OPEN ACCESS Freely available online

Journal of Data Mining in Genomics & Proteomics

Research Article

Comparative Studies of Vertebrate Mitochondrial Carbonic Anhydrase (CA5) Genes and Proteins: Evidence for Gene Duplication in Mammals with CA5A Being Liver Specific and CA5B Broadly Expressed and Located on the X-Chromosome

Roger S Holmes*

Griffith Research Institute for Drug Discovery (GRIDD) and School of Environment and Science, Griffith University, Nathan, QLD 4111, Australia

ABSTRACT

At least fifteen families of mammalian carbonic anhydrases (CA) (E.C. 4.2.1.2) catalyse the hydration of carbon dioxide and related functions. CA5A and CA5B genes encode distinct mitochondrial enzymes and perform essential biochemical roles, including ammonia detoxification and glucose metabolism. Bioinformatic methods were used to predict the amino acid sequences, secondary structures and gene locations for CA5A and CA5B genes and proteins using data from vertebrate genome projects. CA5A and CA5B genes usually contained 7 coding exons for each of the vertebrate genomes examined. Human CA5A and CA5B subunits contained 305 and 317 amino acids, respectively, with key amino acid residues including mitochondrial transit peptides; three Zinc binding sites (His130, His132, His155); and a Tyr164 active site. Phylogenetic analyses of vertebrate CA5 gene families suggested that it is an ancient gene in vertebrate evolution which had undergone a gene duplication event in a mammalian ancestral genome forming the CA5A and CA5B gene families in monotreme, marsupial and eutherian mammals. CA5A was predominantly expressed in liver whereas CA5B had a wide tissue distribution profile, was localized on the X-chromosome and was more highly conserved during mammalian evolution.

KEYWORDS: Mitochondrial enzymes; Vertebrate genome; Vertebrate; Carbon dioxide

INTRODUCTION

At least fifteen families of mammalian carbonic anhydrases (CA) (E.C. 4.2.1.2) catalyse the hydration of carbon dioxide and related functions, and are involved in a range of biological functions, including respiration, bodily fluid formation, calcification, regulating acid base balance and bone reabsorption [1-5]. CA genes are differentially expressed in the body and have diverse tissue and subcellular distribution profiles [2,4,5]. These include the CA5A and CA5B genes, which encode distinct mitochondrial enzymes and perform essential biochemical roles, including ammonia detoxification and glucose metabolism, and have similar 3D structures with the other CA isozymes [6-20]. Targeted mutagenesis studies have shown that both enzymes play important metabolic roles although Ca5a 'null' mice showed more significant deleterious effects than CA5B 'null' mice, with Ca5A/Ca5b double knockouts showing more substantial effects [7]. Genetic analyses of human CA5A and CA5B have shown that these enzymes are encoded by distinct genes, with CA5A and CA5B localized on separate chromosomes, chromosome 16 and the X-chromosome, respectively [8,20]. Moreover, genetic variants of CA5A in human populations have caused hyperammonia in early childhood [20] and mutagenesis studies of zebrafish CA5 suggested that this enzyme regulates acid-base homeostasis in this organism [14].

This paper reports the predicted amino acid sequences, gene locations and exon structures for CA5-like vertebrate genes and proteins, including primates (human (*Homo sapiens*) and baboon (*Papio anubis*), other eutherian mammals (mouse (*Mus musculus*), rat (*Rattus norvegicus*) and cow (*Bos Taurus*)), a marsupial mammal (opossum) (*Monodelphis domestica*), a monotreme mammal (platypus) (*Ornithorhynchus anatinus*)), and representatives of birds (chicken (*Gallus gallus*)), reptiles [alligator (*Alligator mississippiensis*)), frogs (*Xenopus tropicalis* and *Xenopus laevis*), fish [zebra fish (*Danio rerio*), medaka (*Oryzias latipes*) and coelacanth (*Latimeria chalumnae*)) and elephant shark (*Callorhinchus millii*). The phylogenetic and

Correspondence to: Roger S Holmes, Griffith Research Institute for Drug Discovery (GRIDD) and School of Environment and Science, Griffith University, Nathan, QLD 4111, Australia, E-mail: r.holmes@griffith.edu.au

Received: June 04, 2020; **Accepted:** June 19, 2020; **Published:** June 25, 2020

Citation: Holmes RS (2020) Comparative Studies of Vertebrate Mitochondrial Carbonic Anhydrase (CA5) Genes and Proteins: Evidence for Gene Duplication in Mammals with CA5A Being Liver Specific and CA5B Broadly Expressed and Located on the X-Chromosome. J Data Mining Genomics Proteomics 10:223. doi: 10.35248/2165-7556.20.11.223.

evolutionary relationships of these genes and enzymes are described with a hypothesis for a gene duplication event for an ancestral mammalian CA5 gene, generating CA5A and CA5B genes, which are separately localized on mammalian genomes, including the X-chromosome for CA5B, and are differentially expressed in tissues of the body.

METHODS

Vertebrate CA5 gene and protein identification

BLAST (Basic Local Alignment Search Tool) studies were undertaken using web tools from the National Center for Biotechnology Information (NCBI) [21]. Protein BLAST analyses used the human and mouse CA5A and CA5B amino acid sequences deduced from reported sequences for these genes [7,8,15]. Non-redundant protein sequence databases for several mammalian and other vertebrate genomes were obtained using the blastp algorithm for the following genome sequences: human (Homo sapiens) [22]; baboon (*Papio anubis*) [23], cow (*Bos Taurus*) [24]; mouse (*Mus musculus*); [25] rat (*Rattus norvegicus*); [26] opossum (*Monodelphis domestica*); [27] platypus (*Ornithyrinchus anatinus*); [28] chicken (*Gallus gallus*); [29] alligator (*Alligator mississippiensis*); frog (*Xenopus tropicalis*); [30] and zebrafish (*Danio rerio*) [31].

BLAT analyses were subsequently undertaken for each of the predicted CA5-like amino acid sequences using the UC Santa Cruz web browser [32] to obtain the predicted locations for each of the vertebrate CA5-like genes, including exon boundary locations and gene sizes . Structures for human CA5A and CA5B isoforms were obtained using the AceView website to examine predicted gene and protein structures to interrogate this database of human mRNA sequence [33] (Table 1).

Predicted structures and properties of vertebrate CA5 subunits

Alignments of predicted CA5-like amino acid sequences and estimates of sequence identities were undertaken using a ClustalW method [34]. Predicted secondary structures for vertebrate CA5 subunits were obtained using alignments with the reported tertiary structures for mouse CA5A9 and CA5B [10]. Predictions of the CA5A, CA5B and CA5 protein N-terminal sequences serving as mitochondrial targeting peptides, and the cleavage site for this peptide, were undertaken using MITOPROT [35].

