Comparative and Evolutionary Studies of *ALDH18A1* Genes and Proteins

Roger S Holmes

The Eskitis Institute for Drug Discovery and School of Natural Sciences, Griffith University,
Nathan 4111 QLD AUSTRALIA
Email: r.holmes@griffith.edu.au

ABSTRACT

Vertebrate *ALDH18A1* genes encode a bifunctional mitochondrial enzyme, catalyzing a 2-step conversion of glutamate to glutamyl semialdehyde, subsequently converted into proline, ornithine and arginine. Bioinformatic analyses of vertebrate and invertebrate genomes were undertaken using known *ALDH18A1* amino acid sequences. G5K (glutamyl kinase) and GPR (glutamyl phosphate reductase) domain sequences were identified for all vertebrate and invertebrate genomes examined, whereas bacterial sequences encoded separate enzymes. Vertebrate *ALDH18A1* (also called P5CS) sequences were highly conserved throughout vertebrate evolution. A mechanism for generating two major vertebrate *ALDH18A1* isoforms is proposed with 'a' isoform containing Asn239-Val240 with wide tissue expression, whereas the 'b' isoform lacking the dipeptide has been reported in gut tissues. Phylogenetic analyses describe the relationships and potential origins of the *ALDH18A1* gene during vertebrate and invertebrate evolution and a proposal for generating the bifunctional vertebrate and invertebrate *ALDH18A1* gene from a bacterial operon (*proBA*) encoding G5K and GPR. A more recent *Aldh18a1* gene duplication event has apparently occurred with a primordial rat genome.

Keywords: Aldehyde dehydrogenases; bifunctional enzyme; *ALDH18A1*; glutamyl kinase; GK; glutamyl phosphate reductase; GPR; vertebrates; invertebrates; evolution; phylogeny; primordial gene; gene integration; transcription factor binding sites; CpG islands.

Abbreviations: ALDH: aldehyde dehydrogenase; P5CS: delta 1-pyrroline-5-carboxylate synthase; GK: gamma-glutamyl kinase; GPR: gamma-glutamyl phosphate reductase; EC: Enzyme Commission; BLAST: Basic Local Alignment Search Tool; BLAT: Blast-Like Alignment Tool; NCBI: National Center for Biotechnology Information; AceView: NCBI Based representation of public mRNAs; TFBS: transcription factor binding sites; UTR: Untranslated Gene Region; CpG: region of high density of guanine-cytosine dinucleotides; mRNA: messenger RNA.

INTRODUCTION

The aldehyde dehydrogenase 18A1 gene (*ALDH18A1*) is one of 19 human *ALDH* genes encoding enzymes with diverse metabolic roles in aldehyde metabolism in the body (<http://aldh.org/superfamily.php>) [1]. Human *ALDH18A1* encodes a bifunctional enzyme, designated as delta 1-pyrroline-5-carboxylate synthase (P5CS), which catalyses the first two steps in proline, ornithine and arginine biosynthesis [2-3]:

Step 1: glutamate 5-kinase (domain G5K) EC=2.7.2.11

Step 2: gamma-glutamyl phosphate reductase (domain GPR) EC=1.2.1.41

Subsequent GSA metabolism results in the formation of proline, via pyrroline-5-carboxylate reductase (P5CR) (EC=1.5.12) [4]; ornithine, via ornithine aminotransferase (OAT) (EC=2.6.1.13) [5]; and arginine, via arginase (ARG) (EC=3.5.3.1) [6].

The human *ALDH18A1* and mouse *Aldh18a1* genes encode two major isoforms which differ by a 2 amino acid deletion within the G5K domain (239Asn-240Val), designated as *ALDH18A1long* and *ALDH18A1short* [3,7,8]. These isoforms differ in tissue distribution and kinetic properties, with the long version being widely distributed in most tissues of the body and insensitive to product inhibition, whereas the short isoform is restricted to gut tissues and is inhibited by L-ornithine at physiological concentrations. Several mutations for human *ALDH18A1* have been reported which confirm the key role for this gene and protein in the body. These include a rare autosomal recessive cutis laxa type 3A and Marburg Micro syndrome [9,10]; an autosomal dominant form of cutis laxa [11]; and an inherited spastic paraplegia condition [12].

This study describes the predicted sequences, structures and phylogeny of vertebrate and invertebrate *ALDH18A1* genes and enzymes and compares these results with those previously reported for the human and mouse genes and proteins [2,3,7,8]. Evidence is presented concerning the generation of *ALDH18A1* isoforms from several vertebrate species, and for distinct modes of gene regulation and expression with transcription factor binding sites (TFBS) and CpG islands identified for the human *ALDH18A1* gene. Two *Aldh18a1* genes were observed on chromosome 1 of the rat genome, apparently resulting from tandem duplication. Phylogenetic analyses also describe the relationships and potential origins of the *ALDH18A1* gene during vertebrate and invertebrate evolution and a proposal for generating the bifunctional vertebrate and invertebrate *ALDH18A1* gene from a bacterial operon (*proBA*) encoding G5K and GPR.

