Journal of Data Mining in Genomics & Proteomics

Holmes, J Data Mining Genomics Proteomics 2017, 8:2
DOI: 10.4172/2153-0602.1000212

Research Article Open Access

Vertebrate Arylsulfatase K (ARSK): Comparative and Evolutionary Studies of the Lysosomal 2-Sulfoglucuronate Sulfatase

Roger S Holmes*

Griffith Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan, QLD, Australia

Abstract

Arylsulfatase K (ARSK) is one of 17 sulfatase gene family members encoded on the human genome for which a role has been recently identified as a lysosomal 2-sulfoglucuronate sulfatase. Vertebrate ARSK sequences shared 60-82% identity but only <27% identities with other arylsulfatase family members. Comparative enzyme structures were studied, including residues with predicted roles in forming N-glycosylation sites, Ca^{2+} binding and active site residues. Vertebrate *ARSK* genes usually contained 8 coding exons. A human *ARSK* gene promoter comprised CpG61 and multiple TFBS, which may be involved in signal transduction, transcription activation or regulating entry into cell division. Phylogenetic analyses examined evolutionary changes for the vertebrate *ARSK* and the invertebrate *SUL1* genes. In summary, a major role for this enzyme as a 2-sulfoglucuronate sulfatase is supported which has been conserved throughout vertebrate evolution.

Keywords: Vertebrates; Arylsulfatase K; Amino acid sequence; 2-sulfoglucuronate sulfatase; Chromosome 5

Abbreviations: ARSK: Arylsulfatase K; kbps: Kilobase Pairs; CpG Island: Multiple C (Cytosine)-G (Guanine) Dinucleotide Region; miRNA: microRNA Binding Region; BLAST: Basic Local Alignment Search Tool; BLAT: Blast-Like Alignment Tool; NCBI: National Center for Biotechnology Information; SWISS-MODEL: Automated Protein Structure Homology-Modeling Server

Introduction

Seventeen human sulfatase gene families and fourteen mouse sulfatase gene families encode sulfatases which catalyse the hydrolysis of a range of biological sulfate esters in the body [1-5]. The gene encoding arylsulfatase K (ARSK; EC 3.1.6.13) (*ARSK* in vertebrates; *Arsk* in rodents) was initially identified using bioinformatic methods through its conserved sulfatase active site signature sequence [1,6] and subsequently cloned, expressed in human cells, purified and biochemically characterized as an arylsulfatase [7], recently identified as catalysing the hydrolysis of the 2-O-sulfate group from 2-sulfoglucuronate [8]. Kinetic properties for human ARSK were similar to those of several other lysosomal sulfatases and an established role in the degradation of sulfated glycosaminoglycans has been identified [8]. In addition, low sequence identities with other human sulfatases have been reported, indicating that this gene and enzyme represents a distinct form of human sulfatase. Broad *ARSK* mRNA expression in human tissues has been reported which suggested that a ubiquitous biological arylsulfate substrate was the target for ARSK physiologically [7]. Previous studies have shown that it comprises several domains: an N-terminus signal peptide (residues 1-22); five Ca^{2+} binding sites (1 Ca^{2+} per subunit); two active site residues (313Asp and 314His); and multiple N-glycosylation sites [7].

This paper examines comparative amino acid sequences for vertebrate ARSK proteins, particularly in relation to the identification of key structures for this conserved enzyme; exonic structures for vertebrate *ARSK* genes; potential sites for regulating human *ARSK* gene expression; alignment and evolutionary studies of vertebrate ARSK and invertebrate arylsulfatase (SUL1) and comparisons of ARSK structures with other sulfatase gene families.

Methods

Gene (*ARSK*) and protein (ARSK) structures

Vertebrate ARSK amino acid sequences were derived from BLAST studies using NCBI web tools (http://www.ncbi.nlm.nih.gov/) [9] using the human ARSK sequence [1,7] (Table 1). BLAT analyses were subsequently undertaken for each of the predicted ARSK amino acid sequences using the UC Santa Cruz (UCSC) Genome Browser to obtain predicted locations, exon boundaries and gene sizes for each of the vertebrate *ARSK* and *SUL1* genes (Table 1) [10]. Structural features for the major human mRNA *ARSK* isoform were also examined using AceView [11].