Human CA5A and CA5B gene expression and predicted gene regulation sites

The human genome browser was used to examine predicted CpG islands [36] and transcription factor binding sites (TFBS) (ORegAnno IDs: Open Regulatory Annotations) [37] for human CA5A and CA5B using the UC Santa Cruz Genome Browser [32]. The GTEx web browser was used to examine the human tissue expression profiles for CA5A, CA5Aps and CA5B [38].

Phylogenetic studies and sequence divergence

Mammalian CA5A and CA5B and other vertebrate CA5 sequences were subjected to phylogenetic analysis using the http://www.phylogeny.fr/ portal to enable alignment (MUSCLE), curation (Gblocks), phylogeny (PhyML) and tree rendering (TreeDyn) to reconstruct phylogenetic relationships [39]. Mammalian sequences were identified as members of the CA5A or CA5B (mitochondrial) groups, whereas non-mammalian vertebrate sequences were identified as members of the CA5 group.

Table 1: Mammalian CA5A and CA5B and other vertebrate CA5 genes and proteins

*predicted sequence; ^scaffold IDs are shown; transcript IDs were derived from NCBI sources http://www.ncbi.nlm.nih.gov/genbank/; UNIPROT refers to UniprotKB/Swiss-Prot IDs for individual CA5A, CA5B and other vertebrate CA5 subunits (see http://kr.expasy.org); the number of coding exons are listed; 'na' means data not available; single CA5 sequences were observed from lower vertebrate sources (birds; reptiles; amphibians; fish; and sharks).

Vertebrate	Species	CA Gene	Gene Location	Transcript ID*	Exons (Strand)	UNIPROT ID	Amino acids
Human	*Homo sapiens*	CA5A	16:87,888,132-87,936,450	L19297	7 (-)	P35216	305
		CA5B	X:15,750,024-125,782,661	BC028142	7 (+)	Q9Y2D0	317
Baboon	*Papio anubis*	CA5A	20:69,860,214-69,909,041	*XP_003917341	7 (-)	A0A5F7ZRP5	307
		CA5B	X:13,133,143-13,166,842	*XP_003917488	7 (+)	A0A096N1V1	317
Mouse	*Mus musculus*	CA5A	8:121,916,367-121,944,793	BC030174	7 (-)	P23589	299
		CA5B	X:163,979,194-164,014,944	BC034413	7 (-)	Q9QZA0	317
Rat	*Rattus norvegicus*	CA5A	19:54,732,112-54,761,585	BC088147	7 (-)	P43165	304
		CA5B	X:32,244,503-32,290,304	BC081872	7 (+)	Q66HG6	317
Cow	*Bos taurus*	CA5A	18:13,345,984-13,368,318	*NP_001179338	7 (-)	na	310
		CA5B	X:127,727,046-127,743,377	*NP_001074377	7 (-)	na	317
Opossum	*Monodelphis domestica*	CA5A	1:693,776,094-693,836,235	*XP_007477337	7 (-)	F6W8Y6	298
		CA5B	7:24,269,877-24,307,423	*XP_007500932	7 (-)	F6XEV7	313
Platypus	*Ornithorhynchus anatinus*	CA5A	^DS180954v1:2,170,173-2,206,662	*XP_001508964.3	7 (-)	na	307
		CA5B	^DS181337v1:8,449,952-8,469,447	*XP_028935902.1	7 (+)	na	315
Chicken	*Gallus gallus*	CA5	11:17,993,314-18,005,353	*XP_414195	7 (-)	F1N986	314
Alligator	*Alligator mississippiensis*	CA5	^JH738261:154,749-178,224	*XP_019346721	7 (-)	A0A151ND60	306
Tropical frog	*Xenopus tropicalis*	CA5	4:68,200,520-68,227,829	*XP_012816720	7 (-)	Q28BX3	309
Clawed toad	*Xenopus laevis*	CA5	4L:54,653,647-54,676,100	*XP_018112094.1	7 (-)	Q6NTY3	311
Zebra fish	*Danio rerio*	CA5	25:12,798,498-12,808,955	*NP_001104671	7 (-)	na	310
Medaka	*Oryzias latipes*	CA5	6:19,186,720-19,195,404	*XP_004069790	7 (+)	na	314
Coelacanth	*Latimeria chalumnae*	CA5	^JH126597:1,194,229-1,249,345	*XP_014340058	7 (-)	H3B5R4	310
Shark	*Callorhinchus milii*	CA5	^KI635866:5,082,728-5,091,173	*XP_007887550	7 (+)	A0A4W3J095	290

RESULTS AND DISCUSSION

Alignments and biochemical features of vertebrate CA5 amino acid sequences

Amino acid sequence alignments for opossum (marsupial) CA5A and CA5B, chicken and zebrafish CA5 amino acid sequences are shown in Figure 1, together with the previously reported sequences for human and mouse CA5A and CA5B [7,8,15]. The vertebrate CA5-like sequences exhibited >50% identities, suggesting that these protein subunits are products of the same gene family (Table 2).

Amino acid sequences for the mammalian CA5-like proteins examined contained 299-310 (CA5A) and 313-317 (CA6B) residues, whereas the lower vertebrate CA5 sequences examined contained 290-314 residues. The elephant shark CA5 was the smallest among the vertebrate CA5-like proteins examined with 290 amino acid residues (Table 1). The N-termini showed the lowest levels of sequence identity among the sequences examined perhaps due to the presence of the transit peptide for facilitating mitochondrial localization (Figure 1).

X-ray crystallographic studies for mouse CA5A and CA5B have enabled the identification of key structural and catalytic residues among those aligned for vertebrate CA-like sequences examined (Figure 1) [7,15]. These included mouse Tyr94, Tyr158 and Tyr161 which were identified as catalytic residues; His124, His126 and His149, which were responsible for chelating the Zinc residue at the active site; and 229Thr-230Thr, involved in substrate binding. Tyr 94 is conserved for all of the vertebrate CA5 sequences examined, whereas Tyr158 underwent a substitution with phenylalanine for human and mouse CA5B; and Tyr161 underwent a similar phenylalanine substitution in the mammalian CA5B and lower vertebrate CA5 sequences examined. Genetic substitution of mouse CA5A Ser233 with Pro 233 resulted in markedly reduced activity, [20] which suggests a significant role in catalysis for this conserved amino acid.

