40 and protease inhibitors was incubated with 10 mM GSSG at 4 °C for 24 h before the enzyme assay. COS-7 cell membrane extracts overexpressing mouse CA IV were used as a positive control.

The transfected COS-7 cells, just after chloroquine treatment, were incubated with different concentrations of PBA (4-phenylbutyric acid; Sigma–Aldrich) and dorzolamide (Trusopt, Walgreen's Pharmacy, Deerfield, IL, U.S.A.) for 72 h. The cell lysates were used to isolate the cell membranes after

© 2005 Biochemical Society

centrifugation at 30 000 g. The cell membranes were washed with 20 mM sodium acetate buffer, pH 5.5, to remove bound CA inhibitor. The membrane pellets were recovered after washing and suspended in 20 mM Tris/SO_4 buffer, pH 7.5, for the CA activity measurement. The enzyme activity was measured in duplicate, and average enzyme activity was used to calculate the fold increase in the CA activity.

Expression and purification of recombinant CA XV produced in *E. coli*

The fragment of *CA15* was amplified by PCR from a full-length *CA15* into a pCXN vector. The NdeI and BamHI sites were introduced. The amplified product encoding amino acids 22–292 of the full-length native sequence was cloned into TA vector (Invitrogen) and sequenced. The DNA fragment, isolated by restriction-enzyme digestion, was cloned into the bacterial expression vector pET11a (Novagen) at the NdeI/BamHI sites to construct CA XV P293X pET11a. *E. coli* host cells [BL21(DE3)-pLysS or Origami (DE3)], from Novagen, were transformed with the vector.

E. coli host cells BL21(DE3)pLysS or Origami (DE3) containing CA XV P293X pET11a were grown as described for human CA II [31]. For Origami (DE3) cells, the antibiotics used were kanamycin, tetracycline and ampicillin. The Origami host strains are K-12 derivatives that have mutations in both the thioredoxin reductase (*trxB*) and glutathione reductase (*gor*) genes, which enhance disulphide-bond formation in the cytoplasm. The *trx* and *gor* mutations are selectable on kanamycin and tetracycline respectively. For BL21(DE3)pLysS cells, the culture medium contained ampicillin and chloramphenicol. The overnight cultures were diluted 100–200-fold in Luria–Bertani medium containing appropriate antibiotics, and grown at 37 °C with shaking to an attenuance at 600 nm (D_{600}) of 0.5. The expression of mouse CA XV was induced by 0.5 mM IPTG (isopropyl β-D-thiogalactoside) and 0.6 mM zinc sulphate, and the culture was allowed to continue for a further 3 h. The bacterial cells were recovered after centrifugation.

The bacterial cell pellet was lysed in 10 mM Hepes/NaOH, pH 7.5/0.5 % (v/v) Triton X-100 containing 1 mM each of PMSF, *o*-phenanthroline, iodoacetamide and EDTA, as well as 6 μM $ZnSO_4$, using a Brinkman Polytron at 4 °C. The mouse CA XV in the cell lysate was refolded by incubating with 10 mM oxidized glutathione at 4 °C for 72 h, as described previously for CA IV [30]. The clear supernatant recovered after centrifugation at 30 000 g for 30 min was mixed with one-tenth the volume of 1 M Tris/SO_4 buffer, pH 9.0. The cell supernatant containing CA XV was applied to a p-AMBS Affigel column equilibrated with 10 mM Hepes/NaOH buffer, pH 7.5. The unbound proteins were removed by equilibration buffer, followed by 50 mM NaCl in equilibration buffer and 100 mM Tris/SO_4 buffer, pH 7.5. The affinity-resin-bound proteins were eluted with 100 mM sodium acetate, pH 5.5, containing 0.5 M sodium perchlorate. The fractions containing the enzymes were concentrated on Amicon tubes and dialysed against 10 mM Tris/SO_4 buffer, pH 7.5. The enzyme was purified further on a Sephacryl S-300 sizing column. The fractions were analysed by SDS/PAGE. The polypeptides were visualized by silver staining or by Western blot analysis.

RESULTS

Bioinformatics analyses

The sequence analyses were applied to 10 species in order to reveal those that could possess an active *CA15* gene in their genome. The results of the sequence analyses showed that eight animal species are likely to have an active *CA15*. The alignment of the predicted CA XV sequence of these species is shown in Figure 1. The amino acid sequence for CA XV, as well as its exon structure, seems to be well maintained throughout evolution, because species from different taxons showed conserved sequences. Surprisingly, humans were shown to have three copies of *CA15*, and chimpanzees at least two copies of the gene. Chimpanzees may also possess one more copy, since there was a sequencing gap in the region where the third gene might reside. However, all of these human and chimpanzee *CA15* genes seem to have become pseudogenes for various reasons: an AluY repeat splits exon 8, and there are several frame-shifts resulting in the loss of many CA hallmark residues conserved in all active isoenzymes. There is also an insertion in the middle of the active site, and one intron has lost the essential GT dinucleotide in the beginning. In addition, each copy shows unique point mutations and frame-shifts. Figure 2 shows those disrupting features that were common to all of these five gene copies of *CA15*. Each of these defects is alone likely to be able to make this CA isoenzyme non-functional. More detailed information on these errors can be found in Supplementary Figure 1 (http://www.BiochemJ.org/bj/392/bj3920083add.htm), where the sequences for all of these gene copies have been presented. Supplementary Figure 2 shows an alignment where a reconstruction of human CA XV has been included. Supplementary Figure 3 presents this gene in the human genome, and shows that the automatic gene prediction algorithms in genome databases had been unsuccessful in finding the correct exons for these pseudogenes. In databases, no mRNAs or EST sequences have been reported for any of these genes. In all species, except humans and chimpanzees, the three histidine residues critical for CA activity have been conserved. Thus it can be summarized that humans and chimpanzees seem to have lost their CA XV during evolution.

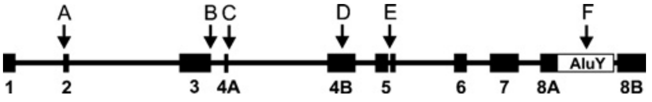

Figure 2 Organization of human and chimpanzee CA XV pseudogenes

Numbered black boxes show reconstructed exons. Exon and intron lengths are presented in the correct scale. Arrows highlight those defects that are common to all five copies. Arrows A and B: frame-shifts. Arrow C, the beginning of the intron after exon 4A has GA instead of the conserved GT dinucleotide. Arrow D, a 9 bp insertion in exon 4B, disrupting the active centre. Arrow E, a 4-bp insertion in exon 5, leading to a frame-shifted sequence in a region which is highly conserved in all CAs. Arrow F, insertion of an AluY-repeat sequence which splits exon 8, duplicating 17 bp of the exon sequence (duplicated part seen as a gap after AluY).

The draft assembly of the genome of rhesus monkey (*Macaca mulatta*) was inspected at the UCSC genome browser. The quality of the genomic sequence was low and therefore not all exons were found. At the active site, the second histidine was mutated to asparagine, identical with human and chimpanzee pseudogenes, and in addition, there were several in-frame stop codons and one frame-shift in the sequences. There was no AluY sequence insertion corresponding to human and chimpanzee sequences. However, from the pieces of *CA15* sequence that we could retrieve, we concluded that *CA15* has probably become a pseudogene also in the rhesus monkey.

Because humans did not show active CA XV, mouse CA XV was characterized in order to reveal the properties of this new enzyme. The mouse CA XV protein sequence contains three potential glycosylation sites (Asn-not Pro-Ser/Thr) at asparagine residues 189, 201 and 210 (numbering shown in Figure 1). The NetNGlyc 1.0 server predicted all three sites to be capable

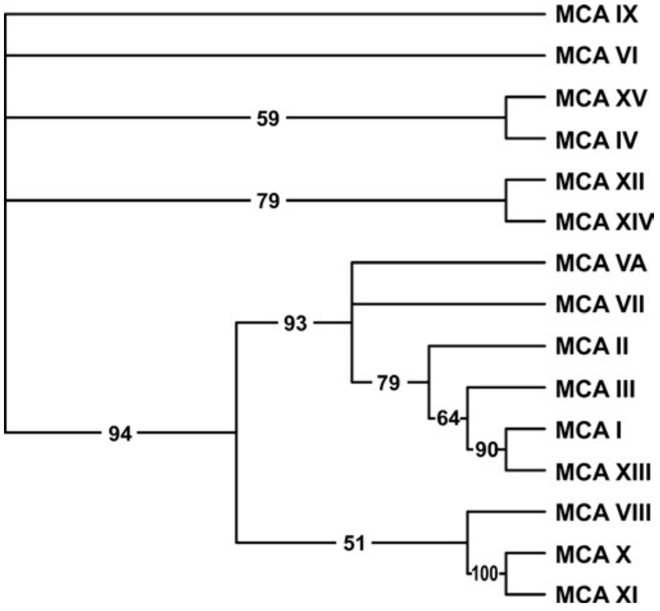

Figure 3 The phylogenetic tree of 15 mouse CAs and CA-RPs

The numbers at the branches show confidence levels in the bootstrap analysis. The tree implies that mouse CA XV is most closely related to CA IV.

of being glycosylated (potential > threshold 0.5). The potentials were 0.5977, 0.6636 and 0.8056 respectively. This result is also supported by endoglycosidase digestion data in the present paper.

A phylogenetic analysis was performed to estimate evolutionary relationship of CA XV to the other known CA isoforms. The results in Figure 3 show that CA XV is most closely related to the extracellular, GPI-anchored CA IV.

A prediction for the structure of mouse CA XV was carried out using computer modelling. The results are illustrated in Figure 4. Panels (a) and (b) show the surfaces of mouse CA XV and human CA IV coloured according to their hydrophobicity profiles. The green sphere represents the zinc atom in the active site of the enzyme. The pictures illustrate that the active site is conserved in mouse CA XV. In panel (c), all the cysteine residues are shown. In the molecular model the distances between cysteine pairs 26/34, 46/242 and 103/107 (numbering according to Figure 1) are compatible with disulphide-bridge formation, even though the disulphides were not forced in our model. The C-terminal part of the molecule, which contains a cysteine residue at position 326, is cleaved off in the process of GPI anchoring and therefore is not available to form a disulphide bond with the cysteines present in the mature molecule. The locations of the asparagine residues that were predicted to be glycosylated are pointed out in panel (d). All of these residues are predicted to be located on the surface of the molecule, and thus they have the potential to have oligosaccharide chains attached.

CA15 mRNA expression in murine tissues

In mouse tissues the most abundant *CA15* mRNA expression was found in the kidney. Sequencing confirmed that the positive band detected by PCR corresponded to the correct amplification product. The PCR results in the kidney and testes are shown

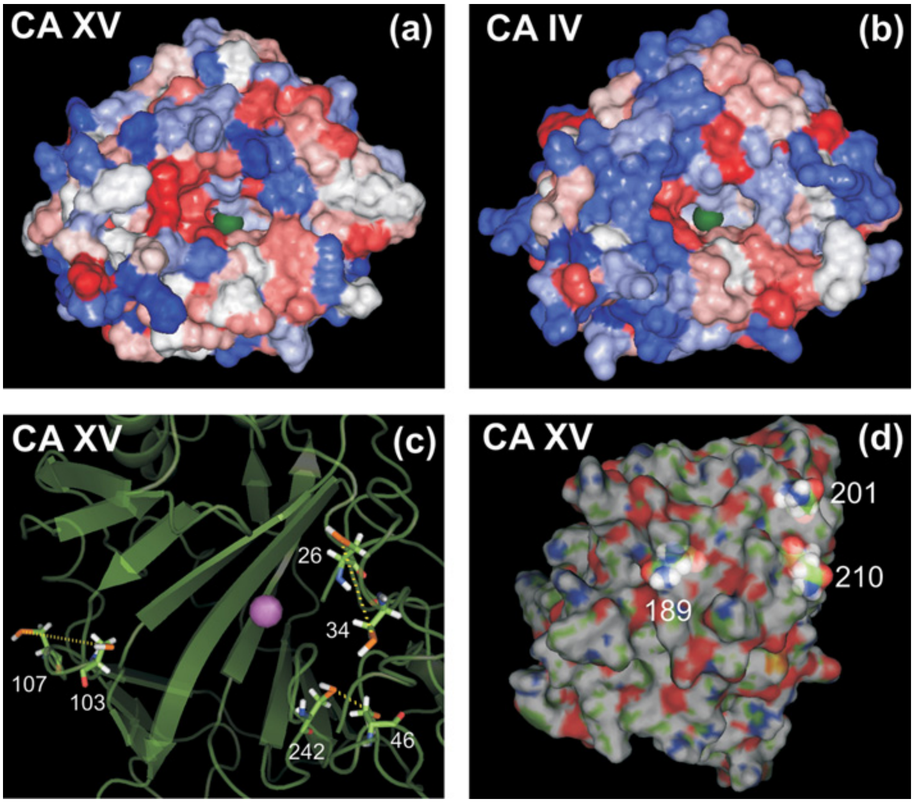

Figure 4 The prediction of the structure for CA XV

Panels (**a**) and (**b**): comparison of the surfaces of CA XV and CA IV, coloured according to hydropathy (blue represents hydrophilic, and red hydrophobic). The green sphere represents the zinc atom crucial for CA activity. Panel (**c**): secondary structures and cysteine pairs. The zinc atom is shown by a purple sphere. Panel (**d**): highlighted asparagines predicted to represent glycosylation sites; residue numbers are the same as shown as in Figure 1.