MATERIALS AND METHODS

***ALDH18A1* gene and enzyme identification**

ALDH18A1 amino acid sequences for vertebrate and invertebrate species were retrieved from databases, NCBI (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) and ExPASy (<http://www.expasy.org>) [13], using human, mouse, fruit fly and worm *ALDH18A1* sequences to seed searches. Identification of the genes was based on high predictive scores (>850) and sequence coverage (>98%) for *ALDH18A1* proteome

sequences (as listed by NCBI) in each case (Table 1 and Supplementary Table 1). BLAT searches were performed using relevant ALDH18A1 protein sequences to confirm the gene and enzyme sequences among the species examined using the UCSC Genome Browser [14]. Gene locations, predicted gene structures and protein subunit sequences were obtained for each ALDH18A1 examined showing identity with the respective ALDH sequences (Tables 1 and Supplementary Table 1). Representations of human *ALDH18A1* gene structures were obtained using the AceView (<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/>) web browser. Identification of potential *ALDH18A1* gene regulatory sites, including transcription factor binding sites (TFBS) and CpG islands within the respective gene promoter regions, was undertaken using the UCSC Human Genome Browser [14].

Predicted structures and properties of vertebrate ALDH18A1 subunits

Predicted secondary and tertiary structures for vertebrate G5K and GPR domain sequences were obtained using SWISS MODEL web tools [15]. The *Escherichia coli* G5K (pd:2J5T) [16] and human GPR [17] tertiary structures served as references for obtaining these structures.

Amino acid sequence alignments and phylogenetic analyses

Alignments of identified vertebrate G5K and GPR domain sequences were undertaken using Clustal Omega, a multiple sequence alignment program [18]. Phylogenetic analyses used several bioinformatic programs, coordinated using the <http://www.phylogeny.fr/> bioinformatic portal, to enable alignment (MUSCLE), curation (Gblocks), phylogeny (PhyML) and tree rendering (TreeDyn), to reconstruct phylogenetic relationships [19]. Sequences were identified as ALDH18A1 members, including the proposed primordial *ALDH18A1* genes and ALDH18A1 proteins identified for the fruit fly, *Drosophila melanogaster*, and round worm, *Caenorhabditis elegans*.

RESULTS AND DISCUSSION

Alignments of vertebrate and invertebrate G5K and GPR amino acid sequences

Amino acid sequence alignments for human, zebra fish (*Danio rerio*), round worm (*Caenorhabditis elegans*), and fruit fly (*Drosophila melanogaster*) G5K domain sequences (see Table 1) are shown in Figure 1. Comparisons of these sequences with the bacterial (*Campylobacter jejuni*) G5K sequence, for which the tertiary structure has been described (template pdb:2AKO) [20], enabled prediction

of secondary structures and likely key residues contributing to catalysis, structure and function. Active site residues (human G5K numbers used) binding the substrate (Ser117, Asp223, Asn246) or nucleotide (Ala80, Ser266-Asp267-Val268, Leu271-Phe272, Met305-Gly-306, Gly307, Met308, Glu309-Ala310-Lys311, Lys348) were conserved or underwent conservative substitutions for all of these G5K sequences. Twelve predicted α -helices and 9 β -sheets were observed for the human G5K sequence, including α 3 in the region for ATP binding, which was fully conserved among the sequences examined. (Figure 1). Percentage identities for these aligned vertebrate and invertebrate G5K domain sequences were >53% identical indicating that these sequences have been strongly conserved throughout evolution.

Amino acid sequence alignments for GPR domain sequences from these vertebrate and invertebrate species are shown in Figure 2. A comparison of these sequences with the human GPR sequence, for which the tertiary structure has been reported (template pdb:2H5G) [17], enabled identification of key residues for this domain (human GPR numbers used), including substrate binding (Glu581 and Cys612) or stabilizing the transition state for the catalyzed reaction (Asn499). In addition, a fully conserved dinucleotide-binding motif and NAD(P)⁺ binding domain (Leu501-Leu-Leu-Lys-Gly-Gly-Lys-Glu-Ala509) was observed for these vertebrate and invertebrate GPR sequences. Eighteen α -helices and 15 β -sheets were observed for the human GPR sequence, including α 19 and β 13 in the NAD(P)⁺ binding site, which was highly conserved among the sequences examined. (Figure 2). Overall percentage identities for these aligned vertebrate and invertebrate GPR domain sequences were >58% identical indicating that these were also strongly conserved.

Predicted gene locations and exonic structures for *ALDH18A1* genes and proteins

Table 1 and Supplementary Table 1 summarize the predicted locations, sizes and number of coding exons for vertebrate *ALDH18A1* genes examined, and of the encoded ALDH18A1 subunit amino acid sequences. These were based on BLAST interrogations of *ALDH18A1* databases (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) using the reported sequences for human and mouse ALDH18A1 [2,3,8] and BLAT analyses of vertebrate genomes using the UC Santa Cruz Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgBlat>) [13]. In all cases for the vertebrate *ALDH18A1* genes examined, there were 17 coding exons observed (Table 1 and Supplementary Table 1), whereas fruit fly and

roundworm *ALDH1A1* genes contained fewer coding exons (12 and 9, respectively). This suggested that the primordial vertebrate *ALDH18A1* gene encoded both G5K and GPR domain sequences, as for the invertebrate *ALDH18A1* genes examined, which is in contrast to the corresponding genes (*ProB* and *ProA*) in a bacterial species (*Escherichia coli*), which are localized together as separate genes within the PRO operon [21].

Table 1 and Supplementary Table 1 show comparative locations and coding exon compositions for vertebrate and invertebrate *ALDH18A1* genes. Two *ALDH18A1* genes were observed on rat chromosome 1 in closely located positions and with the same orientations, which is suggestive of a gene duplication event of an ancestral rat *ALDH18A1* gene.