Predicted gene locations, exon structures and tissue expression for vertebrate CA5 genes

Table 1 and Figure 1 summarize the predicted locations and exon structures for vertebrate CA5-like genes based upon BLAT interrogations of several vertebrate genomes using the sequences for human [6,8] and mouse [7] CA5A and CA5B, and the predicted sequences for other vertebrate CA5 subunits (Table 1) and the UC Santa Cruz Web Browser [32]. Vertebrate CA5 genes contained 7 coding exons with the predicted exon start sites in identical or similar positions. Figure 2 describes the tissue expression profiles for CA5A and CA5B, as well as a CA5A pseudogene (CA5Aps) [38] (Figure 2).

Human CA5A was predominantly expressed in liver, as compared with CA5Aps which was detected at low levels only in human testis, with both genes located on chromosome 16: CA5A covering 48.5kb from 87,970,135-87,921,623 on the reverse strand, while CA5Aps covered 17.5kb from 29,618,785-29,636,328, also on the reverse

Table 2: Percentage identities for mammalian CA5A and CA5B and other vertebrate CA5 amino acid sequences.

	Human CA5A	Mouse CA5A	Opossum CA5A	Human CA5B	Mouse CA5B	Opossum CA5B	Chicken CA5	Zebra fish CA5
Human CA5A	100	72	67	59	59	61	64	50
Mouse CA5A	72	100	63	57	57	57	63	50
Opossum CA5A	67	63	100	63	62	67	66	54
Human CA5B	59	59	63	100	89	80	69	54
Mouse CA5B	59	57	62	89	100	79	69	53
Opossum CA5B	61	57	67	80	79	100	74	57
Chicken CA5	64	57	66	69	69	74	100	58
Zebra fish CA5	50	50	54	54	53	57	58	100

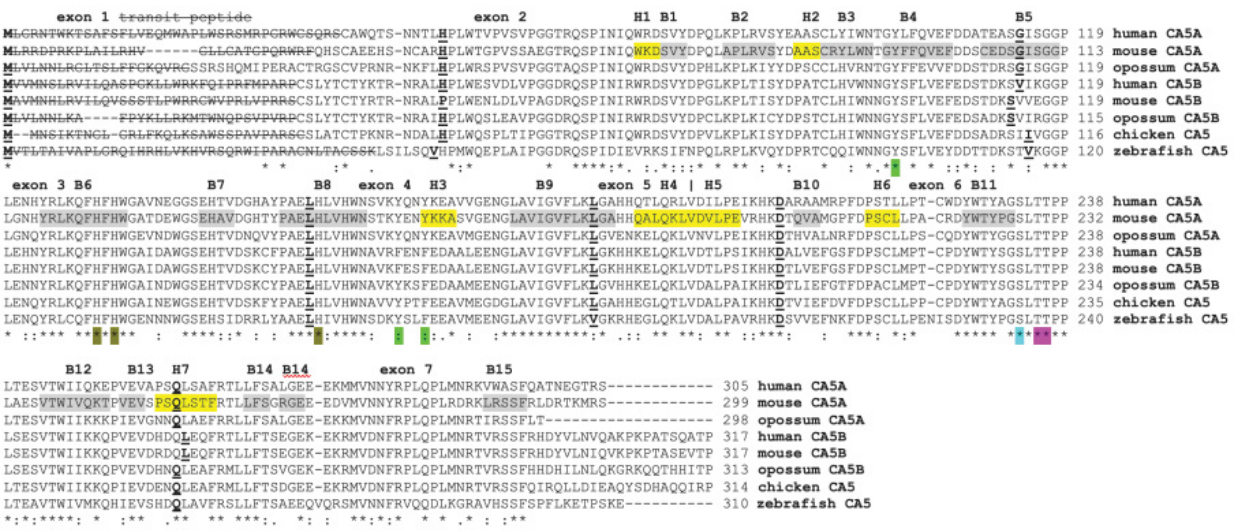

Figure 1: Amino acid sequence alignments for vertebrate CA5A, CA5B and CA5 sequences
See Table 1 for sources of CA5A, CA5B and CA5 sequences; * identical residues; : 1 or 2 conservative substitutions; . 1 or 2 non-conservative substitutions; active site His residues for binding Zn2+; catalytic active site residues; helices H1, H2 etc; sheets B1 B2 etc; conserved serine and threonine residues are involved in substrate binding; bold underlined font shows predicted exons (numbered) junctions; and mitochondrial transit peptides are shown.