© 2005 Biochemical Society

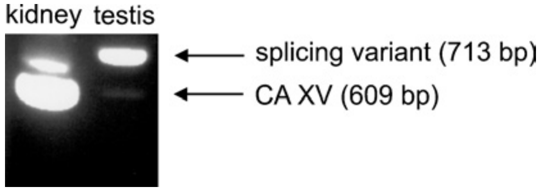

Figure 5 Results of RT-PCR

The results of the kidney and testis RT-PCR are shown in the Figure. The kidney has a strong band for CA XV (609 bp), whereas in the testis the band is much weaker. In testis, however, there is a stronger band for a splicing variant (713 bp). The rest of the results are summarized in Table 1.

Table 1 Expression of *CA15* mRNA in mouse tissues

Scores in RT-PCR (denoted by the band intensity) were as follows: +, strong band; +/− weak band; − no band observed.

Tissue	Band intensity
Heart	−
Skeletal muscle	−
Brain	+/−
Kidney	+
Liver	−
Lung	−
Spleen	−
Testis	+/−
Stomach	−
Duodenum	−
Jejunum	−
Ileum	−
Colon	−
Blood	−
7-Day embryo	+/−
11-Day embryo	−
15-Day embryo	−
17-Day embryo	+/−

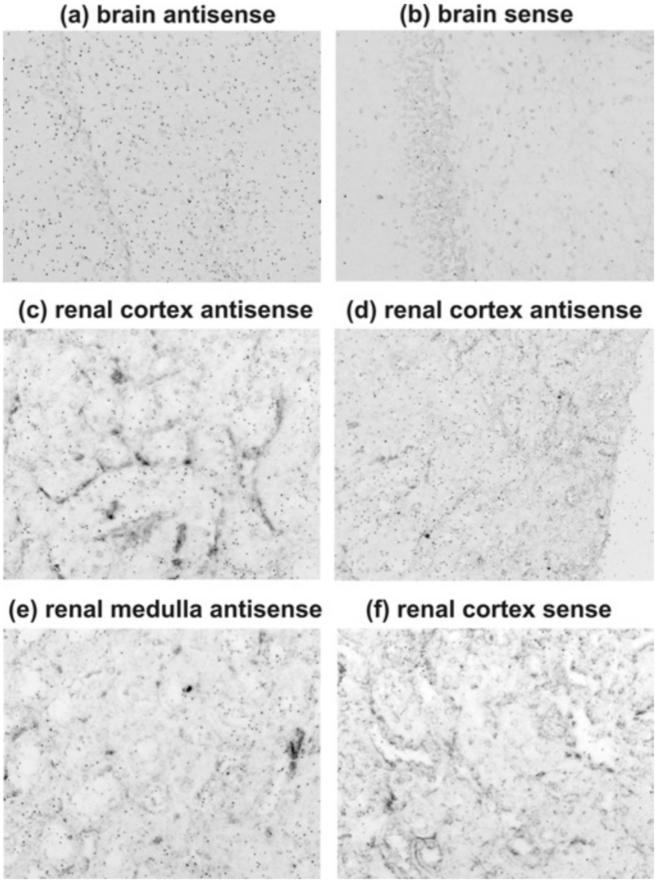
(a) brain antisense
(b) brain sense
(c) renal cortex antisense
(d) renal cortex antisense
(e) renal medulla antisense
(f) renal cortex sense

Figure 6 *In situ* hybridization revealing *CA15* mRNA in mouse tissues

The signal was present in brain (**a**), whereas the negative control had a much lower signal density (**b**). High expression was observed in the renal cortex (**c, d**) and lower expression in the medulla (**e**). The sense control shows the background level of the signal (**f**). Original magnifications of the panels: **a**, **b**, **c**, **e** and **f**, × 400; **d**, × 200.

in Figure 5. The rest of the results are summarized in Table 1. The kidney showed the strongest PCR reaction; weak bands were observed in the brain, testes, 7-day-old embryo and 17-day-old embryo. Testes contained an additional band, which was 104 bp longer than the expected *CA15* product. This PCR product was also sequenced, and it appeared to represent a *CA15* splicing variant which contains a longer second exon, i.e. 104 nt of the following intron are included in this exon, which causes a frameshift and thus results in a stop codon near the beginning of the third exon. Therefore this splicing variant would not produce a functional CA.

In addition to RT-PCR, *in situ* hybridization was also performed to reveal the murine tissues that express *CA15* mRNA. In accordance with the RT-PCR results, the kidney and brain showed a positive signal (Figure 6). The signal density varied within the renal parenchyma such that high expression was observed in the renal cortex and lower expression in the medulla.

The following human tissues were screened for the expression of three *CA15* candidate genes: kidney, heart, lung, brain, pancreas, spleen, thymus, small intestine, colon, skeletal muscle, prostate, testis, ovary, placenta and peripheral blood leukocytes. All the human tissues remained negative in RT-PCR experiments using a primer pair designed for detection of all the candidate genes (results not shown). In addition, gene-specific primers for the putative *CA15* genes were used to re-examine the kidney, brain and heart samples, and also these results remained negative (results not shown). Taken together, these findings and the sequence data indicate that all three *CA15* candidate genes are pseudogenes in the human genome.

Expression of CA XV in COS-7 cells

CA XV was expressed in COS-7 cells transfected with mouse *CA15* cDNA (Figure 7a). The CA XV polypeptides migrated on SDS/PAGE with an apparent molecular mass of 34–36 kDa. For a positive control, mouse CA IV, a glycosylated GPI-anchored protein, was also expressed in COS-7 cells. The results in Figure 7(b) show that the majority of the CA XV protein was associated with the membrane fraction, suggesting a membrane localization for the enzyme. The affinity of the CA XV enzyme towards p-AMBS was studied using p-AMBS Affigel resin. The results in Figure 7(c) show that 9 % of the total membrane-associated mouse CA XV and 50 % of the mouse CA IV were retained on the affinity resin. This difference could be due to the presence of an extra pair of thiol groups in CA XV, and/or weak retention of CA XV over p-AMBS affinity resin. Part of this hypothesis has been tested in the latter part of the present study.

Mouse CA XV is a glycosylated and GPI-anchored protein

Figure 8(a) shows that CA XV without EndoH treatment contains one major polypeptide of 36 kDa and three minor polypeptides

© 2005 Biochemical Society

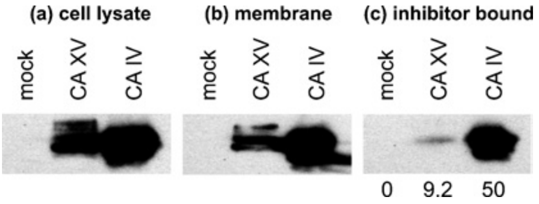

Figure 7 Expression of functional mouse CA XV in the membrane of transfected COS-7 cells

(**a**) Total cell lysates of non-transfected COS-7 cell (mock) and transfected COS-7 cells (with *CA15* or *CA4* cDNAs) were analysed by Western blotting using anti-mouse CA IV antibodies. (**b**) The cell lysates were centrifuged at 100 000 ***g*** for 30 min. The membranes were analysed by Western blotting. (**c**) The membrane extracts in 10 mM Hepes/NaOH, pH 7.5, containing 1 % NP40 and protease inhibitors were incubated with CA inhibitor-affinity resin, and bound enzyme was analysed by Western blotting. The numbers at the bottom of the gel show percentages of inhibitor-bound CA XV or CA IV.

of 34, 32 and 29 kDa respectively. Upon EndoH treatment, the higher-apparent-molecular-mass forms of CA XV were reduced to 29 kDa. These results suggested that expression of CA XV results in glycosylated (36, 34 and 32 kDa) and non-glycosylated (29 kDa) forms, and also that the fully glycosylated mouse CA XV contains at least three N-linked oligosaccharides. This result is in accordance with the bioinformatics predictions. Mouse CA IV contains two N-linked oligosaccharides, as was evident from one intermediate polypeptide between 34 and 29 kDa. The results in Figure 8(b) show that, after treatment with PI-PLC, the membrane pellet contained little residual CA XV compared with non-treated membranes, indicating that the extracellular domain of CA XV can be sheared from the COS-7 cell membrane by PI-PLC. The mouse CA IV was similarly removed from the cell membrane by PI-PLC treatment. From these results, we conclude that CA XV, like CA IV, is a GPI-anchored protein.

Mouse CA XV forms disulphide-bond-linked, high-molecular-mass quaternary structures

CA XV forms several high-molecular-mass aggregates in COS-7 cells under non-reducing conditions (Figure 8c). However, in the presence of a reducing compound, DTT (dithiothreitol), CA XV migrated as a diffuse 29–34 kDa protein band. These results suggested that CA XV in COS-7 cells could form disulphide-bond-linked aggregates. When affinity-purified mouse CA XV from *E. coli* was analysed on SDS/PAGE in the absence and presence of DTT, CA XV migrated as 66 kDa disulphide-linked dimer, which turned into a 31 kDa monomer upon reduction with DTT. Proteolytically cleaved fragments of 14–16 and 21 kDa were also seen (Figure 8d).

Functional activity and enhancement by chemical chaperones

The results in Figure 9(a) show that the cell lysates expressing mouse CA XV contained a detectable amount of CA activity, which was enriched in membrane pellets. Mouse CA IV, a positive control, showed significantly higher activity than CA XV in the membrane suspension and cell lysates. Since CA XV contains an additional pair of thiol residues, a slower or less perfect folding or a stronger tendency to form aggregates may reduce the activity of the enzyme. This can especially occur in a cell culture overexpressing the protein. When the membrane lysates were treated with 10 mM GSSG to facilitate rearrangement of disulphide bonds, mouse CA XV showed significantly higher enzymatic activity, due to deaggregation and/or refolding. Similarly, mouse CA IV activity was also enhanced, but not as much as CA XV. The results in Figure 9(b) show that the enzymatic activity of CA XV was also significantly higher when the transfected COS-7 cells were incubated with 2 or 4 mM PBA and 10 or 20 μM dorzolamide. Mouse CA IV activity was also increased by PBA and dorzolamide treatment. However, the enhancement in CA IV activity by dorzolamide was lower than that seen for CA XV. This difference could be due to stronger binding of dorzolamide with CA IV, which prevented it from being completely removed by washing before the CA assay.