Figure 3 presents the genomic sequences for the two major isoforms derived from vertebrate *ALDH18A1* genes, confirming previous reports for human and mouse *ALDH18A1* genes and proteins, for which a dipeptide (Asn239-Val240) is either present (*ALDH18A1*long or *ALDH18A1a*) or absent (*ALDH18A1*short or *ALDH18A1b*). As previously proposed [3,8], it would appear that these isoforms have resulted from slippage in gene splicing, with two GT sequences, separated by conserved AAAT/C nucleotide sequences, for all vertebrate *ALDH18A1* genes examined. This high level of conservation is extended into the proximate intronic sequence for 5 or 6 nucleotides, following which sequence variability was observed until the AG dinucleotide sequence, required for directing splicing specificity. It is apparent that these multiple *ALDH18A1* isoforms were present for all vertebrate genomes examined, which extends the biochemical rationale for these isoforms being differentially distributed and inhibited by ornithine to all vertebrate *ALDH18A1* genes examined.

Predicted transcription factor binding sites (TFBS) located in the promoter region for the human *ALDH18A1* gene were observed: two N-myc proto-oncogene binding sites located together in sequence, which have been implicated in activating growth-related genes [22]; three cyclic AMP-dependent transcription factors (ATF-1, ATF-2 and ATF-6 alpha) which trigger cell proliferation and transformation [23]; and GATA-1 or erythroid transcription factor, which serves as a general switch factor for erythroid development [24]. These binding sites are in addition to two previously reported p53-binding consensus sequences in the promoter region and in intron 1 of the human *ALDH18A1* gene [8]. The *ALDH18A1* promoter also contains a large CpG island (CpG94: genome size of 866 nucleotides) which may contribute

to the high level and wide tissue expression of this gene previously reported

(<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/>).

Phylogeny and evolution of vertebrate and invertebrate ALDH18A1 sequences

A phylogram (Figure 4) was calculated by the progressive alignment of vertebrate ALDH18A1 amino acid sequences, using invertebrate ALDH18A1 sequences (from *D. melanogaster* and *C. elegans*) (Table 1; Supplementary Table 1) to ‘root’ the tree. ALDH18A1 sequences were identified for all mammalian, bird, lizard, frog, bony fish and cartilaginous fish genomes examined. The phylogram demonstrates separation of these sequences into distinct groups consistent with their relatedness during vertebrate evolution, and suggests that these genes have been derived from an ancestral invertebrate *ALDH18A1* gene. The two rat *Aldh18a1* genes reported earlier in this paper encoded identical rat ALDH18A1 proteins with almost identical genomic sequences located in tandem on rat chromosome one, suggesting that these genes are a product of a recent gene duplication event not shared with the mouse *Aldh18a1* gene.

Figure 5 summarizes a working hypothesis for the evolution of vertebrate and invertebrate *ALDH18A1* genes:

1. A proposed primordial *ALDH18A1*-like gene was derived from a bacterial ancestor, for which G5K and GPR domains were encoded together on a *PRO* operon, which was responsible for proline biosynthesis in bacteria [21].
2. A gene fusion step for the G5K and GPR encoding genes took place within the primordial invertebrate genome, resulting in the formation of the *G5K-GPR* (or *ALDH18A1*) gene, which was integrated into subsequent invertebrate and vertebrate genes. This has been also described for *ALDH1L*-like genes and proteins, encoding mitochondrial and cytosolic folate dehydrogenases [26]; and for Frizzled-like CRD (cysteine-rich domain) [27] and ATP citrate lyase genes and proteins [28].
3. A subsequent gene duplication event occurred for an ancestral rat *Aldh18a1* gene, apparently via tandem duplication, resulting with the presence of consecutive *Aldh18a1* genes, on rat

chromosome one.

CONCLUSIONS

BLAST and BLAT analyses of several vertebrate and invertebrate genome databases were undertaken using amino acid sequences reported for human and mouse ALDH18A1 for interrogation of vertebrate and invertebrate genome sequences. Predicted amino acid sequences for these vertebrate and invertebrate ALDH18A1 subunits showed a high degree of sequence identities and the presence, in each case, of encoded G5K and GPR domain sequences. Secondary structure and key residue identification were undertaken using previous reports for bacterial G5K and human GPR. Bioinformatic analyses enabled the identification of putative gene regulation sites, including transcription factor binding sites (TFBS) and a CpG island within the human *ALDH18A1* gene promoter, in addition to the p53-binding consensus sequences, previously reported [8].

Sequences for vertebrate *ALDH18A1* genes were examined in the region enabling two isoforms to be generated, which play key roles in the regulation of proline and ornithine biosynthesis in the body [3,8]. The nucleotide sequence conservation observed in the region of the splicing junction responsible for the formation of the 2 major isoforms suggested that *ALDH18A1* isoform formation is subject to selection during vertebrate evolution to enable the ALDH18A1short isoform (minus the dipeptide sequence) to be expressed only in gut cells. Intestinal epithelial cells play a major role in glutamine and glutamate metabolism, particularly following heavy protein ingestion in the diet [29]. Given that the ALDH18A1short isoform is subject to feedback inhibition by ornithine [3,8], its presence in the gut during periods of high protein feeding, may contribute to increased availability of glutamate for circulation within the body.