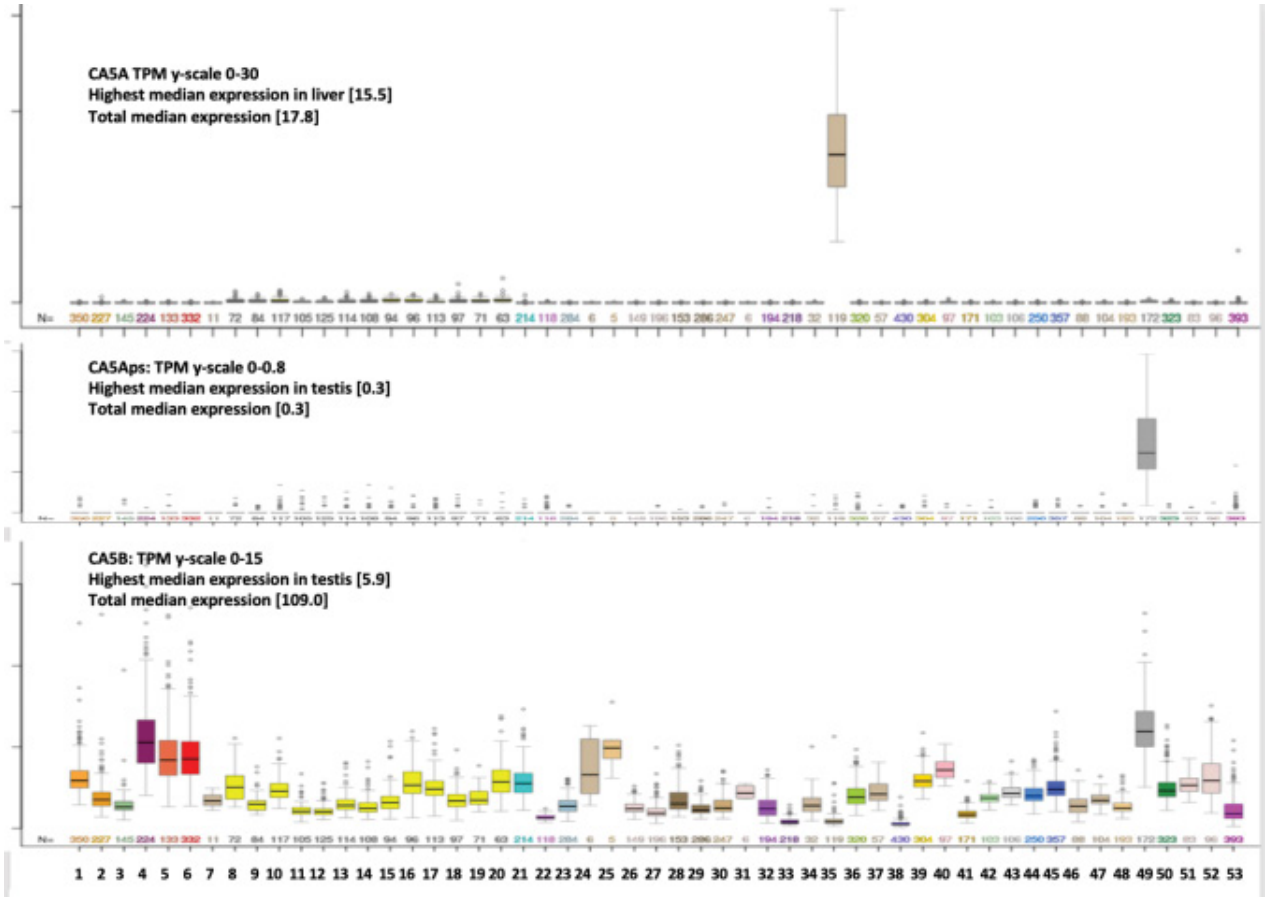

Figure 2: Comparative tissue expression levels for human *CA5A*, *CA5Aps* and *CA5B*
RNA-seq gene expression profiles across 53 selected tissues (or tissue segments) were examined from the public database for human *CA5*, *CA5Aps* and *CA5B* based on expression levels for 175 individuals[38] (http://www.gtex.org). Tissues: 1. Adipose-Subcutaneous; 2. Adipose-Visceral (Omentum); 3. Adrenal gland; 4. Artery-Aorta; 5. Artery-Coronary; 6. Artery- Tibial; 7. Bladder; 8. Brain-Amygdala; 9. Brain-Anterior cingulate Cortex (BA24); 10. Brain-Caudate (basal ganglia); 11. Brain-Cerebellar Hemisphere; 12. Brain- Cerebellum; 13. Brain-Cortex; 14. Brain-Frontal Cortex; 15. Brain-Hippocampus; 16. Brain-Hypothalamus; 17. Brain-Nucleus accumbens (basal ganglia); 18. Brain- Putamen (basal ganglia); 19. Brain-Spinal Cord (cervical c-1); 20. Substantia nigra; 21. Breast-Mammary Tissue; 22. Cells-EBV-transformed lymphocytes; 23. Cells- Transformed fibroblasts; 24. Cervix-Ectocervix; 25. Cervix-Endocervix; 26. Colon-Sigmoid; 27. Colon-Transverse; 28. Esophagus-GastroesophagealBrain-Junction; 29. Esophagus-Mucosa; 30. Esophagus-Muscularis; 31. Fallopian Tube; 32. Heart-Atrial Appendage; 33. Heart-Left Ventricle; 34. Kidney-Cortex; 35. Liver; 36. Lung; 37. Minor Salivary Gland; 38. Muscle-Skeletal; 39. Nerve-Tibial; 40. Ovary; 41. Pancreas; 42. Pituitary; 43. Prostate; 44. Skin-Not Sun Exposed (Suprapubic); 45. Skin-Sun Exposed (Lower leg); 46. Small Intestine-Terminal Ileum; 47. Spleen; 48. Stomach; 49. Testis; 50. Thyroid; 51. Uterus; 52. Vagina; 53. Whole Blood. TPM, Transcripts per million, calculated from a gene model with isoforms collapsed to a single gene. Box plots show a median and 25th and 75th percentiles; points are shown as the outliers if they are above or below 1.5 times the interquartile range.

strand of chromosome 16. The human CA5B gene comprised two consecutive components (CA5BP1 and CA5B) located on the plus strand of the X-chromosome (15,693,048-15,806,528) (Figure 3) and was expressed with a broad tissue distribution profile, with highest levels of expression observed in testis and arteries (Figure 2). Comparative levels of total median expression for these genes showed nearly 6 times higher levels for human CA5B as compared with CA5A, but with human liver showing > 2 times the liver specific expression of the CA5A gene as compared with tissue specific expressions of the CA5B gene.

Figure 3 presents the predicted structures of human CA5A and CA5B gene transcripts [33] (Figure 3).

There were 7 coding exons for the CA5A isoform 2 precursor mRNA (RefSeq:NM_001367225.1) sequence which contained several transcription factor binding sites in the 5' region: an early growth response gene (EGR1), involved in regulating synaptic plasticity [40]; hepatocyte nuclear factor 4A (HNF4A), a nuclear transcription factor which controls the expression of several hepatic genes [41]; and CCAAT enhancer binding protein alpha (CEBPA), a transcription factor that coordinates proliferation arrest and the differentiation of hepatocytes [42]. There were also 7 coding exons for the human CA5B gene on the X-chromosome, which included a gene enhancer regulatory element (GERE) at the 5'end of the gene [43], located near CpG47 and several transcription factor binding sites: STAT3 (a tyrosine phosphorylated transcription factor) [44]; EGR140 (shared with CA5A); FOXH1 (a fork head DNA binding transcription factor which is essential for development of the anterior heart field) [45]; and ZEB1 (Zinc finger E-box-binding homeobox 1, which is required for neural differentiation of human embryonic stem cells [46].

Phylogeny of primate CA5A and CA5B and vertebrate CA5 sequences

A phylogenetic tree (Figure 4) was constructed from alignments of mammalian CA5A and CA5B amino acid sequences with other lower vertebrate CA5 sequences, with representatives of bird (chicken), reptile (alligator), amphibian (frogs), fish and a shark species. The phylogram showed clustering into 2 major groups of the mammalian CA5 sequences consistent with previous reports for mammalian CA5A and CA5B genes and enzymes [7-11], [15-20]. In contrast, evidence was obtained for single copies of