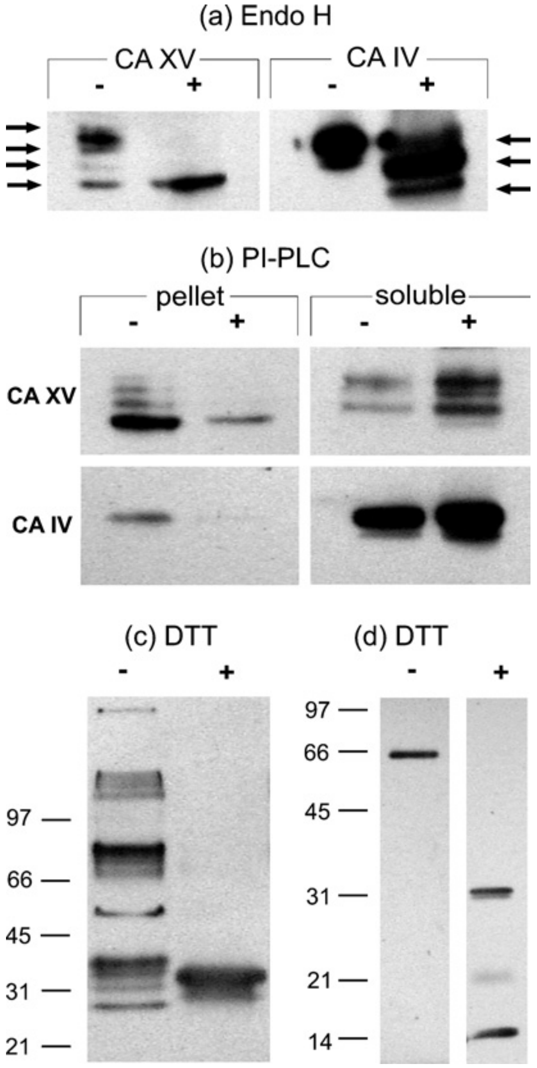

Figure 8 Biochemical properties of mouse CA XV

(**a**) EndoH treatment resulted in shifts of high-molecular-mass polypeptide of CA XV and CA IV. The CA XV showed two intermediate polypeptides, whereas CA IV showed only one intermediate polypeptide, as indicated by arrows. (**b**) COS-7 cell membranes were treated without (−) and with (+) PI-PLC enzyme. The enzymes remaining in the membrane pellet and soluble fraction after PI-PLC treatment were analysed by Western blotting. The decrease in the membrane-associated CA XV was compensated by an increase in the soluble fraction after PI-PLC treatment, suggesting that CA XV is a GPI-anchored protein. (**c**) COS-7 cell lysates expressing mouse CA XV were analysed by Western blotting under non-reducing (−) and reducing (+) conditions. Under non-reducing conditions, the mouse CA XV formed high-molecular-mass aggregates, which upon reducing with DTT resulted in a single immunoreactive polypeptide. (**d**) Affinity pure recombinant mouse CA XV from *E. coli* was treated with non-reducing (−) and reducing (+) sample buffer and analysed by SDS/PAGE, followed by Western blotting.

Recombinant mouse CA XV from *E. coli* is functionally active

Since CA XV from the COS- 7 cell extract showed CA activity that was increased by treatment with oxidized glutathione (Figure 10),

© 2005 Biochemical Society

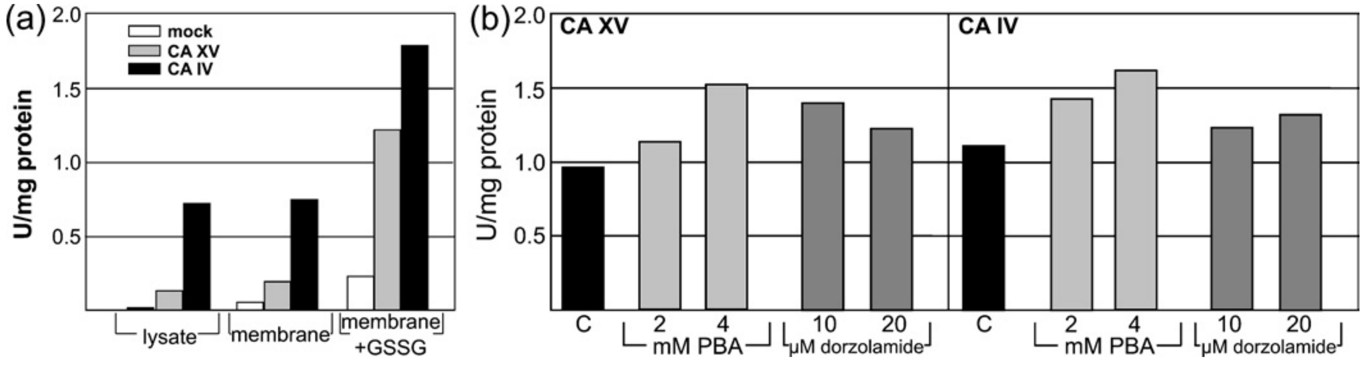

Figure 9 Refolding of CA XV by oxidized glutathione and chemical chaperones

(**a**) The cell lysates of COS-7 cells transfected with *CA15* or *CA4* cDNAs and untransfected cells (mock) were analysed for CA activity. The membrane suspensions from each cell line were also used for CA activity measurement. The results suggested that the cell membranes contained detectable CA XV and IV activities. Upon treatment with 10 mM GSSG, both CA XV and CA IV were refolded further into more active enzymes. (**b**) The COS-7 cells, after transfection with *CA15* or *CA4* cDNAs, were treated with 2 and 4 mM PBA or 10 and 20 μM dorzolamide for 72 h. After removing the PBA or dorzolamide, the cell membranes were used for CA activity measurement. Both non-specific (PBA) and specific (dorzolamide) chemical chaperones were able to refold the mouse CA XV and IV into more active enzymes. Abbreviation: U, units.

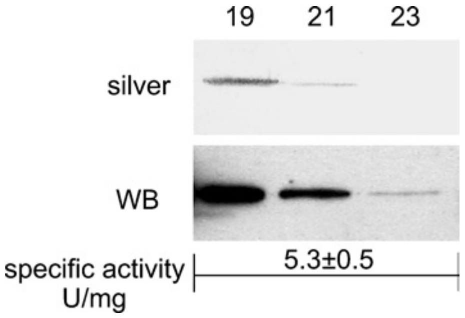

Figure 10 Recombinant affinity-purified mouse CA XV from *E. coli*

The fractions from the Sephacryl S-300 sizing column were analysed by SDS/PAGE. The polypeptides were visualized by silver staining (silver) or Western blot analysis (WB). The numbers show the fractions used for SDS/PAGE. The homogeneous pooled enzyme showed a specific activity of 5.3 ± 0.5 units/mg.

we expressed recombinant mouse CA XV in *E. coli*. The CA XV expressed in *E. coli* was refolded with oxidized glutathione and purified to homogeneity using a combination of CA-inhibitor affinity resin and sizing column chromatography over Sephacyl S-300. The results in Figure 10 show that homogeneous mouse CA XV from *E. coli* is functionally active, showing specific activity of 5.3 ± 0.5 units/mg of enzyme. By way of comparison, the high-activity enzymes such as human CA II and CA IV show activities between 2000–3000 units/mg [32]. The activity of CA XV is in the same range as that of CA III (1–5 units/mg).

DISCUSSION

New tools of bioinformatics have allowed genome-wide analyses in which novel genes can be efficiently identified and aligned with other homologous counterparts. The present study was initiated when we found a new CA-related gene, named *CA15*, in the databases (submitted by Hewett-Emmett and Shimmin in 2000; see GenBank® accession no. AF231122). Sequence comparisons indicated that the *CA15* gene can be found in genomes of several species. Interestingly, three copies of *CA15* genes were identified in the human chromosome band 22q11.21. In the chimpanzee genome, we found two copies, and it is possible that one copy of the gene is missing, due to incomplete genomic data. We concluded that, in both species, all the *CA15* genes represent pseudogenes, because of frame-shifts, insertions, point mutations and the lack of mRNAs and EST sequences. In contrast, all the other genomes exhibited only single *CA15* genes, which apparently encode catalytic CA XV enzymes with conserved active-site residues. This prediction was verified in mice by studies reported here. The full-length murine cDNA produced enzymatically active CA XV in COS-7 cells. Thus CA XV is the only active CA isoenzyme thus far known that is expressed in several vertebrate species, but has been lost in humans and chimpanzees. This is a novel and unique observation in terms of evolution. In a recent paper, only 27 genes that are active in rodents have been reported to become pseudogenes in humans and chimpanzees; *CA15* was not among the reported genes [33].

CAs are ancient enzymes [34]. The efficient regulation of acid–base balance mediated by CAs appears to be extremely important in biological processes, since numerous isoenzymes of this family catalyse the same fundamental reaction in different organs, tissues and cell compartments. Prior to this example, 12 active isoenzymes had been characterized in humans and mice, and all are highly conserved in evolution. CA XV is the first CA isoenzyme that is expressed in many other vertebrates, but not in chimpanzees and humans. Conservation of all key residues and a well-conserved intron–exon structure suggest that CA XV is functionally important in these non-primate vertebrates. Perhaps functional redundancy may explain why it was lost in chimpanzees and humans.

Phylogenetic studies suggested that CA XV and CA IV are closely related isoenzymes, which share distinct sequence similarity. The relatedness of CA XV and CA IV was supported by several experimental results. Until now, CA IV was a unique member of the CA family because of its GPI anchorage to cell membranes. CA XV was shown to be similarly anchored to cell membranes by a GPI anchor. Two disulphide linkages were shown to be critical to stabilize the conformation of the extracellular CA IV protein [19]. The experiments using oxidized glutathione reported here showed that the correct formation of disulphide bridges is also important for maturation of CA XV. Finally, like murine CA IV, CA XV is an N-glycosylated protein.

The expression of CA IV has been studied most intensively in humans and rats. In both species, CA IV is expressed in a variety of tissues: in kidney, CA IV is present mainly on the apical brush border membrane of the proximal tubular cells and on the cells of

© 2005 Biochemical Society

the thick ascending limbs of Henle, where its physiological role is to facilitate bicarbonate reabsorption [27,35]. In lung, CA IV is localized on the luminal surface of pulmonary endothelial cells, where it catalyses the dehydration of bicarbonate in the serum to yield CO_2 [27,36]. CA IV is localized in the capillary endothelium of skeletal and heart muscle, and in the latter it can be also found in special sarcolemmal structures and sarcoplasmic reticulum [37,38]. In distal small and large intestine, CA IV participates in ion and fluid transport [5]. CA IV participates in acidification of epididymal fluid, and is also expressed in the capillary endothelial cells of brain [39]. In addition, CA IV has been reported to be expressed in human pancreas, salivary glands [40], gall-bladder epithelium [41], choriocapillaris of the eye [42] and erythrocytes [43]. CA IV has been shown to form physical complexes with chloride/bicarbonate exchange proteins, and therefore it facilitates the rate of bicarbonate transportation [44]. CA IV is also crucial to the function of NBC1 (the Na^+–HCO_3^- co-transporter) [45]. Recently, an apoptosis-inducing mutation has been identified in the signal sequence of *CA4* gene that causes an RP17 form of retinitis pigmentosa [12,46].

The similarity in properties of CA IV and CA XV raises the question of whether CA IV makes CA XV redundant and dispensable in chimpanzees and humans. For instance, the reabsorption of bicarbonate in the kidney is a major process that may require the assistance of CA XV in mice. Mice have a relatively low activity of CA IV. The high activity of human CA IV may have made CA XV redundant in humans, and explains why this gene has become a pseudogene in the course of evolution. In further studies, it would be interesting to study the expression of these isoenzymes in several species in order to gain a better understanding of their roles in normal physiology. It would be attractive to develop single knockout mice for CA IV and CA XV, as well as double knockout mice in which both of these enzyme activities are missing, in order to deepen our knowledge of the physiological importance of these two GPI-anchored enzymes in physiology.

This work was supported by grants from the Finnish Cultural Foundation, Sigrid Juselius Foundation, Academy of Finland and National Institutes of Health (GM34182 and DK40163). The use of computing and software resources provided by CSC, the Finnish IT Center for Science, partly funded by the Finnish Ministry of Education, is gratefully acknowledged. We also thank Dr Markku Heikinheimo for advising with *in situ* hybridization.