Phylogenetic analyses also suggested that vertebrate and invertebrate *ALDH18A1* genes were derived from a gene integration event, during which *G5K (ProB)* and *GPR (ProA)* genes, located together on a bacterial operon [21], underwent gene fusion to form the primordial invertebrate *ALDH18A1* gene. A subsequent tandem gene duplication event of an ancestral rat *ALDH18A1* gene, apparently generated two rat *Aldh18a1* genes on rat chromosome one.

ACKNOWLEDGEMENTS

I thank Dr Laura Cox from the Texas Biomedical Research Institute and Dr Thomas Hurley from the Indiana University School of Medicine for useful discussions.

REFERENCES

- [1] Black W, Vasiliou V (2009) The aldehyde dehydrogenase gene superfamily resource center. *Human Genomics* 4:136-142.
- [2] Aral B, Schlenzig JS, Liu G, Kamoun P (1996) Database cloning human delta 1-pyrroline-5-carboxylate synthetase (P5CS) cDNA: a bifunctional enzyme catalyzing the first 2 steps in proline biosynthesis. *C R Acad Sci III* 319:171-8.
- [3] Hu CA, Lin WW, Obie C, Valle D (1999) Molecular enzymology of mammalian delta 1-pyrroline-5-carboxylate synthase. Alternative splice donor utilization generates isoforms with different sensitivity to ornithine inhibition. *J Biol Chem* 274:6754-62.
- [4] Merrill MJ, Yeh GC, Phang JM (1989) Purified human erythrocyte pyrroline-5-carboxylate reductase. Preferential oxidation of NADPH. *J Biol Chem* 264:9352-8.
- [5] Shen BW, Hennig M, Hohenester E, Jansonius JN, Schirmer T (1998) Crystal structure of human recombinant ornithine aminotransferase. *J Mol Biol* 277:81-102.
- [6] Wu G, Morris SM (1998) Arginine metabolism: nitric oxide and beyond. *Biochem J* 336:1-17.
- [7] Baumgartner MR, Hu CA, Almashanu S, Steel G, Obie C, Aral B, Rabier D, Kamoun P, Saudubray JM, Valle D (2000) Hyperammonemia with reduced ornithine, citrulline, arginine and proline: a new inborn error caused by mutation in the gene delta(1)-pyrroline-5-carboxylate synthase. *Human Mol Genet* 9:2853-8.
- [8] Hu CA, Khalil S, Zhaorigetu S, Liu Z, Tyler M, Wan G, Valle D (2008) Human delta 1-pyrroline-5-carboxylate synthase: function and regulation. *Amino Acids* 35:665-72.
- [9] Bicknell LS, Pitt J, Aftimos S, Ramadas R, Waw MA, Robertson SP (2008) A missense mutation in ALDH18A1, encoding delta1-pyrroline-5-carboxylate synthase (P5CS), causes an autosomal recessive neurocutaneous syndrome. *Eur J Hum Genet* 16:1176-86.
- [10] Skidmore DL, Chitayat D, Morgan T, Hinek A, Fischer B, Dimopoulou A, Somers G, Halliday W, Blaser S, Diambomba Y, Lemire EG, Kornak U, Robertson SP (2011) Further expansion of the phenotypic spectrum associated with mutations in ALDH18A1, encoding Δ^1 -pyrroline-5-carboxylate synthase (P5CS). *Amer J Hum Genet* 155A: 1848-56.
- [11] Fischer-Zirnsak B, Escande-Beillard N, Ganesh J, Tan YX, Al Bughaili M, Lin AE, Sahai I, Bahena P, Reichert SL, Loh A, Wright GD, Liu J, Rahikkala E, Pivnick EK, Choudhri AF, Krüger U, Zemojtel T, van Ravenswaaij-Arts C, Mostafavi R, Stolte-Dijkstra I, Symoens S, Pajunen L, Al-Gazali L, Meierhofer D, Robinson PN, Mundlos S, Villarroel CE, Byers P, Masri A, Robertson SP, Schwarze U, Callewaert B, Reversade B, Kornak U (2015) Recurrent De Novo Mutations Affecting Residue Arg138 of Pyrroline-5-Carboxylate Synthase Cause a Progeroid Form of Autosomal-Dominant Cutis Laxa. *Am J Hum Genet* 97:483-92.
- [12] Coutelier M, Goizet C, Durr A, Habarou F, Morais S, Dionne-Laporte A, Tao F, Konop J, Stoll M, Charles P, Jacoupy M, Matusiak R, Alonso I, Tallaksen C, Mairey M, Kennerson M, Gaussen M, Schule R, Janin M, Morice-Picard F, Durand CM, Depienne C, Calvas P, Coutinho P, Saudubray JM, Rouleau G, Brice A, Nicholson G, Darios F, Loureiro JL, Zuchner S, Ottolenghi C, Mochel F, Stevanin G (2015) Alteration of ornithine metabolism leads to dominant and recessive hereditary spastic paraplegia. *Brain* 138:2191-205.
- [13] Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H (2012) ExPASy: SIB bioinformatics portal. *Nucleic Acids Res* 40:W597-603.
- [14] Karolchik D, Bejerano G, Hinrichs AS, Kuhn RM, Miller W, Rosenbloom KR, Zweig AS, Haussler D, Kent WJ (2009) Comparative genomic analysis using the UCSC genome browser. *Methods Mol. Biol.* 395:17-34.
- [15] Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modelling server. *Nucleic Acids Res* 31:3381-5.