Structures and properties of vertebrate ARSK

Predicted structures for vertebrate ARSK proteins were obtained using the SWISS-MODEL web-server (http://swissmodel.expasy.org/) [12] and a tertiary structure reported for a putative sulfatase from *Bacteroides thetaiotaomicron* (PDB:1b5q), with a modelling residue range of 35-475 for human ARSK [13]. Predicted signal peptide cleavage sites, N-glycosylation sites, molecular weights and theoretical isoelectric points for vertebrate ARSK proteins were obtained using Expasy web tools (http://au.expasy.org/tools/pi_tool.html).

Human ARSK tissue expression

Comparative tissue expression levels for human *ARSK* mRNA were obtained using a GTEx database [14] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (http://www.gtex.org).

***Corresponding author:** Roger S Holmes, Griffith Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan, QLD, Australia, Tel: +61 410583348; E-mail: r.holmes@griffith.edu.au

Received July 12, 2017; **Accepted** August 14, 2017; **Published** August 21, 2017

Citation: Holmes RS (2017) Vertebrate Arylsulfatase K (ARSK): Comparative and Evolutionary Studies of the Lysosomal 2-Sulfoglucuronate Sulfatase. J Data Mining Genomics Proteomics 8: 212. doi: 10.4172/2153-0602.1000212

Amino acid sequence alignments, identities and phylogenetic analyses

Clustal Omega was used to generate amino acid sequence alignments and percentage identities for vertebrate ARSK and other human arylsulfatase sequences (Tables 1 and 2) [15]. Phylogenetic analyses used the http://www.phylogeny.fr/ bioinformatic portal [16] to reconstruct vertebrate ARSK evolutionary relationships with a *C. elegans SUL1* gene and protein (Table 1).

Results and Discussion

ARSK and other human sulfatase genes and proteins: Proposed classification

Table 2 summarises a classification scheme for 17 human sulfatase genes and proteins previously proposed, based on phylogenetic and amino acid sequence comparisons for human SULF1 and SULF2 (extracellular sulfatases 1 and 2) [5] and the other 15 human sulfatases, including the human *ARSK* gene and protein. Seven groups of these genes and proteins were identified, including ARS group 5 for a single human *ARSK* gene, which encodes a distinct 526 amino acid sequence enzyme and shares <27% sequence identities with other human arylsulfatases. This is in contrast to the sequence identities observed for other human enzyme ARS groups: Group 1 (ARSA, ARSG and GALNS) (33-39% identical); group 2 (ARSB, ARSI and ARSJ) (54-59%) [2]; group 3 (ARSD, ARSE, ARSF, ARSH and STS) (57-64%) [3]; and group 4 ARS sequences (SULF1, SULF2 and GNS) (42-67%) [5]. Groups 6 (SGSH) and 7 (IDS) [4] enzymes, however, exhibited distinct amino acid sequences which were 22% and 26% identical, respectively, with the human ARSK amino acid sequence (Table 2). With the exception of the group 3 gene cluster previously reported on the human X-chromosome for *ARSD, ARSE, ARSF, ARSH* and *STS* [3], other human ARS genes are separately located on the human genome, including ARSK on human chromosome 5 [6,7].

Gene	Organism	Species	Chromosome^ Location	Coding Exons (strand)	Gene Size bps	GenBank ID*	UNIPROT ID	Amino acids	Subunit MW (pI)	Leader Peptide
ARSK	Human	*Homo sapiens*	5:95,555,279-95,603,523	8 (+ve)	48,245	NM_198150	Q6UWY0	536	61,450 (9.0)	1…22
ARSK	Baboon	*Papio anubis*	6:89,585,690-89,631,202	8 (+ve)	45,513	XP_003899979*	A0A096MMD3	536	61,368 (8.8)	1…15
ARSK	Mouse	*Mus musculus*	13:76,062,259-76,098,499	8 (-ve)	36,241	NM_029847	Q9D2L1	556	63,100 (8.7)	1…18
ARSK	Cow	*Bos taurus*	7:97,276,428-97,331,337	8 (+ve)	54,910	BC118383+	Q148F3	540	61,361 (9.1)	1…15
ARSK	Tasmanian devil	*Sarcophilus harrisii*	^GL841273:981,226-1,016,664	8 (-ve)	35,439	XP_003759881*	G3WFM9	553	63,325 (8.3)	1…21
ARSK	Chicken	*Gallus gallus*	Z:57,238,252-57,248,657	8 (-ve)	10,406	NM_001031415	Q5ZK90	535	61,341 (8.7)	1…24
ARSK	Frog	*Xenopus tropicalis*	^KB021649:36,197,268-36,213,144	8 (-ve)	15,877	XP_002933485*	Q0IHJ2	536	60,608 (6.5)	1…24
ARSK	Zebra fish	*Danio rerio*	5:50,252,467-50,263,734	8 (+ve)	11,268	NM_001077157	Q08CJ7	523	59,470 (6.3)	1…16
SUL1	Worm	*Caenorhabditis elegans*	X:3,267,384-3,270,766	16 (-ve)	3,231	NM_076159	A8XJG0	704	83,303 (8.8)	1…20