REFERENCES

1 Lindskog, S. and Silverman, D. N. (2000) The catalytic mechanism of mammalian carbonic anhydrases. In The Carbonic Anhydrases: New Horizons (Chegwidden, W. R., Carter, N. D. and Edwards, Y. H., eds.), pp. 175–195, Birkhäuser Verlag, Basel

2 Sly, W. S. and Hu, P. Y. (1995) Human carbonic anhydrases and carbonic anhydrase deficiencies. Annu. Rev. Biochem. **64**, 375–401

3 Hewett-Emmett, D. (2000) Evolution and distribution of the carbonic anhydrase gene families. In The Carbonic Anhydrases: New Horizons (Chegwidden, W. R., Carter, N. D. and Edwards, Y. H., eds.), pp. 29–76, Birhkhäuser Verlag, Basel

4 Lehtonen, J., Shen, B., Vihinen, M., Casini, A., Scozzafava, A., Supuran, C. T., Parkkila, A. K., Saarnio, J., Kivela, A. J., Waheed, A. et al. (2004) Characterization of CA XIII, a novel member of the carbonic anhydrase isoenzyme family. J. Biol. Chem. **279**, 2719–2727

5 Fleming, R. E., Parkkila, S., Parkkila, A. K., Rajaniemi, H., Waheed, A. and Sly, W. S. (1995) Carbonic anhydrase IV expression in rat and human gastrointestinal tract regional, cellular, and subcellular localization. J. Clin. Invest. **96**, 2907–2913

6 Pastorekova, S., Parkkila, S., Parkkila, A. K., Opavsky, R., Zelnik, V., Saarnio, J. and Pastorek, J. (1997) Carbonic anhydrase IX, MN/CA IX: analysis of stomach complementary DNA sequence and expression in human and rat alimentary tracts. Gastroenterology **112**, 398–408

7 Kyllonen, M. S., Parkkila, S., Rajaniemi, H., Waheed, A., Grubb, J. H., Shah, G. N., Sly, W. S. and Kaunisto, K. (2003) Localization of carbonic anhydrase XII to the basolateral membrane of H^+-secreting cells of mouse and rat kidney. J. Histochem. Cytochem. **51**, 1217–1224

8 Parkkila, S., Parkkila, A. K., Rajaniemi, H., Shah, G. N., Grubb, J. H., Waheed, A. and Sly, W. S. (2001) Expression of membrane-associated carbonic anhydrase XIV on neurons and axons in mouse and human brain. Proc. Natl. Acad. Sci. U.S.A. **98**, 1918–1923

9 Leinonen, J., Parkkila, S., Kaunisto, K., Koivunen, P. and Rajaniemi, H. (2001) Secretion of carbonic anhydrase isoenzyme VI (CA VI) from human and rat lingual serous von Ebner's glands. J. Histochem. Cytochem. **49**, 657–662

10 Shah, G. N., Hewett-Emmett, D., Grubb, J. H., Migas, M. C., Fleming, R. E., Waheed, A. and Sly, W. S. (2000) Mitochondrial carbonic anhydrase CA VB: differences in tissue distribution and pattern of evolution from those of CA VA suggest distinct physiological roles. Proc. Natl. Acad. Sci. U.S.A. **97**, 1677–1682

11 Nishimori, I. (2004) Acatalytic CAs: Carbonic anhydrase-related proteins. In Carbonic Anhydrase: Its Inhibitors and Activators (Supuran, C. T., Scozzafava, A. and Conway, J., eds.), pp. 25–43, CRC Press, Boca Raton

12 Bonapace, G., Waheed, A., Shah, G. N. and Sly, W. S. (2004) Chemical chaperones protect from effects of apoptosis-inducing mutation in carbonic anhydrase IV identified in retinitis pigmentosa 17. Proc. Natl. Acad. Sci. U.S.A. **101**, 12300–12305

13 Kent, W. J. (2002) BLAT – the BLAST-like alignment tool. Genome Res. **12**, 656–664

14 Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. Genome Res. **14**, 988–995

15 Notredame, C., Higgins, D. G. and Heringa, J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**, 205–217

16 Reference deleted

17 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673–4680

18 Swofford, D. L. (2002) PAUP*, Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4. Sinauer Associates, Sunderland, MA

19 Stams, T., Nair, S. K., Okuyama, T., Waheed, A., Sly, W. S. and Christianson, D. W. (1996) Crystal structure of the secretory form of membrane-associated human carbonic anhydrase IV at 2.8-Å resolution. Proc. Natl. Acad. Sci. U.S.A. **93**, 13589–13594

20 Heikinheimo, M., Scandrett, J. M. and Wilson, D. B. (1994) Localization of transcription factor GATA-4 to regions of the mouse embryo involved in cardiac development. Dev. Biol. **164**, 361–373

21 Whittington, D. A., Grubb, J. H., Waheed, A., Shah, G. N., Sly, W. S. and Christianson, D. W. (2004) Expression, assay, and structure of the extracellular domain of murine carbonic anhydrase XIV: implications for selective inhibition of membrane-associated isoenzymes. J. Biol. Chem. **279**, 7223–7228

22 Lopata, M. A., Cleveland, D. W. and Sollner-Webb, B. (1984) High level transient expression of a chloramphenicol acetyl transferase gene by DEAE-dextran mediated DNA transfection coupled with a dimethyl sulfoxide or glycerol shock treatment. Nucleic Acids Res. **12**, 5707–5717

23 Luthman, H. and Magnusson, G. (1983) High efficiency polyoma DNA transfection of chloroquine treated cells. Nucleic Acids Res. **11**, 1295–1308

24 Lowry, O. H., Rosebrough, N. J., Farr, A. L. and Randall, R. J. (1951) Protein measurement with the Folin phenol reagent. J. Biol. Chem. **193**, 265–275

25 Sundaram, V., Rumbolo, P., Grubb, J., Strisciuglio, P. and Sly, W. S. (1986) Carbonic anhydrase II deficiency: diagnosis and carrier detection using differential enzyme inhibition and inactivation. Am. J. Hum. Genet. **38**, 125–136

26 Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature (London) **227**, 680–685

27 Zhu, X. L. and Sly, W. S. (1990) Carbonic anhydrase IV from human lung. Purification, characterization, and comparison with membrane carbonic anhydrase from human kidney. J. Biol. Chem. **265**, 8795–8801

28 Waheed, A., Parkkila, S., Zhou, X. Y., Tomatsu, S., Tsuchihashi, Z., Feder, J. N., Schatzman, R. C., Britton, R. S., Bacon, B. R. and Sly, W. S. (1997) Hereditary hemochromatosis: effects of C282Y and H63D mutations on association with beta2-microglobulin, intracellular processing, and cell surface expression of the HFE protein in COS-7 cells. Proc. Natl. Acad. Sci. U.S.A. **94**, 12384–12389

29 Waheed, A., Zhu, X. L. and Sly, W. S. (1992) Membrane-associated carbonic anhydrase from rat lung. Purification, characterization, tissue distribution, and comparison with carbonic anhydrase IVs of other mammals. J. Biol. Chem. **267**, 3308–3311

30 Waheed, A., Pham, T., Won, M., Okuyama, T. and Sly, W. S. (1997) Human carbonic anhydrase IV: *in vitro* activation and purification of disulfide-bonded enzyme following expression in Escherichia coli. Protein Expr. Purif. **9**, 279–287

31 Hu, P. Y., Waheed, A. and Sly, W. S. (1995) Partial rescue of human carbonic anhydrase II frameshift mutation by ribosomal frameshift. Proc. Natl. Acad. Sci. U.S.A. **92**, 2136–2140

32 Karhumaa, P., Parkkila, S., Waheed, A., Parkkila, A. K., Kaunisto, K., Tucker, P. W., Huang, C. J., Sly, W. S. and Rajaniemi, H. (2000) Nuclear NonO/p54(nrb) protein is a nonclassical carbonic anhydrase. J. Biol. Chem. **275**, 16044–16049

33 International Human Genome Sequencing Consortiuim (2004) Finishing the euchromatic sequence of the human genome. Nature (London) **431**, 931–945

© 2005 Biochemical Society

34 Tripp, B. C., Smith, K. and Ferry, J. G. (2001) Carbonic anhydrase: new insights for an ancient enzyme. J. Biol. Chem. **276**, 48615–48618

35 Brown, D., Zhu, X. L. and Sly, W. S. (1990) Localization of membrane-associated carbonic anhydrase type IV in kidney epithelial cells. Proc. Natl. Acad. Sci. U.S.A. **87**, 7457–7461

36 Fleming, R. E., Crouch, E. C., Ruzicka, C. A. and Sly, W. S. (1993) Pulmonary carbonic anhydrase IV: developmental regulation and cell-specific expression in the capillary endothelium. Am. J. Physiol. **265**, L627–L635

37 Sender, S., Gros, G., Waheed, A., Hageman, G. S. and Sly, W. S. (1994) Immunohistochemical localization of carbonic anhydrase IV in capillaries of rat and human skeletal muscle. J. Histochem. Cytochem. **42**, 1229–1236

38 Sender, S., Decker, B., Fenske, C. D., Sly, W. S., Carter, N. D. and Gros, G. (1998) Localization of carbonic anhydrase IV in rat and human heart muscle. J. Histochem. Cytochem. **46**, 855–861

39 Ghandour, M. S., Langley, O. K., Zhu, X. L., Waheed, A. and Sly, W. S. (1992) Carbonic anhydrase IV on brain capillary endothelial cells: a marker associated with the blood-brain barrier. Proc. Natl. Acad. Sci. U.S.A. **89**, 6823–6827

40 Fujikawa-Adachi, K., Nishimori, I., Sakamoto, S., Morita, M., Onishi, S., Yonezawa, S. and Hollingsworth, M. A. (1999) Identification of carbonic anhydrase IV and VI mRNA expression in human pancreas and salivary glands. Pancreas **18**, 329–335

41 Parkkila, S., Parkkila, A. K., Juvonen, T., Waheed, A., Sly, W. S., Saarnio, J., Kaunisto, K., Kellokumpu, S. and Rajaniemi, H. (1996) Membrane-bound carbonic anhydrase IV is expressed in the luminal plasma membrane of the human gallbladder epithelium. Hepatology **24**, 1104–1108

42 Hageman, G. S., Zhu, X. L., Waheed, A. and Sly, W. S. (1991) Localization of carbonic anhydrase IV in a specific capillary bed of the human eye. Proc. Natl. Acad. Sci. U.S.A. **88**, 2716–2720

43 Wistrand, P. J., Carter, N. D., Conroy, C. W. and Mahieu, I. (1999) Carbonic anhydrase IV activity is localized on the exterior surface of human erythrocytes. Acta Physiol. Scand. **165**, 211–218

44 Sterling, D., Alvarez, B. V. and Casey, J. R. (2002) The extracellular component of a transport metabolon. Extracellular loop 4 of the human AE1 Cl^-/$HCO3^-$ exchanger binds carbonic anhydrase IV. J. Biol. Chem. **277**, 25239–25246

45 Alvarez, B. V., Loiselle, F. B., Supuran, C. T., Schwartz, G. J. and Casey, J. R. (2003) Direct extracellular interaction between carbonic anhydrase IV and the human NBC1 sodium/bicarbonate co-transporter. Biochemistry **42**, 12321–12329

46 Rebello, G., Ramesar, R., Vorster, A., Roberts, L., Ehrenreich, L., Oppon, E., Gama, D., Bardien, S., Greenberg, J., Bonapace, G. et al. (2004) Apoptosis-inducing signal sequence mutation in carbonic anhydrase IV identified in patients with the RP17 form of retinitis pigmentosa. Proc. Natl. Acad. Sci. U.S.A. **101**, 6617–6622

Received 11 July 2005/3 August 2005; accepted 8 August 2005
Published as BJ Immediate Publication 8 August 2005, doi:10.1042/BJ20051102

© 2005 Biochemical Society

Bioorganic & Medicinal Chemistry 21 (2013) 1503–1510

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Analysis of evolution of carbonic anhydrases IV and XV reveals a rich history of gene duplications and a new group of isozymes

Martti E. E. Tolvanen [a,*], Csaba Ortutay [a], Harlan R. Barker [a], Ashok Aspatwar [a,b], Maarit Patrikainen [a,b], Seppo Parkkila [a,b]

[a] *Institute of Biomedical Technology, University of Tampere, Finland and BioMediTech, FI-33014 Tampere, Finland*
[b] *Fimlab Laboratories Ltd, Tampere, Finland*

ARTICLE INFO

Article history:
Available online 10 September 2012

Keywords:
Bioinformatics
Phylogenetics
GPI anchor
N-glycosylation
Carbonic anhydrase 17

ABSTRACT

Carbonic anhydrase (CA) isozymes CA IV and CA XV are anchored on the extracellular cell surface via glycosylphosphatidylinositol (GPI) linkage. Analysis of evolution of these isozymes in vertebrates reveals an additional group of GPI-linked CAs, CA XVII, which has been lost in mammals. Our work resolves nomenclature issues in GPI-linked fish CAs. Review of expression data brings forth previously unreported tissue and cancer types in which human CA IV is expressed. Analysis of collective glycosylation patterns of GPI-linked CAs suggests functionally important regions on the protein surface.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Our present study focuses on the glycosylphosphatidylinositol (GPI) linked isozymes of carbonic anhydrase (CA) in vertebrates. This type of linkage is formed during the secretory pathway of certain proteins. A hydrophobic sequence with specific features in the C-terminus of the protein is recognized by the GPI transamidase complex, which exchanges the C-terminal peptide for the glycolipid anchor. The anchor typically consists of ethanolamine, phosphate, three mannose residues, one glucosamine, and a phosphatidylinositol lipid, with various optional additions. GPI anchors, their biosynthesis, and mechanism of attachment have been reviewed in great detail by Orlean and Menon.[1]

GPI-anchored proteins and the machinery to make them are found in all eukaryotes.[1] This particular linkage leads to membrane proteins which are devoid of a transmembrane domain, and only attached to one leaflet of the membrane. Similar to other glycolipids, glycosylphosphatidylinositol is preferentially located in specific cholesterol-enriched membrane microdomains known as lipid rafts.[2] As a consequence, GPI-linked proteins have the same special distribution on the membrane and share the same specific, non-clathrin mediated, endocytosis pathway.