- [16] Marco-Marin C, Gil-Ortiz F, Pérez-Arellano I, Cervera J, Fita I, Rubio V (2007) A novel two-domain architecture within the amino acid kinase enzyme family revealed by the crystal structure of *Escherichia coli* glutamate 5-kinase. *J Mol Biol* 367:1431-46.
- [17] Papagrigoriou, E., Shafqat, N., Turnbull, A.P., Berridge, G., Hozjan, V., Kavanagh, K., Gileadi, O., Smee, C., Bray, J., Gorrec, F., Sundstrom, M., Arrowsmith, C., Weigelt, J., Edwards, A., Oppermann, U., Structural Genomics Consortium (SGC) (2006) Crystal structure of human pyrroline-5-carboxylate synthetase. Protein Data Bank: 2H5G.
- [18] Sievers F, Higgins DG (2014) Clustal omega. *Curr Protoc Bioinformatics* 48:3.13.1-16.
- [19] Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 36:W465-469.
- [20] Gorman J, Shapiro L, Burley SK, New York SGX Research Center for Structural Genomics (2009) Crystal structure of glutamate 5-kinase from *Campylobacter jejuni*. Protein Data Bank: 2AKO.
- [21] Deutsch AH, Rushlow KE, Smith CJ (1984) Analysis of the *Escherichia coli* proBA locus by DNA and protein sequencing. *Nucleic Acids Res* 12:6337-55.
- [22] Pession A1, Tonelli R, Fronza R, Sciamanna E, Corradini R, Sforza S, Tedeschi T, Marchelli R, Montanaro L, Camerin C, Franzoni M, Paolucci G (2004) Targeted inhibition of NMYC by peptide nucleic acid in N-myc amplified human neuroblastoma cells: cell-cycle inhibition with induction of neuronal cell differentiation and apoptosis. *Int J Oncol* 24:265-72.
- [23] Samten B1, Townsend JC, Weis SE, Bhounik A, Klucar P, Shams H, Barnes PF (2008) CREB, ATF, and AP-1 transcription factors regulate IFN-gamma secretion by human T cells in response to mycobacterial antigen. *J Immunol* 181:2056-64.
- [24] Kadri Z, Lefevre C, Goupille O, Penglong T, Granger-Locatelli M, Fucharoen S, Maouche-Chretien L, Leboulch P, Chretien S (2015) Erythropoietin and IGF-1 signaling synchronize cell proliferation and maturation during erythropoiesis. *Genes Dev* 29(24):2603-16.
- [25] Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. *Nature Rev. Genetics* 14:204-220.
- [26] Krupenko NI, Holmes RS, Tsybovsky Y, Krupenko SA (2015) Aldehyde dehydrogenase homologous folate enzymes: Evolutionary switch between cytoplasmic and mitochondrial localization. *Chem Biol Interact* 234:12-17.
- [27] Yan J, Jia H, Ma Z, Ye H, Zhou M, Su L, Liu J, Guo AY (2014) The evolutionary analysis reveals domain fusion of proteins with Frizzled-like CRD domain. *Gene* 533:229-39.
- [28] Gawryluk RM1, Eme LI, Roger AJ (2015) Gene fusion, fission, lateral transfer, and loss: Not-so-rare events in the evolution of eukaryotic ATP citrate lyase. *Mol Phylogenet Evol* 91:12-6.
- [29] Smith C, Marks AD, Lieberman MA (2005) Mark's Basic Medical Biochemistry: A clinical approach (second edition). Lippincott Williams & Wilkins.

LEGENDS FOR FIGURES

Figure 1: Amino acid sequence alignments for human, zebra fish, worm (*C. elegans*) and fruit fly (*D. melanogaster*) G5K domain (ALDH18A1) sequences

G5K refers to the glutamate kinase domain; see Table 1 for details of *ALDH18A1* genes and proteins; “*” shows identical residues for G5K domain sequences; “:” similar alternate residues; “.” less similar alternate residues; ATP binding residues (based on [16]) are shown; **bold** font shows known or predicted exon junctions; exon numbers refer to the human *ALDH18A1* gene; ^ designates residues involved in nucleotide binding; # for substrate binding residues; & designates dipeptide (Asn239-Val240) residues involved in isoform differences; predicted α -helices (white enclosed in black) and β -sheets (shaded) are numbered from the amino-terminus end.

Figure 2: Amino acid sequence alignments for human, zebra fish, worm (*C. elegans*) and fruit fly (*D. melanogaster*) GPR domain (ALDH18A1) sequences

GPR refers to the glutamyl phosphate reductase domain; see Table 1 for details of *ALDH18A1* genes and proteins; “*” shows identical residues for GPR domain sequences; “:” similar alternate residues; “.” less similar alternate residues; NAD(P)H binding residues and likely active site (A) residues (Asn499; Lys504; 581Glu; and 612Cys) (based on [17])

are shown; **bold** font shows known or predicted exon junctions; exon numbers refer to the human *ALDH18A1* gene; predicted α -helices (white enclosed in black) and β -sheets (shaded) are numbered from the amino-terminus end.

Figure 3: Sequences for vertebrate *ALDH18A1* gene regions encoding isoforms a (*ALDH18A1.long*) and b (*ALDH18A1.short*)

See Tables 1 and Supplementary Table 1 for sources of vertebrate *ALDH18A1* gene sequences; isoform a encodes the longer sequence containing the dipeptide (Asn239-Val240); the isoform b sequence lacks these residues; nucleotide sequences are aligned, showing identical residues (*) for the exonic (shown in capitals) and intronic (in lower case) sequences; 'splicing junction' specific nucleotide residues (GT (or gt)---AG) are shaded; human and coelacanth *ALDH18A1* amino acid sequences are shown, with differences from human sequence shaded.