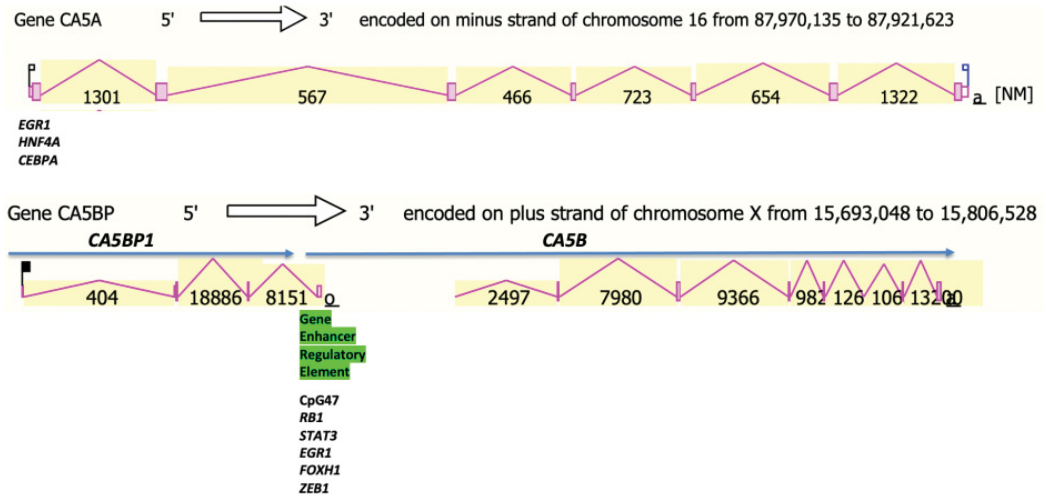

Figure 3: Gene structures for human *CA5A* and *CA5B*
From AceView website[33] http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/ The major isoforms are shown for the *CA5A* and *CA5B* transcripts; note that the *CA5B* transcript was split into 2 components with *CA5B* encoding the CA5B enzyme and *CA5BP1* encoding a pseudogene; capped 5'- and 3'-ends for the predicted mRNA sequences are identified; a predicted CpG island (CpG47), a gene enhancer regulatory element (GERE)[43]; and transcription factor binding sites (*EGR1*,[40] *HNF4A*,[41] *CEBPA*,[42] *STAT3*,[44] *FOXH1*,[45] *ZEB1*[46]) are shown. The numbers of nucleotides separating exons are also shown.

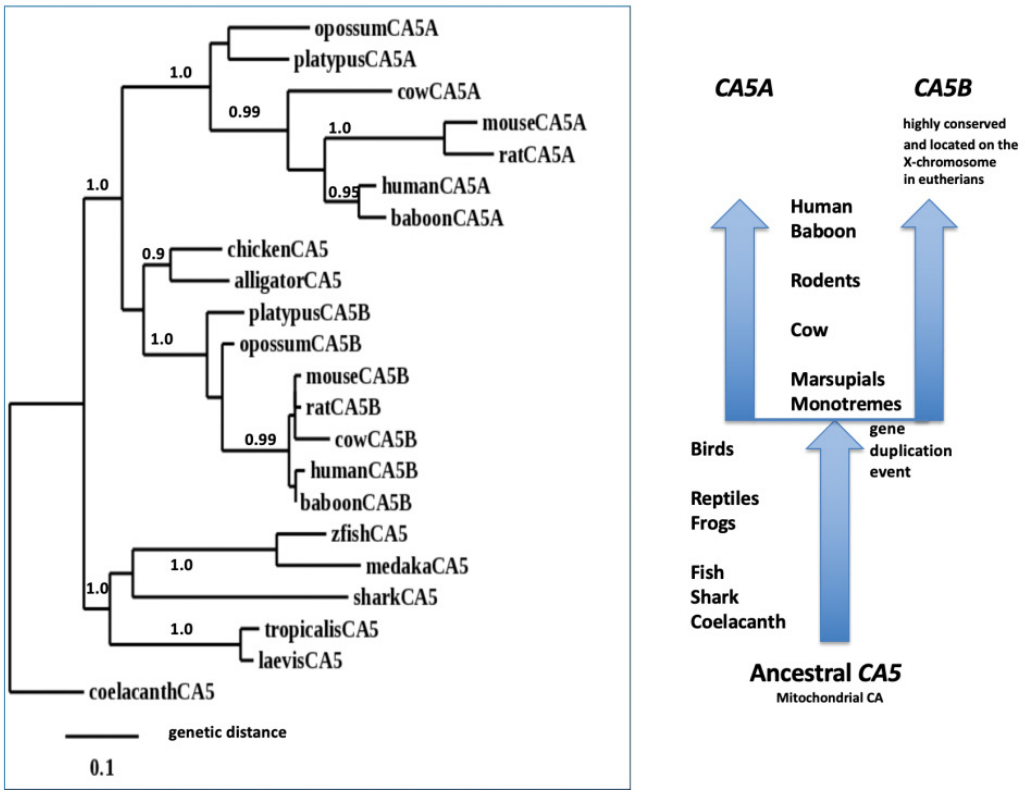

Figure 4: Phylogenetic tree of mammalian CA5A and CA5B sequences and other vertebrate CA5 sequences
The tree is labeled with the *CA5* gene name and the name of the vertebrate; note the major clusters include the lower vertebrate CA5 group and two groups for the mammalian CA5A and CA56B enzymes; a gene duplication event generating the mammalian *CA5A* and *CA5B* gene families is proposed to have occurred in a mammalian *CA5* ancestral gene leading to the formation of the monotreme, marsupial and eutherian mammal groups. A genetic distance scale is shown. The number of times a clade (sequences common to a node or branch) occurred in the bootstrap replicates are shown. Only replicate values of 0.9 or more which are highly significant are shown. 100 bootstrap replicates were performed in each case. Note the higher level of sequence conservation observed for the eutherian mammalian CA5B sequence. Sequences were derived from those reported in Table 1.

CA5 genes and enzymes among lower vertebrates, including the Xenopus laevis (clawed toad) species, for which multiple copies of genes have been reported due to the genome undergoing tetraploidization [47]. The phylogram suggested an ancestral relationship between lower vertebrate CA5 and mammalian CA5, with the latter gene undergoing a gene duplication event resulting in the appearance of CA5A and CA5B genes, in the ancestor leading to the appearance of monotremes, and subsequently marsupial and eutherian mammals. Monotremes have been described as arising from primitive birds which diverged from marsupials and eutherians about 163 to 186 Ma (million years ago) [48]. Platypus and opossum genomic sequences have been reported [27,28] and incorporated into the genome browser, enabling identification of CA5A and CA5B-like genes and enzymes sequences in these

Table 3: Primate CA5A genes and proteins

*predicted sequence; ^scaffold IDs are shown; transcript IDs were derived from NCBI sources http://www.ncbi.nlm.nih.gov/genbank/; UNIPROT refers to UniprotKB/Swiss-Prot IDs for individual CA5A subunits (see http://kr.expasy.org); the number of coding exons are listed; 'na' means data not available.