There is only one GPI-linked CA known in human, CA IV.[3,4] For a long time, homologs to CA IV remained as the only known GPI-linked CAs in mammals, but related isozymes were found in crab,[5] in fish,[6] and in yellow fever mosquito.[7]

In addition to the normal CA function in pH control, CA IV has been shown to facilitate ion transport in many contexts. These include catalytic assistance to AE1 in erythrocytes,[8] to MCT1 in skeletal muscle,[9] to AE3 in neurons,[10] and to NBC1 in the choriocapillaris;[11] and non-catalytic assistance to MCT2.[12] CA IV also facilitates MCT-mediated acid/base flux in the astrocytes of rat brain.[13,14]

CA4 has been known to be expressed at low level in nearly all tissues, and highly expressed in kidney, heart, colon, and lung.[3,15–18] Regarding lung expression, it is of interest to note that similar GPI-anchored isozymes are also found in the gills of crabs[5,19] and fishes.[6,20,21] A number of studies have shown anatomical details of CA IV expression in male reproductive tissues,[22] in gallbladder,[23] and in the eye.[24] Recently CA IV was identified as the principal CO_2 taste sensor in sour-sensing taste cells.[25]

Several mutations in *CA4* are associated with eye disease Retinitis Pigmentosa 17 (RP17).[26–28] The pathogenetic mechanism is characterized by apoptosis, for which there are two alternative explanations that may work concurrently: unfolded protein stress,[29] and disturbed pH homeostasis caused by lack of CA IV on cell surface in the choriocapillaris.[11] In another eye disease, glaucoma, therapy with CA inhibitors has been used for nearly 60 years.[30] At the time when this approach was started there was no knowledge of the variety of CA isozymes, but we know now that CA IV is one of the target CAs in the eye, besides CA II and CA XII.[31]

Several three-dimensional structural models of CA IV have been elucidated, often in cocrystals with inhibitors.[32–34] The structure with the highest resolution is called 3FW3 in PDB, at 1.72 Å.[34]

[*] Corresponding author. Tel.: +358 407671053.
E-mail address: martti.tolvanen@uta.fi (M.E.E. Tolvanen).

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.08.060

In 2005, Hilvo et al.[35] published the discovery of the second mammalian GPI-linked isozyme, CA XV, which was shown to be present only as one or more non-functional pseudogenes in human, chimpanzee, and macaque genomes, and assumed to have been lost in the primate lineage. RT-PCR and in situ hybridization showed high expression of *car15* in the kidney and weak expression in the brain and in testis (RT-PCR only).[35] Immunohistochemistry detects CA XV in mouse kidney only, with a distribution similar to that of human CA IV.[36] The inhibition and activation properties of purified mouse CA XV have been investigated in several studies, reviewed by Hilvo et al.[37]

With the explosion of genome data in the last 10 years, more and more GPI-linked proteins, including CAs, have been detected by bioinformatic methods. This includes GPI-anchored CAs in the malaria mosquito[7] and Drosophila.[38] See also Syrjänen et al., this issue, pp. xxx–xxx.

Contrary to mammals, fish genomes have many more than just one or two GPI-anchored CA isozymes. Our database searches and analyses have revealed 6–9 GPI-linked isozymes in each of the six fish genomes which are available in the Ensembl database.[39] There is a general confusion in their annotation. Even one specific study of zebrafish CAs, and their phylogenetics,[20] lacks some isozymes, and authors arrive at some classifications and gene names with which we cannot agree. Our phylogenetic studies, performed in a wider context than previous studies, have shown us that GPI-linked CAs differentiated into *CA4* and *CA15* in early vertebrates, and thereafter the lineage of *CA4* has split into two major groups. Because of the early origin of the latter two groups, and complex duplications in both groups, we need two distinct CA designations: CA4 for the group containing the well-characterized mammalian *CA4* genes, and CA17 for the other group. Primarily, this paper aims at providing our analysis of the evolution of this group and our new suggestions for nomenclature, but we also present other bioinformatic analyses of these isozymes, and a review of previous results and existing expression data.

2. Materials and methods

2.1. Sequences

All available CA IV-like and CA XV protein sequences were retrieved from Ensembl[39] release 67 (www.ensembl.org, June 5th, 2012) starting from the Gene Tree for human CA4 (ENSG00000167434). We selected a parent node at the root of a subtree containing all vertebrate CA4 and CA15 orthologs and a group of four Ciona homologs, a total of 123 genes. An alignment of the corresponding proteins was created, and low-quality sequences were eliminated in several steps. First, all sequences with X characters to indicate missing regions, from supercontigs, were removed. Next, sequences with gaps of more than 20 residues within mainly gapless regions were removed. Then, sequences with insertions of more than 20 residues between conserved blocks were removed. After these steps, the alignment was remade with ClustalW[40] and, finally, sequences with gaps longer than 15 residues within conserved blocks were removed. Of the original 123 proteins this final selection contained 75 sequences. Four additional protein sequences from phylogenetically interesting and less represented groups were retrieved from the NCBI protein database[41] (http://www.ncbi.nlm.nih.gov/protein/ June 18th, 2012). Coding regions of the corresponding 79 transcripts were retrieved from Ensembl and NCBI. A full list of the sequences we used is provided in Supplementary data 1.

Sequences were not trimmed or corrected for phylogenetic analysis, but for signal peptide prediction, we evaluated the N-termini of all protein sequences for spuriously predicted translations which would seem to precede the most likely (consensus) starting methionine positions. Thirteen sequences were considered to have 1–27 N-terminal residues in excess, which were removed up to the next available initial methionine. Details of the trimmed sequences are also found in Supplementary data 1.

2.2. Sequence analyses

SignalP 4.0[42] at CBS, Denmark (http://www.cbs.dtu.dk/services/ Jun 27th 2012) was used to predict N-terminal secretion signal peptides in the protein sequences, trimmed as explained in Section 2.1. C-terminal glycosylphosphatidylinositol anchor attachment sites (GPI-anchor sites) were predicted in the Pred-GPI server[43] (http://gpcr.biocomp.unibo.it/predgpi/ Jun 27th 2012) with the general model and taking 'Highly probable' and 'Probable' predictions (99.5% specificity cutoff) as positive.

Potential N-glycosylation sites were analyzed in the NetNGlyc server (http://www.cbs.dtu.dk/services/NetNGlyc/, unpublished). The 75 Ensembl protein sequences (Section 2.1) were aligned with ClustalW.[40] The glycosylation sites, except sites with a sterically hindering Proline in the second position, were subsequently identified in the alignment and mapped to residue numbers in human CA IV structure 3FW3[34] from PDB.[44] In case of alignment matches to a gap in human CA IV, we mapped the site to the closest residue on the edge of the gap. The total number of sequences with a predicted glycosylation was tabulated for each position. A 3FW3-based glycosylation model was generated for CA IV, XV, and XVII protein groups (as defined by our phylogenetics results) individually, as well as for a composite of all three. Positions in CA IV matching these sites were colored by the frequency of glycosylation sites in the corresponding position in each collection of sequences. Side chain surface exposure and secondary structure environment of each position were evaluated. Molecular images were prepared with PyMol (PyMOL Molecular Graphics System, Version 1.51, Schrödinger, LLC).

2.3. Phylogenetics

A total of 79 selected CA IV-like and CA XV protein sequences were aligned by ClustalW.[40] PAL2NAL web server[45] was used to produce codon aligned cDNA sequences using the provided protein alignment as a guide. This alignment was used in MrBayes v 3.2[46] to estimate the phylogeny of the sequences using Bayesian inference. Bayesian estimation was run for 35,000 generations, with flat a priori distribution of base frequencies, substitution rates, proportion of invariable sites, and gamma shape parameter. The average standard deviation of split frequencies after 35,000 generations was 9.7×10^{-2} when the analysis was stopped. The arithmetic mean of the estimated marginal likelihoods for runs sampled was -55637.33. A 50% majority rule consensus tree was created and visualized using the APE R package.[47] The tree was rooted using the *Ciona intestinalis* monophyletic group as outgroup. Syntenic chromosome regions were viewed in Genomicus[48] (http://www.dyogen.ens.fr/genomicus-67.01/ June 16th 2012).

3. Results and discussion

3.1. Genomics and phylogenetics

3.1.1. Origin of GPI-anchored CAs goes back to fungi

GPI anchors are ancient, dating back to earliest eukaryotes. In case of CAs, GPI linkage has not been discussed in literature for anything more primitive than insects.[7,38] However, our studies have found very clearly predicted GPI attachment sites in alpha CAs of nematodes (*Caenorhabditis elegans* cah-5, NP_509186) and even in fungi (*Paracoccidioides brasiliensis* alpha CA, ACA28690.1). Thus far, we have not discovered any plant alpha CAs which would be predicted to be GPI-anchored, so it is reasonable to believe that the first GPI-linked carbonic anhydrases were formed early in the Fungi-Metazoa lineage.

3.1.2. Phylogenetics reveals a novel group of isozymes

Our phylogenetic analysis, as presented in the tree of Figure 1, indicates that the gene duplication which led into CA4 and CA15 groups of genes happened in very early vertebrates, before the radiation of jawed vertebrates. The CA4-like genes of two Agnathostomata species, lamprey (*Petromyzon marinus*) and hagfish (*Eptatretus stoutii*), are not associated with any of the branches in the Gnathostomata part of the tree. Slightly later, but still before the split of the tetrapod and fish lineages, *CA4* was duplicated to give rise to an additional isozyme gene, which we call *CA17*. Either one of these duplication events may have been related to whole-genome duplications in early vertebrate evolution. The evolutionary history, reflected in the branching pattern of the tree, and the fact that both isozymes coexist in many species in distinct