Table 1: Vertebrate *ARSK* and worm *SUL1* genes and proteins. RefSeq: the reference amino acid sequence; *Predicted amino acid sequence; GenBank IDs are derived from NCBI http://www.ncbi.nlm.nih.gov/genbank/; UNIPROT refers to UniprotKB/Swiss-Prot IDs for individual ARSK and SUL1 proteins (http://kr.expasy.org); ^ refers to a scaffold; bps refers to base pairs of nucleotide sequences; pI refers to theoretical isoelectric points; the number of coding exons are listed.

ARS Group	Gene	EC Number	Chromosome Location	Coding Exons (strand)	Gene Size bps	GenBank ID	UNIPROT ID	Amino acids	Subunit MW (pI)	% Identity ARSK
1	*ARSA*	3.1.6.8	22:51,066,606-51,061,176	8 (-ve)	2,626	NM_000487	P15289	507	53,588 (5.6)	25
	ARSG	3.1.6.-	17:68,307,494-68,420,460	11 (+ve)	112,967	NM_001267727	Q96EG1	525	57,061 (6.2)	20
	GALNS	3.1.6.4	16:88,880850-88,923,285	14 (-ve)	42,436	NM_000512	P34059	522	58,026 (6.3)	21
2	*ARSB*	3.1.6.12	5:78,076,223-78,281,071	8 (-ve)	204,849	NM_000046	P15848	533	59,687 (8.4)	19
	ARSI	3.1.6.13	5:150,297,217-150,302,373	2 (-ve)	5,157	NM_001012301	Q5FYB1	569	64,030 (8.8)	19
	ARSJ	3.1.6.-	4:113,902,277-113,978,834	2 (-ve)	76,558	NM_024590	Q5FYB0	599	67,235 (9.2)	21
3	*ARSD*	3.1.6.1	X:2,907,274-2,929,275	10 (-ve)	22,002	NM_009589	P51689	593	64,859 (6.8)	24
	ARSE	3.1.6.1	X:2,934,835-2,958,434	10 (-ve)	23,600	NM_000047	P51690	589	65,669 (6.5)	24
	ARSF	3.1.6.1	X:3,072,024-3,112,553	10 (+ve)	40,530	NM_001201538	P54793	590	65,940 (6.8)	24
	ARSH	3.1.6.1	X:3,006,613-3,033,382	9 (+ve)	26,770	NM_001011719	Q5FYA8	562	63,525 (8.5)	24
	STS	3.1.6.2	X:7,253,194-7,350,258	10 (+ve)	97,065	NM_001320750	P08842	583	65,492 (7.6)	23
4	*SULF1*	3.1.6.-	8:69,563,976-69,638,860	18 (+ve)	74,885	NM_001128204	Q8IWU6	871	101,027 (9.2)	21
	SULF2	3.1.6.-	20:47,659,398-47,757,363	20 (-ve)	97,966	NM_001161841	Q8IWU5	870	100,455 (9.3)	22
	GNS	3.1.6.14	12:64,716,744-64,759,276	14 (-ve)	42,353	NM_002076	P15586	552	62,081 (8.6)	22
5	*ARSK*	3.1.6.-	5:95,555,279-95,603,523	8 (+ve)	48,245	NM_198150	Q6UWY0	526	61,450 (9.0)	100
6	*SGSH*	3.10.1.1	17:80,210,455-80,220,313	8 (-ve)	9,859	NM_000199	P51668	502	56,695 (6.5)	22
7	*IDS*	3.1.6.13	X:149,482,749-149,505,137	9 (-ve)	22,389	NM_000202	P22304	550	61,873 (5.2)	26

Table 2: Proposed classification of ARSK and other human arylsulfatase genes and proteins. Based on a previous proposal for the proposed classification of human arylsulfatase genes and proteins into 7 groups [5]; ARSK data is highlighted in red; pI=Isoelectric point; bps=Base pairs of nucleotide sequence; % identities with the human ARSK amino acid sequence are shown; ARSA: arylsulfatase A; ARSG: Arylsulfatase G; GALNS: N-acetylgalactosamine-6-sulfatase; ARSB: Arylsulfatase B; ARSI: Arylsulfatase I; ARSJ: Arylsulfatase J; ARSD: Arylsulfatase D; ARSE: Arylsulfatase E; ARSF: Arylsulfatase F; ARSH: arylsulfatase H; STS: Sterylsulfatase; SULF1: Extracellular sulfatase 1; SULF2: Extracellular sulfatase 2; GNS: N-acetylglucosamine-6-sulfatase; SGSH: N-sulfatase N-sulphoglucosamine sulphohydrolase; IDS: Iduronate 2-sulfatase.