Primate	Species	Gene Location	Transcript ID*	Exons (Strand)	UNIPROT ID	Amino acids
Human	*Homo sapiens*	16:87,888,132-87,936,450	L19297	7 (-)	P35216	305
Chimp	*Pan troglodytes*	16:73,578,402-73,625,475	*XP_523486.2	7 (-)	H2QBP6	305
Gorilla	*Gorilla gorilla*	^CYUI01015509v1:365,282-412,997	*XP_030858903.1	7 (-)	na	304
Orang-utan	*Pongo abelii*	16:66,031,718-66,080,856	*XP_024089455.1	7 (-)	H2NRR0	304
Gibbon	*Nomascus leucogenys*	2:161,000,175-161,047,105	*XP_030675794.1	7 (-)	na	308
Baboon	*Papio anubis*	20:69,860,214-69,909,041	*XP_003917341.1	7 (-)	A0A096N1V1	307
Green monkey	*Chlorocebus sabaeus*	5:73,284,646-73,332,986	*XP_007992511.1	7 (-)	na	307
Gelada monkey	*Theropithecus gelada*	na	*XP_025226548.1	na	na	307
Rhesus macaque	*Macaca mulatta*	20:74,907,317-74,958,053	*XP_014982249.2	7 (-)	A0A5F7ZRP5	307
Pig-tailed macaque	*Macaca nemestrina*	20:76,285,953-76,343,242	*XP_011751584.1	7 (-)	na	307
Crab-eating macaque	*Macaca fascicularis*	20:76,285,953-76,343,288	*XP_005592801.1	7 (-)	A0A2K5VV27	307
Golden snub-nosed monkey	*Rhinopithecus roxellana*	^KN299711v1: 1,242,256-1,292,347	*XP_010353882.2	7 (+)	A0A2K6RLY9	307
Squirrel monkey	*Saimiri boliviensis*	^JH378111:45,180,556-45,231,845	*XP_003922881.1	7 (-)	na	307
Capuchin monkey	*Cebus capucinus*	na	*XP_017399192.1	na	na	308
Marmoset	*Callithrix jacchus*	20:42,330,747-42,381,907	*XP_002761295.2	7 (-)	na	305

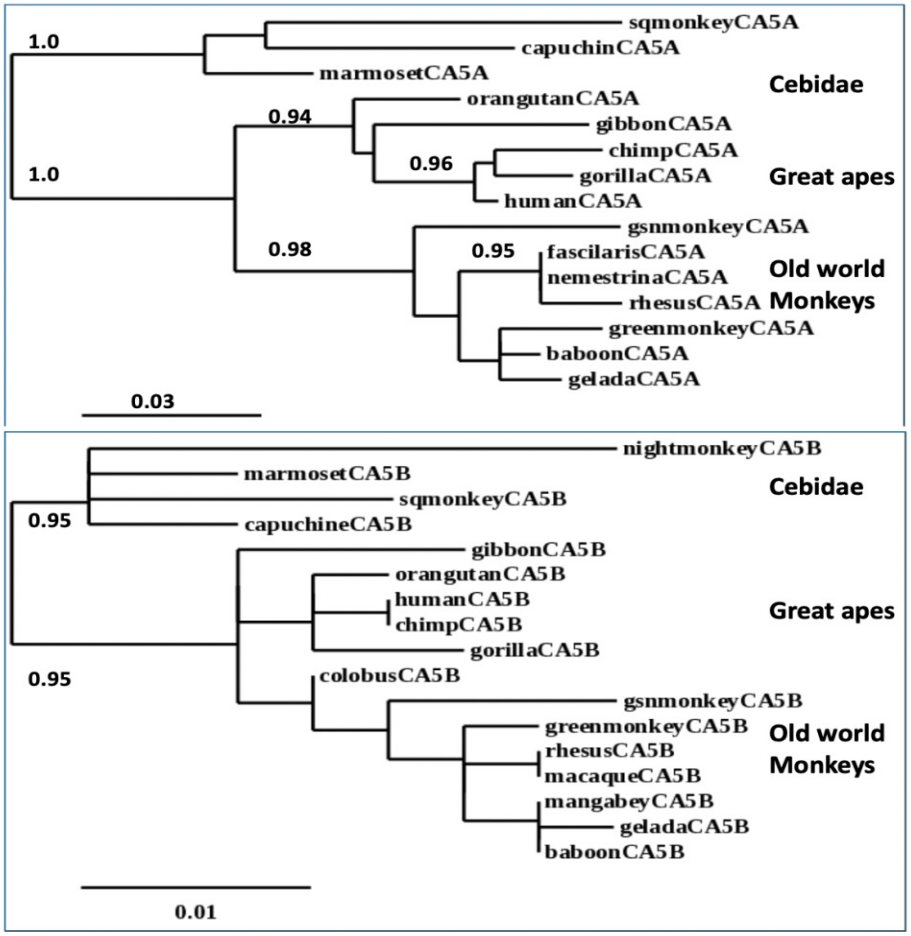

Figure 5: Phylogenetic trees of primate CA5A and CA5B sequences

The trees are labeled with the *CA5* gene name (*CA5A* for upper phylogram; and *CA5B* for lower phylogram) and the name of the primate; note 3 major clusters in each case for primates which are closely related phylogenetically; genetic distance scales are shown for each enzyme, with CA5A showing ~3 times larger genetic distances than for CA5B. The number of times a clade (sequences common to a node or branch) occurred in the bootstrap replicates are shown. Only replicate values of 0.9 or more which are highly significant are shown. 100 bootstrap replicates were performed in each case. Sequences were derived from those reported in Tables 3 and 4.

Table 4: Primate CA5B genes and proteins

*predicted sequence; ^scaffold IDs are shown; transcript IDs were derived from NCBI sources http://www.ncbi.nlm.nih.gov/genbank/; UNIPROT refers to UniprotKB/Swiss-Prot IDs for individual CA5B subunits (see http://kr.expasy.org); the number of coding exons are listed; 'na' means data not available.