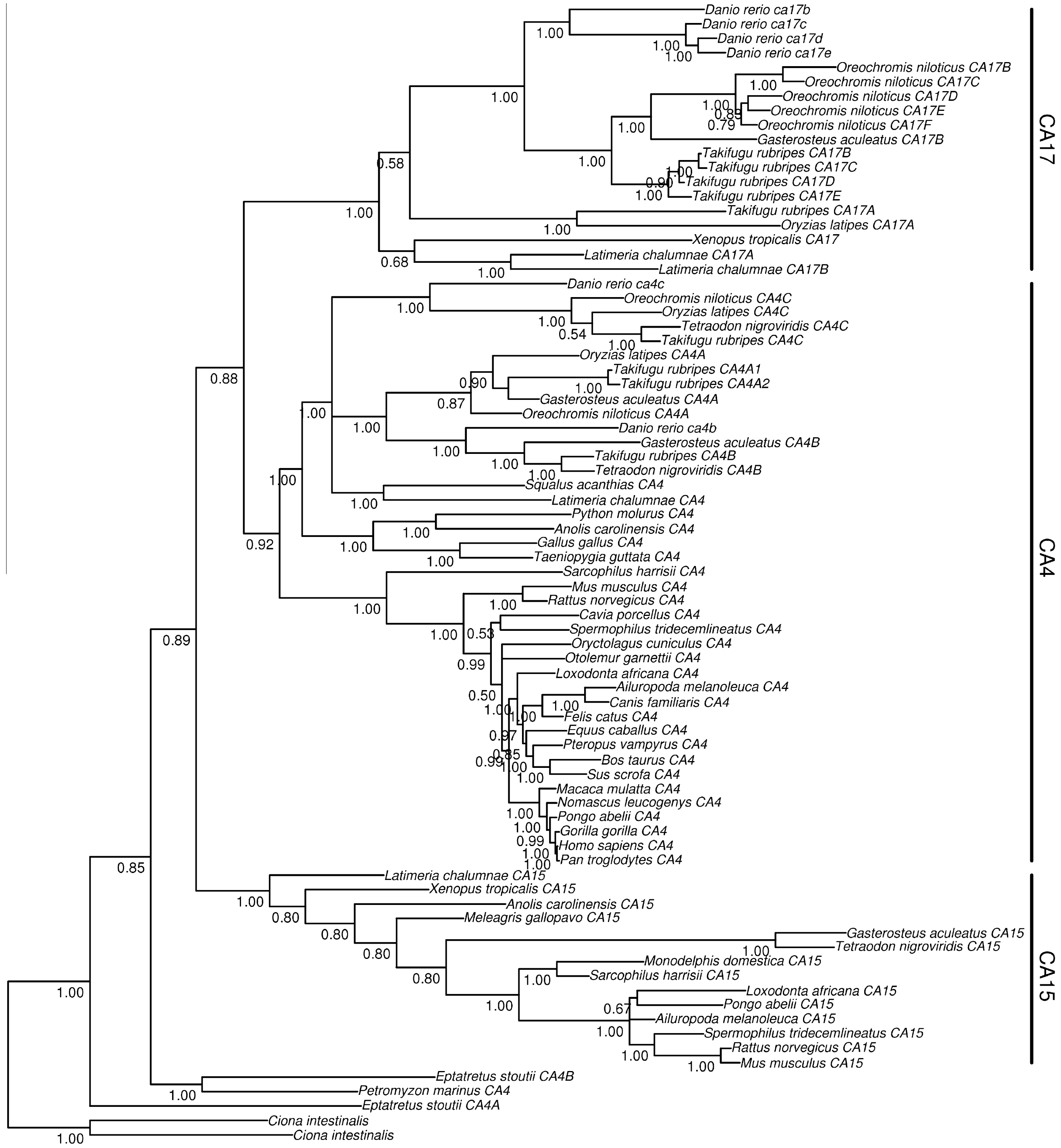

Figure 1. A phylogenetic tree of GPI-linked CAs. A Bayesian phylogenetic tree was created from an alignment of 79 selected proteins. See Section 2.3 for details. The division of CA groups, CA4, CA17, and CA15, is indicated by the vertical lines on the right. Branch lengths are proportional to distances, and the numbers at internal nodes are posterior probabilities.

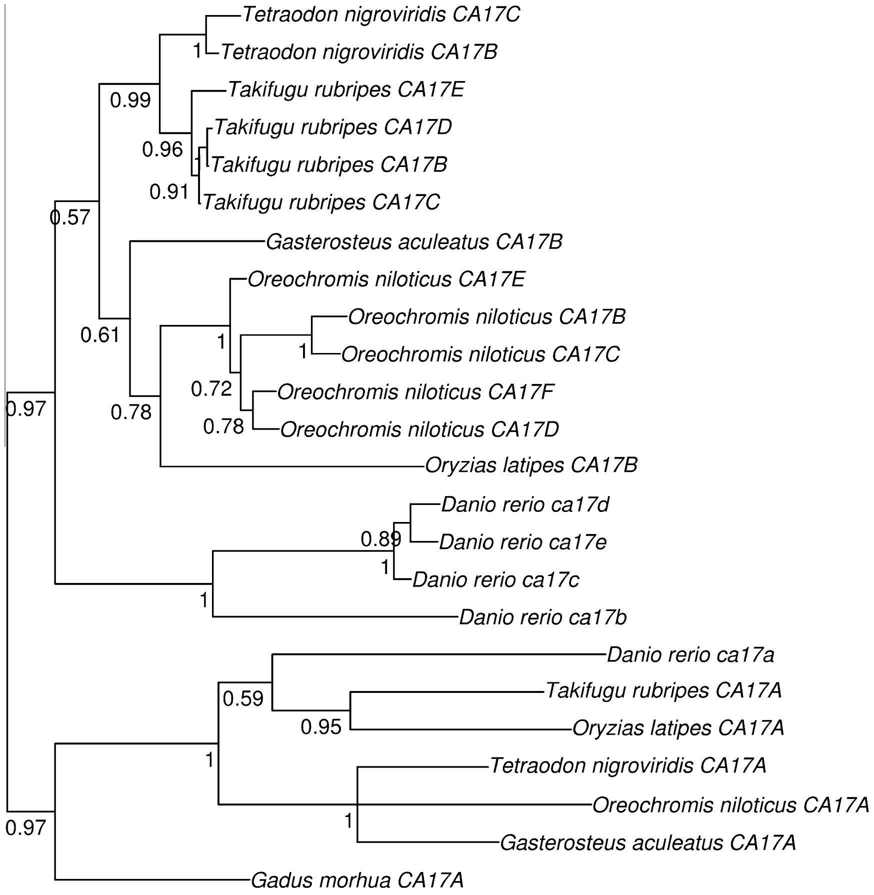

Figure 2. A phylogenetic tree of fish CA17 genes. This Bayesian phylogenetic tree was created using all 24 available sequences within the CA17 family in ray-finned fish species. See Section 2.3 for details. Branch lengths are proportional to distances, and the numbers at internal nodes are posterior probabilities.

Table 1
Gene names and expression data of zebrafish GPI-linked isozymes

Recommended name	Gene in zfin	Expression observed in: (zfin data)	Expression observed in: (UniGene EST data)
ca4a	ca4a	Adult: muscle, brain, eye, heart, intestine, spleen	
ca4b	ca4b		Brain, reproductive system
ca4c	ca4c		
ca15	ca16a		
ca17a	Not assigned		
ca17b	ca15b	At somite stage: primordial germ cell	
ca17c	zgc:153760		Bone, kidney, muscle, reproductive system, and skin
ca17d	ca15a	Day 5: integument and gill Adult: gill, muscle, spleen, and testis	Bone, brain, gills, olfactory rosettes
ca17e	ca15c		Bone, fin, gills

chromosomal locations are two pieces of evidence that clearly show that *CA17* genes code for a novel isozyme, a paralog of *CA4*.

Descendants of both *CA4* and *CA17* are found in fishes, in non-mammalian tetrapods and in the coelacanth, *Latimeria*, whereas therian mammals lack *CA17*. The most plausible hypothesis is that *CA17* was lost in mammals but retained in fish and non-mammalian tetrapods. During the evolution of ray-finned fishes, both *CA4* and *CA17* have undergone multiple further duplications. *CA4* is seen as tripled in fish genomes.

To visualize the duplications within the fish CA17 group we performed a separate phylogenetic analysis with all available sequences, including the lower-quality ones. In the resulting tree (Fig. 2) we observe an early duplication at the root and a series of recent duplications in the upper branch. We gave the name *CA17A* to the genes in the lower branch, which have remained single copies in all known fish genomes. In zfin there is no gene assigned to the corresponding region, chromosome 12: 5041,530-5051,292. The upper branch demonstrates parallel but independent multiplication events in several fish genera, resulting in up to six *CA17* genes per fish species. The genes in the upper branch are labeled starting from *CA17B* in each species. The order of the lettering is essentially random, based on the order seen in earlier trees we made. Because only one of the earlier *CA17* copies has been multiplied further, it would be interesting to study if there are specific genome elements which would have facilitated the duplications. It is also of significance to note that the multiple copies of *CA17* are seen clustered in single chromosomal regions, and the clusters include CA17A as well as a variable number of the genes shown in the upper branch in Figure 2. These recent duplication events suggest that the CA gene family is still evolving rapidly in fishes.

Supplementary data 2 shows an interactive diagram of the clusters and syntenic fragments in which CA17 genes are found. A summary of existing and suggested gene names is presented in Table 1.

All functional *CA15* orthologs are found as single copies in each genome, as shown in Figure 1 and in the Ensembl GeneTree. They should have the name *CA15* in each species (with the exception of historical rodent naming, *car15*). This includes the zebrafish gene labeled as *car15* in Ensembl (ENSDARG00000060829), LOC568143 in NCBI (protein NP_001038604.1), and *ca16a* in zfin.org and in a previous publication.[20] This 'ca16a' is undoubtedly a true one-to-one ortholog to all other known *CA15* genes, including the original mouse car15. The assignment as CA15 is supported by the Ensembl ortholog tables and also by the syntenic location in a chromosome region which contains the genes *DGCR2* and *DGCR14* in the proximity of *CA15*. Supplementary data 3 further demonstrates that the location of this zebrafish gene (in chromosome 10: 43,394,645-43,404,294) is syntenic to *CA15* in all mammals. Therefore, the gene name *ca16a* should be rejected and changed into *ca15*.

There are two other 'ca16' genes in zebrafish gene annotation, *ca16b* and *ca16c*, and these are clearly incorrect names as well. The gene products are in fact protein tyrosine phosphatase receptors (PTPRs) which have a CA-like domain as a minor part in their sequence. The gene *ca16b* is an ortholog to *PTPRG*, and *ca16c* is an ortholog to *PTPRZ*. The CA-related protein domains in PTPRs are not catalytically active CAs, and therefore these genes should not be called carbonic anhydrases, but instead protein tyrosine phosphatase receptors. In zfin.org *ptprgb* refers to D2U7Y2_DANRE in UniProt, a 270-residue fragment which is a perfect match with 248 consecutive residues, of 1382 amino acid residues, of the 'ca16b' gene product F1QWY5_DANRE. The protein D2U7Y2_DANRE is also mapped to the *ca16b* gene locus in Ensembl. Therefore, *ptprgb* would be appropriate instead of *ca16b*. For *ca16c* we would suggest the name *ptprz*, because no other protein tyrosine phosphatase receptor zeta is yet assigned in zfin.org.

However, because these two PTPR proteins contain CA-related protein domains, and we would like to disrupt previous nomenclature as little as possible, we suggest retention of CARP XVI as a descriptor for CA-related domains found in PTPRs. We have chosen the next available number CA17/CA XVII for the newly described group of isozymes.

In the case of CA4 isozymes in zebrafish, our trees retain the identities of *ca4a*, *ca4b*, and *ca4c* as defined in zfin.org, and these would be the names we suggest for their orthologs in other shown fish species too (except for *T. rubripes*, in which there are two copies, labeled CA4A1 and CA4A2 here).

3.1.3. New primate CA15 pseudogenes

Our previous work[35] revealed multiple pseudogene copies of *CA15* in human and chimpanzee genomes, and a discontinuous set of exons in the rhesus macaque genome. We have now revised our knowledge of primate *CA15* pseudogenes in light of new genome data.

In the genomes of gorilla and orangutan we can find sequences which might code for a protein highly similar to CA XV, but not full-length sequences. For orangutan, the protein is ENSPPYP00000009564 in Ensembl, and is missing a signal peptide. In the gorilla genome there is no annotation of a gene, but the coding region is in chromosome 22: 2437,836–2440,405 (in the gorGor3 2011 assembly). The exons we predict (data not shown) would code for a 263-residue fragment, missing the signal peptide and an estimated 40 residues in the C terminus. Both of these predicted proteins have regions which match poorly with mouse CA XV, but match well with the reconstructed human pseudogene sequences,[35] and most importantly, both have a substitution of H–N in the second zinc-binding histidine position of the active site, same as in human and chimpanzee *CA15* pseudogenes. Therefore, the orthologs of *CA15* in gorilla and orangutan cannot code for an enzymatically active CA. However, these sequences are considerably less disrupted than the pseudogenes in human or chimpanzee, in that there are no frameshifts or intervening stop codons in the coding sequences, and no Alu repeat sequence inserted within the exons. Therefore, we cannot exclude the possibility that these sequences would be still transcribed and translated, and perhaps have gained a new function in gorilla and/or orangutan.

The H–N mutation is also seen in the fragmentary *CA15* matches we find in gibbon (GL399200:253-2044).

In marmoset (*Callithrix jacchus*) we find an unannotated and fragmentary *CA15* sequence in chromosome 1: 183,443,677–183,445,044. It seems to contain the second exon (possibly with a signal sequence), third exon, more than half of the fourth exon, and start of the fifth exon. The shortened fourth exon still contains the first two His residues, notably without the H–N mutation in the second histidine that we have seen in all other primates. If we assume that the genome assembly is correct, this is another very broken *CA15* gene in primates.