Figure 4: Phylogenetic tree for vertebrate and invertebrate *ALDH18A1* sequences

The tree is labeled with the gene name and the name of the vertebrate or invertebrate. The tree is 'rooted' with the invertebrate *ALDH18A1* sequences. See Table 1 and Supplementary Table 1 for details of *ALDH18A1* genes and proteins. A genetic distance scale is shown. The number of times a clade (sequences common to a node or branch) occurred in the bootstrap replicates is represented as a fraction out of 100 (shown at each node). Only replicate values of 90 or more are highly significant, with 100 bootstrap replicates performed in each case.

Figure 5: A proposal for the evolutionary appearance of vertebrate and invertebrate *ALDH18A1* genes derived from the fusion of bacterial *ProB* and *ProA* genes

Proposed evolutionary appearance of vertebrate and invertebrate *ALDH18A1* genes and proteins resulting from a fusion of bacterial *ProB* (encoding G5K) and *ProA* (encoding GPR) genes. Gene details are shown in Table 1 (based on [21]).

LEGENDS for TABLES

Table 1: Vertebrate and invertebrate *ALDH18A1* genes and enzymes

RefSeq refers to the NCBI reference sequence; *predicted NCBI sequence; na-not available; ^gene scaffold ID

Supplementary Table 1: Vertebrate *ALDH18A1* genes and enzymes

RefSeq refers to the NCBI reference sequence; *predicted NCBI sequence; ** mRNA sequence; ^gene scaffold ID


```

      exon 9      α13      exon 10 α14      α15
EVKPA GP--TVEQQGEMARSGGRMLATLEPEQRAETIHHLADLLTDORDEITL 404 human
EVKPA GP--TVEQQGEMARSGRSLASLEPEQRSDIICALADLLTERKEEIL 407 zfish
NTKGYE--GPPIEEVAEKCRDAGRQLAALSNKERGAMVRHLAALLVDKEKYII 395 elegans
EATESANAVPVEVMAENARTGSRQMQLTPAQRASAVNTLADLLVSREKFIL 389 fruit fly
...      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

      α16      α17      exon 11      β10
LANKKDLEEA--EGRLAAPLLKRLSLSTSKLNLSLAIGLRQIAASSQDSVGRVLRRTRIAK 462 human
SANKKDMENAVNTGRSLSPAMLRSLSSSKLNSLSIGLRQIAVSSQDSVGRVLRRTRVAN 467 zfish
EANTQDLANAKS--AGLDPQLNRLKMTPEKIQDLHAGLNITADSATLVGRVLKVKVISE 454 elegans
DANAKDLAEAKS--GLAKPLLSRLSLNPAKLNLSVGLKQIAEDSHKNVGRVLRRTRLAD 448 fruit fly
**      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

      β11      β12      α18      β13      NAD(P)H      exon 12 α19
NLELEQVTVPIGVLLVIFESRPDCLPQVAAALATASGNGLLLKGGKEAAHSNRILHLTQE 522 human
NLELEQITVPIGVLLVIFESRPDCLPQVASALAIASGNALLKGGKEAANTNRILHELAQE 527 zfish
GLFLEQVTVPIGSLMVIFESRPDCLPQVASLAMASGNALLKGGKEAESNKALHALVQE 514 elegans
QLELKQVTVPIGVLLVIFESRPDCLPQVAAALAMASGNALLKGGKEAAHSNKALMELVKE 508 fruit fly
*      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

NAD(P)H→α20 | β14      exon 13      β15      α21      β16
ALSIHGV--KEAVQLVNTREEVEDLCRLDKMIDLIIPRGSSSQLVRDIQRAAKGIPVMGHS 580 human
ALSIHGV--KDAIQLVSTREEVEDLCRLDKMIDLIIPRGSSSQLVRDIQRAAKGIPVLGHS 585 zfish
ALGTHGFEMRDVTLVRSREDVADLLQLKDLIDLIIIPRGSSDLVRSMQEKSKGIPVLGHA 574 elegans
ALATVGA--EHAVSLVSTREEISDLLSMENHIDLIIPRGSSDLVRSIQQSLHIPVLGHA 566 fruit fly
**      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

      β17      α22      exon 14      β18      α23      α24
EGICHMYVDSEASVDKVTRLVRDSKCEYPAACNAETLLIHRDLLRTPFDQIIDMLRVE 640 human
EGICHVYVDHEASIDKAIKIIRDSKCDYPACNAETLLVHRDLRTPVFDQIIDMLRTE 645 zfish
EGVCHVYIDKDCDEQKAIQIVRDSKCDYPACNAETILIHKDLATAPFFDSLCSMFKA 634 elegans
EGVCHVYIDRDADLEKALRIARDAKCDYPACNAETLLIHEDLMSGATFGDVCNMLKRE 626 fruit fly
**      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

      β19      α25      exon 15      β20      α25      β21
QVKIHAGPKFASYLTFSPSEVKSLRTEYGDLELCIEVVDNVQDAIDHIHKYGSSTDVIV 700 human
QVKIHAGPKFASYLTFSPSEVKSLRTEYGDLECCIEVVDSMQEADHIHKYGSSTDVIV 705 zfish
GVKIHAGPKLAALLKFAPPPAESMSFEYGSLECTLEVVDNVEEAVAHIIRYGSGHTESII 694 elegans
GVKIIYAGPRLNQLTFGPPAAKSLKHEYGALECCIEVVPSLDEAINHIHYGSSHTDVIV 686 fruit fly
**      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

      α26      exon 16 β22      α27      β23      α28
TEDNTAEFFLQHVDSACVFNWASTRFSDGYRFGLGAEVGISTRIHARGPVGLEGLLT 760 human
TENNTAEQFLQLDSACVFNWASSRFADGYRFGLGAEVGISTARIHARGPVGLEGLLT 765 zfish
TENDNTAEHFLKHVDSACAFHNASTRFADGYRFGLGAEVGISTGRIHARGPVGVEGLLT 754 elegans
TENDNTAEHFLKHVDSACAFHNASTRFADGYRFGLGAEVGISTARIHARGPVGVEGLLT 746 fruit fly
**      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

      β24      α28      exon 17
KWLLRGRGDHVSDFSEHGSLKYLHENLPIQRNTN----- 795 human
KWVLRGEGHTVDFSEHGSMTYLHENLPVAQLLPGYRTTS----- 805 zfish
KWLLRGEGHLVEDFKN--GKYSYLHENLNPSEVYRALDAAGELKKATA 800 elegans
KWILEGQDHAAADFAEGGGRTWLHETLPLD----- 776 fruit fly
**      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *

```

Fig. 2. Amino acid sequence alignments for human, zebra fish, worm (*C. elegans*) and fruit fly (*D. melanogaster*) GPR domain (ALDH18A1) sequences. GPR refers to the glutamyl phosphate reductase domain; see Table 1 for details of ALDH18A1 genes and proteins; "*" shows identical residues for GPR domain sequences; ":" similar alternate residues; "-" less similar alternate residues; NAD(P)H binding residues and likely active site (A) residues (Asn499; Lys504; 581Glu; and 612Cys) (based on [19]) are shown; **bold** font shows known or predicted exon junctions; exon numbers refer to the human ALDH18A1 gene; predicted α -helices (white enclosed in black) and β -sheets (shaded) are numbered from the amino-terminus end.

	Isoform b	Isoform a
Human	AspAlaValValProProAlaGluProAsnSerAspLeuGlnGlyValAsn	
Human	GATGCTGTTGTCCTCCAGCTGAGCCCAACAGTGACCTGCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Mouse	GATGCCGTCGTCCCTCCAGCCGAGCCCAACAGTGATCTTCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Opossum	GATGCTGTTGTCCTCCAGCTGAGCCCAACAGTGACCTGCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Chicken	GATGCTGTTGTCCTCCAGCTGAGCCCAACAGTGATCTTCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Lizard	GATGCTGTTGTCCTCCAGCTGAGCCCAACAGTGATCTTCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Frog	GATGCTGTTGTCCTCCAGCTGAGCCCAACAGTGATCTTCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Zebrafish	GATGCCGTCGTCCCTCCAGCCGAGCCCAACAGTGACCTGCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Coelacanth	GATGCCGTCGTCCCTCCAGCCGAGCCCAACAGTGACCTGCAGGGGTAATgtgggtaaagtatttctaagggttagcatgct..	
Coelacanth	AspAlaValValProProAlaGluProAsnSerAspLeuGlnGlyValAsn	
	** ** *	
Human	ValIleSerValLysAspAsnAspSerLeuAlaAlaArgLeuAlaValGluMetThrAspLeuLeuIleVal	
Human	..tctgcaattctagGTTATTAGTGTAAAGATAATGATAGCTGGCTGCCGACTGGCTGTGGAATGAAACTGATCTCTTGATT	
Mouse	..ccogaattgtagGTATTAGTGTAAAGATAATGATAGCTGGCTGCCGACTGGCTGTGGAATGAAACTGATCTCTTGATT	
Opossum	..tctgaattctagGTATTAGTGTAAAGATAATGATAGCTGGCTGCCGACTGGCTGTGGAATGAAACTGATCTCTTGATT	
Chicken	..tgttaattctagGTATTAGTGTAAAGATAATGATAGCTGGCTGCCGACTGGCTGTGGAATGAAACTGATCTCTTGATT	
Lizard	..tgtatgtttctagGTAATCAGTGTGAAGGACAACGATAGCTTAGCAGCAGCGCTAGCAGTGGAAATGAAGACCGACCTGCTGATC	
Frog	..gggtttttctagGTAATCAGTGTGAAGGACAACGATAGCTTAGCAGCAGCGCTAGCAGTGGAAATGAAGACCGACCTGCTGATC	
Zebrafish	..ttttatccctagGTTATCAGCATTAAGGATAATGACAGCTTGGCTGCAAGGCTTGGCTGTGGAATGAGGGCAGATCTCTTATT	
Coelacanth	..ttttatccctagGTTATCAGTGTAAAGACAACGACAGCTTGGCTGCCGCTGGCTGTGGAATGAGGGCAGATCTCTTATT	
Coelacanth	MetIleSerValLysAspAsnAspSerLeuAlaAlaArgLeuAlaValGluMetLysAlaAsnLeuLeuIle	
	** *	

Fig. 3. Sequences for vertebrate *ALDH18A1* gene regions encoding isoforms a (*ALDH18A1*long) and b (*ALDH18A1*short). See Table 1 for sources of vertebrate *ALDH18A1* gene sequences; isoform a encodes the longer sequence containing the dipeptide (Asn239-Val240); the isoform b sequence lacks these residues; nucleotide sequences are aligned, showing identical residues (*) for the exonic (shown in capitals) and intronic (in lower case) sequences; 'splicing junction' specific nucleotide residues (GT (or gt)---AG) are shaded; human and coelacanth *ALDH18A1* amino acid sequences are shown, with differences from human sequence shaded.