		Location	ID*	(Strand)	ID	acids
Human	*Homo sapiens*	X:15,750,024-15,782,661	BC028142	7 (+)	Q9Y2D0	317
Chimp	*Pan troglodytes*	X:15,712,460-15,742,580	BC028142	7 (+)	H2R4T4	317
Gorilla	*Gorilla gorilla*	^CYUI0101497 5v1:11,824,638-11,857,316	*XP_005063896.1	7 (+)	na	317
Orang-utan	*Pongo abelii*	X:12,309,344-12,341,641	*XP_002831460.1	7 (+)	H2NRR0	317
Gibbon	*Nomascus leucogenys*	X:13,770,923-13,804,629	*XP_030663014.1	7 (+)	G1RE91	317
Baboon	*Papio anubis*	X:13,133,143-13,166,842	*XP_003917488	7 (+)	A0A096NAK3	317
Green monkey	*Chlorocebus sabaeus*	X:14,209,163-14,239,996	*XP_007989305.1	7 (+)	A0A0D9RQX6	317
Gelada monkey	*Theropithecus gelada*	na	*XP_025228522.1	na	na	317
Rhesus macaque	*Macaca mulatta*	X:15,449,497-15,482,570	*XP_014982249.2	7 (+)	I0FMW0	317
Pig-tailed macaque	*Macaca nemestrina*	20:76,285,953-76,343,242	*XP_011751584.1	7 (+)	A0A2K5WNP4	317
Crab-eating macaque	*Macaca fascicularis*	X:13,576,554-13,609,605	EHH60731.1	7 (+)	A0A2K5WNT1	317
Golden snub-nosed monkey	*Rhinopithecus roxellana*	^KN296004v1:1268-16370	*XP_030789461.1	7 (+)	na	317
Squirrel monkey	*Saimiri boliviensis*	^JH378105:71,210,293-71,210,293	*XP_003920396.1	7 (+)	A0A2K6S8K1	317
Capuchin monkey	*Cebus capucinus*	na	*XP_017374775.1	na	A0A2K5RPU0	317
Marmoset	*Callithrix jacchus*	X:13,917,420-13,953,682	*XP_002762698.1	7 (+)	F7A852	317

species. Moreover, this study of other vertebrate CA5-like genes and enzymes is consistent with CA5 being an ancient gene present throughout vertebrate evolution, including sharks, fish, frogs, reptiles and birds, which has undergone a major gene duplication event during the emergence of mammals, generating the separate CA5A and CA5B evolutionary pathways (Figure 4).

It may be noted however that the CA5B gene is consistently located on the X-chromosome in eutherian mammalian genomes but is located on chromosome 7 (an autosome) in the opossum genome (Monodelphis domestica) (Table 1). This may reflect on the evolution of the mammalian X-chromosome which is highly conserved among eutherians due to suppression of recombination between X and Y chromosomes [49,50].

Phylogenetic relationships among primate CA5A (Figure 5 and Table 3) and CA5B (Figure 5 and Table 4) were examined using known and predicted genomic and enzyme sequences for 15 primates which are representative of species separated by > 40 million years of primate evolution [51]. Both phylograms separated into 3 distinct groups, with species representative of Hominidae (great apes, including humans and related species); old world monkeys (including rhesus, baboons and related species); and Cebidae (marmosets and squirrel monkeys).

The results of this study supported previous reports for 2 distinct mammalian CA5-like genes and encoded mitochondrial CA5 enzymes, CA5A and CA5B, which are separately localized on chromosomes 16 and the X-chromosome respectively, in the human genome [6-20]. Reports of the 3D structures for the mouse enzymes have established roles for several residues within both mitochondrial enzymes [9,10,13]. The mouse Ca5a gene and enzyme (CA5A) sequence and structure served as model to study other vertebrate mitochondrial CA5-like genes and enzymes conserved sequences. Key conserved CA5-like amino acid residues and sequences included 3 conserved residues (His124; His 126; His149) which chelate Zinc, together with active site tyrosine residues (Tyr94; Tyr158; Tyr161), which catalyze the reversible carbonic anhydrase reactions [1-20]. N-terminal mitochondrial leader peptides play a key role to facilitate translocation of the enzymes into the mitochondrial matrix [5-7]. These leader peptides undergo significant sequence changes during mammalian and other vertebrate evolution, in direct contrast to those enzyme sequences involved in catalysis and other key functions. Gene expression data [32,38] showed that the human CA5A and CA5B genes are differentially expressed in tissues of the body, with CA5A expressed at a high level only in liver, whereas CA5B is broadly expressed with highest levels in testis and arteries. In addition, CA5B was expressed at higher levels than those for the average [33]. Several transcription factor binding sites were observed within the 5' regions for both human CA5 genes, including hepatocyte nuclear factor 4A (HNF4A) which controls the expression of several hepatic genes [41]; and CCAAT enhancer binding protein alpha (CEBPA), which plays a role in hepatocyte differentation [42].

CONCLUSION

Phylogeny studies examined several vertebrate CA5 subunits and demonstrated that this is an ancient gene in vertebrate evolution which appears to have undergone a gene duplication event in a mammalian ancestral gene prior to the appearance of monotreme, marsupial and eutherian genomes, generating 2 distinct related mitochondrial CA5 genes and enzymes, CA5A and CA5B. These enzymes have been shown to play key roles in ammonia detoxification and glucose metabolism, with similar 3D structures to other CA isozymes.

ACKNOWLEDGEMENT

The advice of Dr Laura Cox of the Centre for Precision Medicine, Wake Forest School of Medicine, Winston Salem NC USA is gratefully acknowledged.

CONFLICT OF INTEREST

The author reports no conflicts of interest.