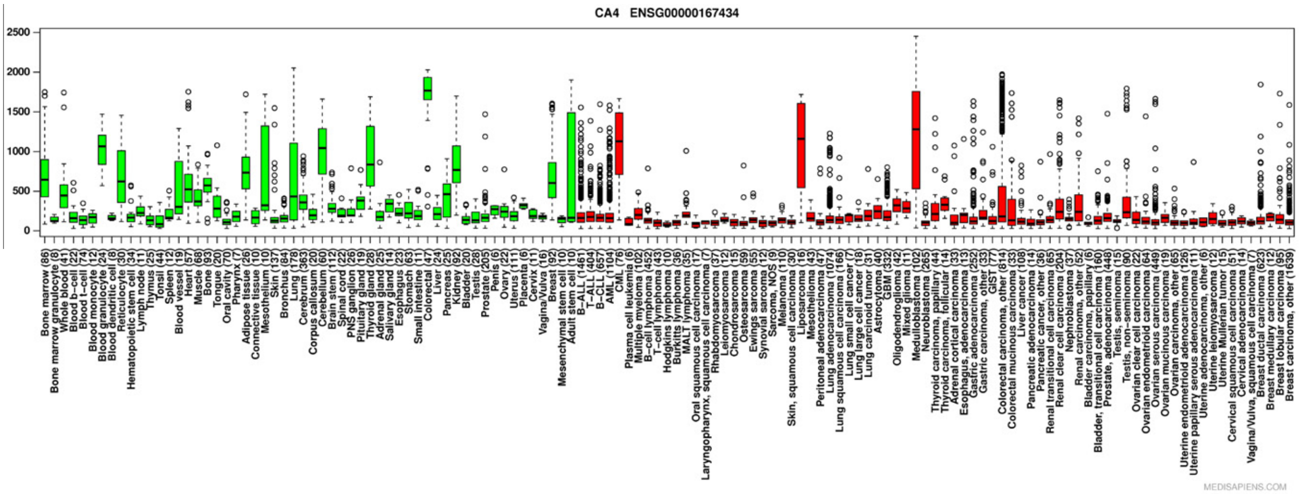

Figure 3. Human CA4 expression values from microarray data. CA4 expression values in 56 healthy tissues (green) and 75 cancer tissues (red), taken from the MediSapiens database, represent data from 20,000 microarrays.

Marmoset is a New World monkey, the only one with a sequenced genome available. The above results indicate that the second active-site His residue was lost early in Old World monkeys, but this was not necessarily the primary event that has inactivated *CA15*. One hypothesis would be that the insertion we see between the second and third histidines of the active site[35] is the original damage that killed *CA15* in primates, but unfortunately this region is not found in the marmoset genome, so the data remains inconclusive in this regard.

3.2. Expression data

In order to understand the functional evolution of GPI-linked CA isozymes, when different species have different numbers of these isozymes, we have compared their expression patterns in human, mouse, and zebrafish.

3.2.1. *CA4* is expressed in previously unreported human tissues

The MediSapiens project (http://medisapiens.com/) has collected data of expression levels of 19,000 human genes from 20,000 Affymetrix DNA microarrays, constituting the world's largest unified gene expression database of human tissues and diseases. They present aggregate data of all genes as expression levels in over 50 normal tissue or organ types and over 70 different cancer types, similar to their predecessor, the GeneSapiens collection, of 10,000 microarrays.[49] Figure 3 presents the expression levels of human *CA4* reported in MediSapiens.

The data verifies known expression of *CA4* in kidney, lungs, heart, colon, and colorectal carcinoma. Expression in reticulocytes is consistent with known expression in erythrocytes. In addition, several previously unreported tissues/organs with high expression of *CA4* emerge from MediSapiens data, including bone marrow, granulocytes, adipose tissue, thyroid gland, and breast. In cancer

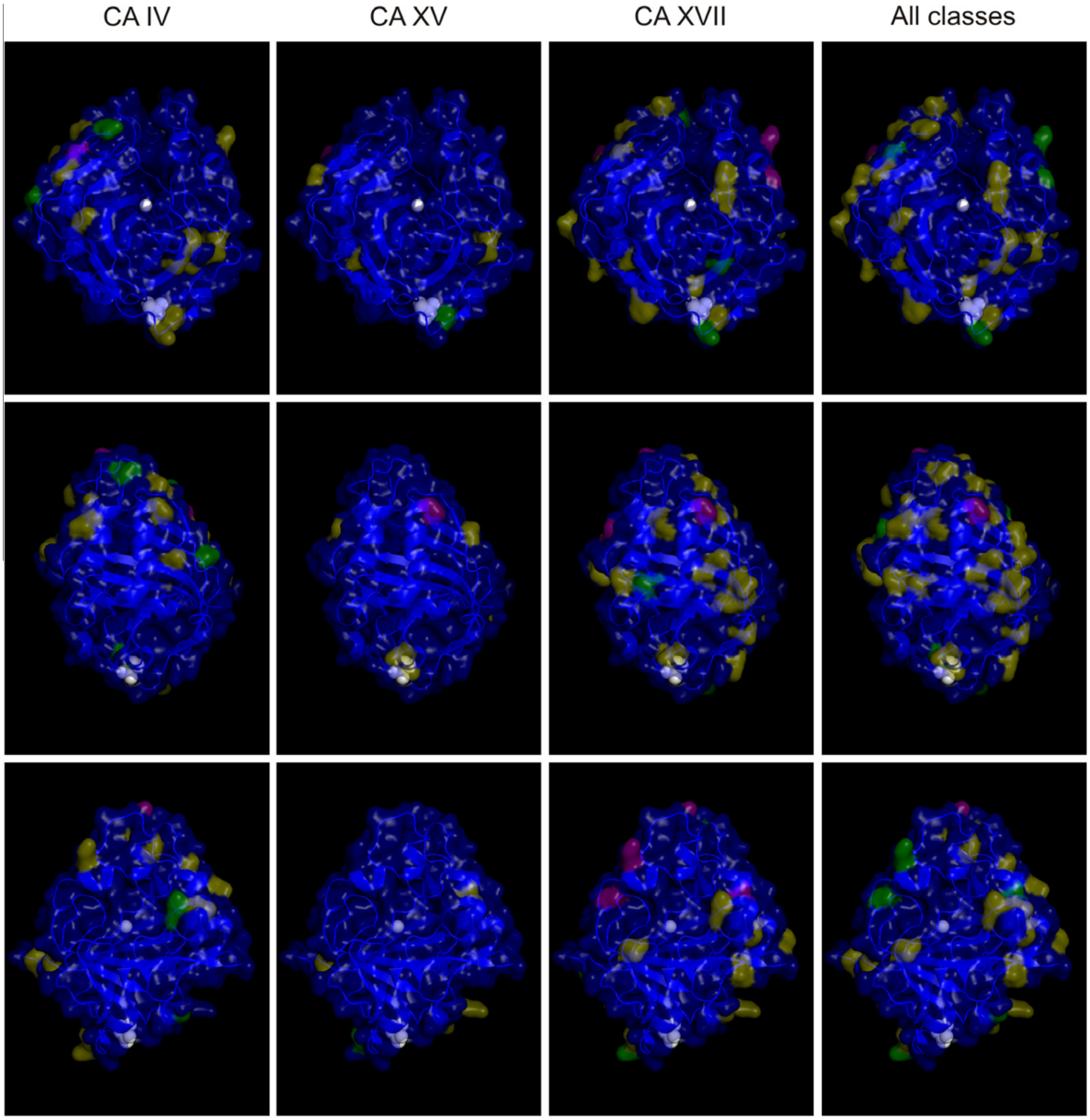

Figure 4. Distribution of N-glycosylation sites in vertebrate GPI-linked CAs. N-glycosylation sites were predicted in 75 protein sequences. Positions of sites were mapped onto chain B of PDB model 3FW3 for three CA classes, CA4, CA17, and CA15, and their composite. Within each CA class the total number of sequences with glycosylation predicted at each position was tabulated. For each CA class the range of scores was divided into thirds and corresponding residues in the 3FW3 model were colored accordingly: top third, magenta; middle third, green; and lower third, yellow. The Zn^{2+} in the active site and the C-terminal residue are indicated in white spheres. Three views are shown for each colored model. The C-terminus, representing the GPI anchor site, is oriented downwards in each view. Top row, a direct view into the active site; middle row, first view rotated 120° vertically to the right; and bottom row, first view rotated 120° vertically to the left. Visualizations were done with Pymol.

expression data we see several cancer types in which the expression was high in the corresponding normal tissue. They include thyroid cancer, liposarcoma, and breast cancer. Especially high expression of *CA4* is also seen in chronic myeloid leukemia and medulloblastoma, both of which are cancers which have not been associated with *CA4* expression previously.

3.2.2. Expression patterns of mouse *car4* and *car15* are similar to that of human *CA4*

We studied the expression of GPI-linked mouse CAs in the Gene Expression Database[50] found in the Mouse Genomics Informatics site (http://www.informatics.jax.org/expression.shtml, accessed Aug 2nd, 2012). The data relevant to CA IV and CA XV consist of histological staining with antisense RNAs, and there are only single-digit numbers of experiments available for each tissue or organ type. The data show more restricted expression of mouse *car4* (CA IV) than is observed for its human counterpart: only 12 of 96 tested histological classes are listed as positive, including kidney, brain (weak staining in few regions), heart, and tail. Most of the *car4* results come only from one study of 14.5-day mouse embryos, so we cannot be sure if the data reflect the expression pattern in the adult mouse. However, mouse *car15* (CA XV) is more widely expressed, and the combined expression patterns of *car4* and *car15* in the mouse overlap the expression pattern of human *CA4*. Positive *car15* antisense-RNA staining was seen in at least one study (out of one to three) in 81 of 95 tested histological classes. The *car15*-positive tissues include brain (strong staining in several regions), thyroid gland, eye, ear, kidney, reproductive system, lungs, heart, pancreas, liver, gastrointestinal tract (several regions), skeleton, and several others. Full details are available in the summary tables at http://www.informatics.jax.org/tissue/marker/MGI:1096574 (*car4*) and http://www.informatics.jax.org/tissue/marker/MGI:1931324 (*car15*).

Many of the tasks of CA XV, which is not encoded by the human genome, seem to have been taken over by human CA IV. This is an interesting example of adaptability and functional evolution within the gene family when one member becomes a pseudogene. However, the overlaps in observed expression patterns are not perfect.

3.2.3. Expression of GPI-linked zebrafish CAs resembles human CA IV expression

To have a preliminary idea of how functional roles are divided between the eight GPI-linked isozymes of zebrafish, we collected information regarding their expression patterns from UniGene and Zfin databases. The available data are summarized in Table 1. Many of the tissues/organs in which the GPI-linked isozymes are expressed, are similar to those in which human *CA4* is found (e.g., eye, kidney, testis, and gill). This resembles the above finding in the mouse, namely that tasks similar to those of human CA IV are distributed to a larger number of CA isozymes.

3.3. Protein-level bioinformatics

3.3.1. Sequence analysis agrees with GPI anchoring

We made signal peptide predictions and GPI anchor site predictions on our collection of complete, or nearly complete, protein sequences. Out of the 79 sequences, 62 were predicted positive for signal peptides. In the 17 negative cases, 14 were incomplete sequences which lack the initial methionine and a short piece of N-terminal sequence. In GPI anchor predictions, only 3 sequences of 79 failed in the prediction (with 99.5% specificity cutoff). Detailed results are available in Supplementary data 1.

3.3.2. Glycosylation sites occur in restricted areas on the protein surface

We wanted to study the distribution of N-glycosylation sites on the surface of GPI-linked isozymes. We identified all potential sites in 75 protein sequences, and mapped their positions back to the known CA IV structure with the help of a multiple sequence alignment. The 'popularity' of each position for glycosylation was tabulated, and color-coded by the number of sequences which share that position as a glycosylation site. We analyzed the predicted sites and saw all of them to have side chains either directly exposed to the surface, or plausibly exposed in the context of longer loops in their actual protein sequence. Out of the total 58 sites, four sites are within the region 125A–125H, presumably a loop, which is not visible in the structure. Out of the visible sites, 35 are found within loop regions, 10 in beta strands, and 9 in alpha helices. All beta strand located sites, except two, are in the ends of the strand elements.