Fig. 4. Phylogenetic tree for vertebrate and invertebrate *ALDH18A1* sequences.

Fig. 5. A proposal for the evolutionary appearance of vertebrate and invertebrate *ALDH18A1* genes derived from the fusion of bacterial *ProB* and *ProA* genes. Proposed evolutionary appearance of vertebrate and invertebrate *ALDH18A1* genes and proteins resulting from a fusion of bacterial *ProB* (encoding GSK) and *ProA* (encoding GPR) genes. Gene details are shown in Table 1 (based on [23]).

Table 1

Vertebrate and Invertebrate *ALDH18A1* genes and enzymes. RefSeq refers to the NCBI reference sequence; *predicted NCBI sequence; ** mRNA sequence; na-not available; 'gene scaffold ID; ('a') refers to a isoform; ('b') refers to b isoform with a 2 amino acid deletion.

Animal/ Organism	Species	Gene	RefSeq ID Prediction*	Chromosome Location	Coding Exons (strand)	UNIPROT ID	Amino acids (a')	Amino acids (b')
Human	<i>Homo sapiens</i>	<i>ALDH18A1</i>	NM_00107423	10:97,366,522-97,413,134	17 (-)	P54886	795	793
Orangutan	<i>Pongo abelii</i>	<i>ALDH18A1</i>	CR861217**	10:94,518,108-94,566,982	17 (-)	Q5R4M8	795	793
Baboon	<i>Papio anubis</i>	<i>ALDH18A1</i>	XP_009213311*	9:87,393,288-87,438,030	17 (-)	A0A096P5R9	795	793
Marmoset	<i>Callithrix jacchus</i>	<i>ALDH18A1</i>	XP_009008311*	12:82,657,023-82,693,391	17 (-)	F7HKL6	795	793
Mouse	<i>Mus musculus</i>	<i>Aldh18a1</i>	NM_019698	19:40,551,151-40,585,577	17 (+)	Q9Z110	795	793
Rat	<i>Rattus norvegicus</i>	<i>Aldh18a1a</i>	NM_001108524	1:266,506,407-266,535,023	17 (-)	D3ZIE9	795	793
		<i>Aldh18a1b</i>	NM_001108524	1:267,086,306-267,114,928	17 (-)	D3ZIE9	795	793
Cow	<i>Bos taurus</i>	<i>ALDH18A1</i>	NM_001046181	26:16,955,013-16,991,894	17 (-)	Q2KJH7	795	793
Dog	<i>Canis familiaris</i>	<i>ALDH18A1</i>	XP_005637610*	28:9,102,413-9,144,300	17 (-)	na	795	793
Rabbit	<i>Oryctolagus cuniculus</i>	<i>ALDH18A1</i>	XP_008268465*	18:44,300,355-44,345,877	17 (-)	G1SQ11	795	793
Tasmanian devil	<i>Sarcophilus harrisii</i>	<i>ALDH18A1</i>	XP_012395865*	G1834453:838,753-881,800	17 (+)	G3WCX8	799	797
Opussum	<i>Monodelphis domestica</i>	<i>ALDH18A1</i>	XP_007478785*	1:75,928,861-75,968,388	17 (+)	F7F6P3	801	799
Chicken	<i>Gallus gallus</i>	<i>ALDH18A1</i>	XP_004942436*	6:34,864,525-34,897,155	17 (+)	E1C8Q3	802	800
Lizard	<i>Anolis carolinensis</i>	<i>ALDH18A1</i>	XP_008104702*	3:38,006,120-38,021,212	17 (-)	na	799	797
Frog	<i>Xenopus tropicalis</i>	<i>ALDH18A1</i>	BC123965*	G1172668:4,179,713-4,197,560	17 (-)	Q05AS6	798	796
Zebra fish	<i>Danio rerio</i>	<i>ALDH18A1</i>	XP_009305061*	12:30,674,720-30,692,083	17 (+)	A4IGC8	805	803
Tetra	<i>Astyanax mexicanus</i>	<i>ALDH18A1</i>	XP_007259270*	na	na	na	801	na
Elephant shark	<i>Callorhynchus milii</i>	<i>ALDH18A1</i>	XP_007908752*	K1636174:238,473-283,477	17 (+)	na	799	na
Coelacanth	<i>Latimeria chalumnae</i>	<i>ALDH18A1</i>	XP_005994897*	JH126786:571,051-599,170	17 (+)	H3ASL2	798	796
Round worm	<i>Caenorhabditis elegans</i>	<i>ALH-13</i>	NM_077732	X:12,782,332-12,786,917	9 (+)	P54889	802	na
Fruit fly	<i>Drosophila melanogaster</i>	<i>CG-7470</i>	NM_001259948	3L:21,972,002-21,982,788	12 (+)	Q9VNW6	776	na
Bacterium	<i>Escherichia coli</i>	<i>PROB</i>	X00786#	16121228+	na	P0A7B5	367	na
		<i>PROA</i>	X00786#	16121229+	na	P07004	417	na