REFERENCES

1. Lindskog S. Structure and mechanism of carbonic anhydrase. Pharmacol Ther. 1997:74:1-20.
2. Sly WS, Hu PY. Human carbonic anhydrases and carbonic anhydrase deficiencies. Ann Rev Biochem. 1995; 64:375-401.
3. Geers C, Gros G. Carbonic dioxide transport and carbonic anhydrase in blood and muscle. Physiol Rev. 2000; 80: 681-715.
4. McDevitt ME, Lambert LA. Molecular evolution and selection pressure in alpha-class carbonic anhydrase family members. Biochim Biophys Acta. 2011;1814:1854-61.
5. Frost SC. Physiological functions of the alpha class of carbonic anhydrases. Subcell Biochem. 2014; 75: 9-30.
6. Dodgson SJ, Forster RE, Storey BT, Mela L. Mitochondrial carbonic anhydrase. Proc Natl Acad Sci USA. 1980; 77: 5562–5566.
7. Nagao Y, Srinivasan M, Platero S, Svendrowski M, Waheed A, Sly WS. Mitochondrial carbonic anhydrase (isozyme V) in mouse and rat: cDNA cloning, expression, subcellular localization, processing and tissue distribution. ProcNatlAcadSci USA. 1994; 91: 10330-10334.
8. Fujikawa-Adachi K, Nishimori I, Taguchi T, Onishi S. Human mitochondrial carbonic anhydrase VB: cDNA cloning, mRNA expression, subcellular localization, and mapping to chromosome X. J Biol Chem. 1999; 274: 21228-21233.
9. Boriack-Sjodin PA, Heck RW, Paipis PJ, Silverman DN, Christianson DW. Structure determination of murine mitochondrial carbonic anhydrase V at 2.45-Ä resolution: implications for catalytic proton transfer and inhibitor design. Proc Nat lAcad Sci USA. 1995; 92: 10949-10953.
10. Jude KM, Wright SK, Tu C, Silverman DN, Viola RE, Christianson DW. Crystal structure of F65A/Y131C-methylimidazole carbonic anhydrase V reveals architectural features of an engineered proton shuttle. Biochem. 2002; 41:2485-91.
11. Nagao Y, Batanian JR, Clemente MF, Sly WS. Genomic organization of the human gene (CA5) and pseudogene for mitochondrial carbonic anhydrase V and their localization to chromosomes 16q and 16p. Genom. 1995; 28: 477-84.
12. Parkkila A-K, Scarim AL, Parkkila S, Waheed A, Corbett JA, Sly WS. Expression of carbonic anhydrase V in pancreatic beta cells suggests role for mitochondrial carbonic anhydrase in insulin secretion. J BiolChem. 1998; 278: 24620-24623.
13. Earnhardt JN, Qian M, Tu C, Laipis PJ, Silverman DN. Intramolecular proton transfer from multiple sites in catalysis by murine carbonic anhydrase V. Biochem. 1998; 37:7649-7655.
14. Postel R, Sonnenberg A. Carbonic anhydrase 5 regulates acid-base homeostasis in zebrafish. PLoS One 2012; 7:e39881.
15. Shah GN, Hewett-Emmett D, Grubb JH, Migas MC, Fleming RE, A Waheed , et al. 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 USA. 2000; 97:1677–1682.
16. Nishimori I, Innocenti A, Vullo D, Scozzafava A, Supuran CT. Carbonic anhydrase inhibitors: The inhibition profiles of the human mitochondrial isoforms VA and VB with anions are very different. Bioorg Med Chem. 2007; 15:6742–6747.
17. Davis RH, Innocenti A, Poulsen SA, Supuran CT. Carbonic anhydrase inhibitors. Identification of selective inhibitors of the human mitochondrial isozymes VA and VB over the cytosolic isoforms I and II from a natural product- based phenolic library. Bioorg Med Chem. 2010; 18:14–18.
18. Idrees D, Shahbaaz M, Bisetty K, Islam A, Ahmad F, Hassan MI. Effect of pH on structure, function, and stability of mitochondrial carbonic anhydrase VA. J Biomol Struct Dyn. 2017; 35: 449-461
19. Queen A, Khan P, Azam A, Hassan MI. Understanding the role and mechanism of carbonic anhydrase V in obesity and its therapeutic implications. Curr Protein Pept Sci. 2018; 19: 909-923.
20. Van Karnebeek CD, Sly WS, Ross CJ, Salvarinova R, Yaplito-Lee J, Santra S, et al. Mitochondrial carbonic anhydrase VA deficiency resulting from CA5A alterations presents with hyperammonemia in early childhood. Am J Human Genet. 2014; 94:453-61.
21. Altschul F, Vyas V, Cornfield A, Goodin S, Ravikumar TS, Rubin EH, et al. Basic local alignment search tool. J Mol Biol. 1990; 215: 403-410.
22. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409: 860-921.
23. Rogers J, Raveedran M, Harris RA, Mailund T, Leppala K, Athanasiadis G, et al. The comparative genomics and complex population history of Papio baboons. Sci Adv. 2019; 5: eaau6947.
24. The bovine genome sequencing and analysis consortium, Elsik CG, Tellam RL, Worley KC. The genome sequence of Taurine cattle: a window to ruminant biology and evolution. Science. 2009; 324: 522-528.
25. Mouse genome sequencing consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002: 420: 520-562.
26. Rat Genome Sequencing Project Consortium. Genome sequence of the brown Norway rat yields insights into mammalian evolution. Nature. 2004: 428: 493-521.
27. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, et al. Genome of the marsupial Monodelphisdomestica reveals innovation in non-coding sequences. Nature. 2007; 447:167-177.
28. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, et al (2008) Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008; 453:175-183.
29. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004; 432: 695-716.
30. Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V et al. The genome of the western clawed frog Xenopustropicalis. Sci. 2010; 328: 633–636.
31. Sprague J, Bayraktaroglu L, Bradford Y, Conlin T, Dunn N, Fashena D, et al. The zebrafish information network: the zebrafish model organism database. Nucleic Acids Res. 2005; 34:D581-D585.
32. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2003; 12:994-1006.
33. Thierry-Mieg D, Thierry-Mieg J. AceView: A comprehensive cDNA-supported gene and transcripts annotation. Genome Biol. 2006; 7:S12
34. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 2003; 31:3497-3500.
35. Claros MG. MITOPROT, a Macintosh application for studying mitochondrial proteins. Comput Appl Biosci. 1995; 11:441-7.
36. Elango N, Yi SV. Functional relevance of CpG island length for regulation of gene expression. Genetics 2011;187: 1077-1083.
37. Lesurf R, Cotto KC, Wang G, Griffith M, Kasaian K, Jones SJ, et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. Nucleic Acids Res. 2016; 44: D126-32.
38. The GTEX Consortium. The genotype-tissue expression (GTEx) project. Nature Genet. 2013; 45: 580-585.
39. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008; 36: W465–469.

40. Duciot F, Kabbaj M. The role of early response 1 (EGR1) in brain plasticity and neuropsychiatric disorders. Behav Neurosci. 2017; 11:35.

41. Mohlke KL, Boehnke M. The role of HNF4A variants in the risk of type 2 diabetes. Curr Diab Rep. 2005; 5:149-156.

42. Lourenco AR, Coffer PJ. A tumor suppressor role for C/EBFalpha in solid tumors: more than fat and blood. Oncogene. 2017; 36: 5221-5230.

43. Pennachio LA, Bickmore W, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013:14: 288-295.

44. Guanizo AC, Fernando CD, Garama DJ, Gough DJ. STAT3: a multifaceted oncoprotein. Growth Factors. 2018: 36: 1-14.

45. Von Both I, Silvestri C, Erdemir T, Lickert H, Walls JR, et al. FoxH1 is essential for the development of the anterior heart field. Dev Cell. 2004; 7:331-345.

46. Jiang Y, Yan L, Xia L, Lu X, Zhu W, Ding D, et al. Zinc finger E-box-binding homeobox 1 (ZEB1) is required for neural differentiation of human embryonic stem cells. J Biol Chem. 2018; 293:19317-29.

47. Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, et al. Genome evolution in the allotetraploid frog Xenopuslaevis. Nature. 2016; 538: 336-343.

48. Messer M, Weiss AS, Shaw DC, Westerman M. Evolution of the monotremes: phylogenetic relationship to marsupials and eutherians, and estimation of divergence days based on alpha-lactalbumin amino acid sequences. J Mammal Evol. 1998; 5:95-105.

49. Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Res. 2005;15: 98-110.

50. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 2005; 309: 613-617.

51. Steiper ME, Seiffert ER. Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. Proc Natl Acad Sci USA. 2012; 109: 6006-6011.