Figure 4 shows the results divided to the three isozyme classes and also combined from all classes in one model. The conspicuous finding is that the glycosylation sites seem to be enriched on the side shown in the view of the middle row of images. In contrast, the region near the active site opening, seen in the top row of images, mainly shows absence of glycosylation. This is to be expected, because sugar chains near the opening might block access to the active site. Surprisingly, we see another area devoid of glycosylation sites in the bottom row of images, in the lower (membrane-proximal) region. This area might be kept clear of glycosylation to provide an essential protein-protein interaction surface. Our working hypothesis is that this region or part of the active-site-proximal region could contribute to the interaction between CA IV and ion transporters.

4. Conclusions

GPI-linked CA isozymes are more ancient than has been previously thought, as they are found even in nematodes and fungi. The pattern of duplications and gene losses in the family of vertebrate GPI-anchored CAs seems more complex than in any other group of CAs studied thus far, especially in fishes. Our analysis shows that an ancestor of *CA4* has given rise to two new isozymes by two duplications that took place before the jawed vertebrate radiation. The earlier duplication created *CA15*, and the second one gave rise to a new group which we have named CA17. Some previous classifications in zebrafish CA genes have been incorrect, and our work leads into renaming of six zebrafish genes and their orthologs in other fishes. The designation CA XVI is reserved to CARP XVI domains in protein tyrosine phosphatase receptors according to previous work. Analysis of glycosylation sites suggests potential protein-protein interaction regions shared between all GPI-linked vertebrate CA isozymes.

Acknowledgments

We thank the Academy of Finland, the Medical Research Fund of Tampere University Hospital, Biocenter Finland, and Sigrid Jusélius Foundation for financial support.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bmc.2012.08.060.

References and notes

1. Orlean, P.; Menon, A. K. *J. Lipid Res.* **2007**, *48*, 993.
2. Levental, I.; Grzybek, M.; Simons, K. *Biochemistry* **2010**, *49*, 6305.
3. Zhu, X. L.; Sly, W. S. *J. Biol. Chem.* **1990**, *265*, 8795.
4. Okuyama, T.; Sato, S.; Zhu, X. L.; Waheed, A.; Sly, W. S. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 1315.
5. Bottcher, K.; Waheed, A.; Sly, W. S. *Arch. Biochem. Biophys.* **1994**, *312*, 429.
6. Tufts, B. L.; Gervais, M. R.; Staebler, M.; Weaver, J. *J. Comput. Physiol. B* **2002**, *172*, 287.

7. Seron, T. J.; Hill, J.; Linser, P. J. *J. Exp. Biol.* **2004**, *207*, 4559.
8. Sterling, D.; Alvarez, B. V.; Casey, J. R. *J. Biol. Chem.* **2002**, *277*, 25239.
9. Wetzel, P.; Hasse, A.; Papadopoulos, S.; Voipio, J.; Kaila, K.; Gros, G. *J. Physiol.* **2001**, *531*, 743.
10. Svichar, N.; Waheed, A.; Sly, W. S.; Hennings, J. C.; Hubner, C. A.; Chesler, M. *J. Neurosci.* **2009**, *29*, 3252.
11. Yang, Z.; Alvarez, B. V.; Chakarova, C.; Jiang, L.; Karan, G.; Frederick, J. M.; Zhao, Y.; Sauve, Y.; Li, X.; Zrenner, E.; Wissinger, B.; Hollander, A. I.; Katz, B.; Baehr, W.; Cremers, F. P.; Casey, J. R.; Bhattacharya, S. S.; Zhang, K. *Hum. Mol. Genet.* **2005**, *14*, 255.
12. Klier, M.; Schuler, C.; Halestrap, A. P.; Sly, W. S.; Deitmer, J. W.; Becker, H. M. *J. Biol. Chem.* **2011**, *286*, 27781.
13. Svichar, N.; Chesler, M. *Glia* **2003**, *41*, 415.
14. Svichar, N.; Esquenazi, S.; Waheed, A.; Sly, W. S.; Chesler, M. *Glia* **2006**, *53*, 241.
15. Carter, N. D.; Fryer, A.; Grant, A. G.; Hume, R.; Strange, R. G.; Wistrand, P. J. *Biochim. Biophys. Acta* **1990**, *1026*, 113.
16. Fleming, R. E.; Parkkila, S.; Parkkila, A. K.; Rajaniemi, H.; Waheed, A.; Sly, W. S. *J. Clin. Invest.* **1995**, *96*, 2907.
17. Mahieu, I.; Benjamin, A.; Stephens, R.; Walters, D.; Carter, N. *Biochem. Soc. Trans.* **1995**, *23*, 320S.
18. Sender, S.; Decker, B.; Fenske, C. D.; Sly, W. S.; Carter, N. D.; Gros, G. *J. Histochem. Cytochem.* **1998**, *46*, 855.
19. Serrano, L.; Halanych, K. M.; Henry, R. P. *J. Exp. Biol.* **2007**, *210*, 2320.
20. Lin, T. Y.; Liao, B. K.; Horng, J. L.; Yan, J. J.; Hsiao, C. D.; Hwang, P. P. *Am. J. Physiol. Cell Physiol.* **2008**, *294*, C1250.
21. Esbaugh, A. J.; Gilmour, K. M.; Perry, S. F. *Respir. Physiol. Neurobiol.* **2009**, *166*, 107.
22. Parkkila, S.; Parkkila, A. K.; Kaunisto, K.; Waheed, A.; Sly, W. S.; Rajaniemi, H. *J. Histochem. Cytochem.* **1993**, *41*, 751.
23. Parkkila, S.; Parkkila, A. K.; Juvonen, T.; Waheed, A.; Sly, W. S.; Saarnio, J.; Kaunisto, K.; Kellokumpu, S.; Rajaniemi, H. *Hepatology* **1996**, *24*, 1104.
24. Hageman, G. S.; Zhu, X. L.; Waheed, A.; Sly, W. S. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 2716.
25. Chandrashekar, J.; Yarmolinsky, D.; von Buchholtz, L.; Oka, Y.; Sly, W.; Ryba, N. J.; Zuker, C. S. *Science* **2009**, *326*, 443.
26. Rebello, G.; Ramesar, R.; Vorster, A.; Roberts, L.; Ehrenreich, L.; Oppon, E.; Gama, D.; Bardien, S.; Greenberg, J.; Bonapace, G.; Waheed, A.; Shah, G. N.; Sly, W. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6617.
27. Alvarez, B. V.; Vithana, E. N.; Yang, Z.; Koh, A. H.; Yeung, K.; Yong, V.; Shandro, H. J.; Chen, Y.; Kolatkar, P.; Palasingam, P.; Zhang, K.; Aung, T.; Casey, J. R. *Invest. Ophthalmol. Vis. Sci.* **2007**, *48*, 3459.
28. Tian, Y.; Tang, L.; Cui, J.; Zhu, X. *Curr. Eye Res.* **2010**, *35*, 440.
29. Datta, R.; Waheed, A.; Bonapace, G.; Shah, G. N.; Sly, W. S. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 3437.
30. Breinin, G. M.; Gortz, H. *AMA Arch. Ophthalmol.* **1954**, *52*, 333.
31. Mincione, F.; Scozzafava, A.; Supuran, C. T. *Curr. Top. Med. Chem.* **2007**, *7*, 849.
32. Stams, T.; Nair, S. K.; Okuyama, T.; Waheed, A.; Sly, W. S.; Christianson, D. W. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13589.
33. Stams, T.; Chen, Y.; Boriack-Sjodin, P. A.; Hurt, J. D.; Liao, J.; May, J. A.; Dean, T.; Laipis, P.; Silverman, D. N.; Christianson, D. W. *Protein Sci.* **1998**, *7*, 556.
34. Vernier, W.; Chong, W.; Rewolinski, D.; Greasley, S.; Pauly, T.; Shaw, M.; Dinh, D.; Ferre, R. A.; Meador, J. W., 3rd; Nukui, S.; Ornelas, M.; Paz, R. L.; Reyner, E. *Bioorg. Med. Chem.* **2010**, *18*, 3307.
35. Hilvo, M.; Tolvanen, M.; Clark, A.; Shen, B.; Shah, G. N.; Waheed, A.; Halmi, P.; Hanninen, M.; Hamalainen, J. M.; Vihinen, M.; Sly, W. S.; Parkkila, S. *Biochem. J.* **2005**, *392*, 83.
36. Saari, S.; Hilvo, M.; Pan, P.; Gros, G.; Hanke, N.; Waheed, A.; Sly, W. S.; Parkkila, S. *PLoS ONE* **2010**, *5*, e9624.
37. Hilvo, M.; Salzano, A. M.; Innocenti, A.; Kulomaa, M. S.; Scozzafava, A.; Scaloni, A.; Parkkila, S.; Supuran, C. T. *J. Med. Chem.* **2009**, *52*, 646.
38. Ortutay, C.; Olatubosun, A.; Parkkila, S.; Vihinen, M.; Tolvanen, M. In *Advances in Medicine and Biology*; Bernhardt, L. E., Ed.; Nova Science Publishers: Hauppauge, NY, 2010; Vol. 7, pp 145–168.
39. Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fairley, S.; Fitzgerald, S.; Gil, L.; Gordon, L.; Hendrix, M.; Hourlier, T.; Johnson, N.; Kahari, A. K.; Keefe, D.; Keenan, S.; Kinsella, R.; Komorowska, M.; Koscielny, G.; Kulesha, E.; Larsson, P.; Longden, I.; McLaren, W.; Muffato, M.; Overduin, B.; Pignatelli, M.; Pritchard, B.; Riat, H. S.; Ritchie, G. R.; Ruffier, M.; Schuster, M.; Sobral, D.; Tang, Y. A.; Taylor, K.; Trevanion, S.; Vandrovcova, J.; White, S.; Wilson, M.; Wilder, S. P.; Aken, B. L.; Birney, E.; Cunningham, F.; Dunham, I.; Durbin, R.; Fernandez-Suarez, X. M.; Harrow, J.; Herrero, J.; Hubbard, T. J.; Parker, A.; Proctor, G.; Spudich, G.; Vogel, J.; Yates, A.; Zadissa, A.; Searle, S. M. *Nucleic Acids Res.* **2012**, *40*, D84.
40. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. *Bioinformatics* **2007**, *23*, 2947.
41. Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Fingerman, I. M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Krasnov, S.; Landsman, D.; Lipman, D. J.; Lu, Z.; Madden, T. L.; Madej, T.; Maglott, D. R.; Marchler-Bauer, A.; Miller, V.; Karsch-Mizrachi, I.; Ostell, J.; Panchenko, A.; Phan, L.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T. A.; Wagner, L.; Wang, Y.; Wilbur, W. J.; Yaschenko, E.; Ye, J. *Nucleic Acids Res.* **2012**, *40*, D13.
42. Petersen, T. N.; Brunak, S.; von Heijne, G.; Nielsen, H. *Nat. Methods* **2011**, *8*, 785.
43. Pierleoni, A.; Martelli, P. L.; Casadio, R. *BMC Bioinformatics* **2008**, *9*, 392.
44. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.
45. Suyama, M.; Torrents, D.; Bork, P. *Nucleic Acids Res.* **2006**, *34*, W609.
46. Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D. L.; Darling, A.; Hohna, S.; Larget, B.; Liu, L.; Suchard, M. A.; Huelsenbeck, J. P. *Syst. Biol.* **2012**, *61*, 539.
47. Paradis, E.; Claude, J.; Strimmer, K. *Bioinformatics* **2004**, *20*, 289.
48. Muffato, M.; Louis, A.; Poisnel, C. E.; Roest Crollius, H. *Bioinformatics* **2010**, *26*, 1119.
49. Kilpinen, S.; Autio, R.; Ojala, K.; Iljin, K.; Bucher, E.; Sara, H.; Pisto, T.; Saarela, M.; Skotheim, R. I.; Bjorkman, M.; Mpindi, J. P.; Haapa-Paananen, S.; Vainio, P.; Edgren, H.; Wolf, M.; Astola, J.; Nees, M.; Hautaniemi, S.; Kallioniemi, O. *Genome Biol.* **2008**, *9*, R139.
50. Finger, J. H.; Smith, C. M.; Hayamizu, T. F.; McCright, I. J.; Eppig, J. T.; Kadin, J. A.; Richardson, J. E.; Ringwald, M. *Nucleic Acids Res.* **2011**, *39*, D835.