Vertebrate ARSK and SUL1 amino acid sequences

The amino acid sequence for human ARSK [6,7] and the deduced amino acid sequences for mouse (*Mus musculus*), chicken (*Gallus gallus*) and zebrafish (*Danio rerio*) ARSK are shown in Figure 1 (Table 1). These and other vertebrate ARSK sequences were more than 60% identical, suggesting that they are members of the same arylsulfatase gene family. In contrast, the human ARSK protein sequence were < 27% identical with other human arylsulfatase families (Table 2), confirming the separate family status for ARSK enzymes. Vertebrate ARSK enzymes contained 523-556 amino acids (Figure 1 and Table 1), including key residues and arylsulfatase domains previously reported [6,7]: Active site residues (human ARSK numbers used) binding calcium ions (Ca^{2+}) (40Asp, 313Asp, 314His) or substrate (80Cys; 128Lys; 208His; 251His; 326Lys) were conserved, including 80Cys, which is post-translationally modified to form C(alpha)-formylglycine by sulfatase modifying factor 1 (SUMF1) within the active site of related sulfatase gene families [1].

The N-terminus leader peptide (1-22) revealed similarities in sequence, containing multiple hydrophobic residues. Seven N-glycosylation sites conserved for the mammalian ARSK sequences, including 108Asn, 166Asn, 193Asn, 262Asn, 375Asn, 413Asn and 498Asn, are designated as N-glycosylation sites 1-7 (Figure 1). These were found for all primate ARSK sequences examined (human, chimp, gorilla, gibbon, squirrel monkey, baboon, rhesus monkey, tarsier and marmoset), although the orangutan (*Pongo abelii*) ARSK sequence, lacked N-glycosylation site 1. Mouse and rat ARSK sequences also contained N-glycosylation sites 1-7, whereas the cow (*Bos taurus*), pig (*Sus scrofa*), horse (*Equus caballus*) and sheep (*Ovis aries*) ARSK sequences lacked N-glycosylation site 2 (data not shown). The chicken ARSK sequence contained seven N-glycosylation sites, although sites 1 and 2 were located in different positions within the protein. The zebra fish ARSK sequence contained six N-glycosylation sites, retaining sites 1, 4 and 6 as for the human sequence, but incorporating 3 other sites throughout the sequence. Multiple N-linked glycosylation sites were a feature of this enzyme ensuring that the vertebrate ARSK glycoprotein is suitably localized within the cell to perform its metabolic functions.

Vertebrate ARSK protein structures

Predictions of secondary and tertiary structures for human ARSK are presented in Figures 1 and 2, generated using the tertiary structure for a putative sulfatase from *Bacteroides thetaiotaomicron* (PDB:1b5qB) (modelling residue range of 35-475). A predicted human ARSK secondary structure for residues 476-536 was obtained using the SWISS-MODEL web-server [12]. The predicted 3D structure for ARSK shows a large active site cleft containing a metal (Ca^{2+}) ion. The enzyme has 2 major domains, with the active site at the base of a cleft on the larger domain, corresponding to the N-terminal region of the enzyme, with the other domain predominantly comprising the C-terminal domain, containing a sequence of beta sheets (β12-14) and a large alpha helix (α13). The C-terminal α13 helix was not fully identified in the predicted ARSK 3D structure since the modeling range fell short of the complete C-terminal sequence, as were two other C-terminal α helices (α14 and α15). 3D structures for other human arylsulfatase sequences have been previously reported, revealing similar tertiary subunit structures, reflecting strong conservation in amino acid sequences and structures [17-20].

Human *ARSK* tissue expression

Figure 3 summarizes the tissues distribution profile for *ARSK* transcripts from human tissues. A previous study has reported a broad *ARSK* mRNA expression profile in human tissues suggesting that a ubiquitous biological arylsulfate substrate was available for ARSK physiologically [4], which has been recognized as lysosomal 2-sulfoglucuronate sulfate [7]. *ARSK* displayed highest expression levels in fibroblasts and tibial nerve cells, and modest expression levels in other tissues of the body but with very low whole blood expression levels.

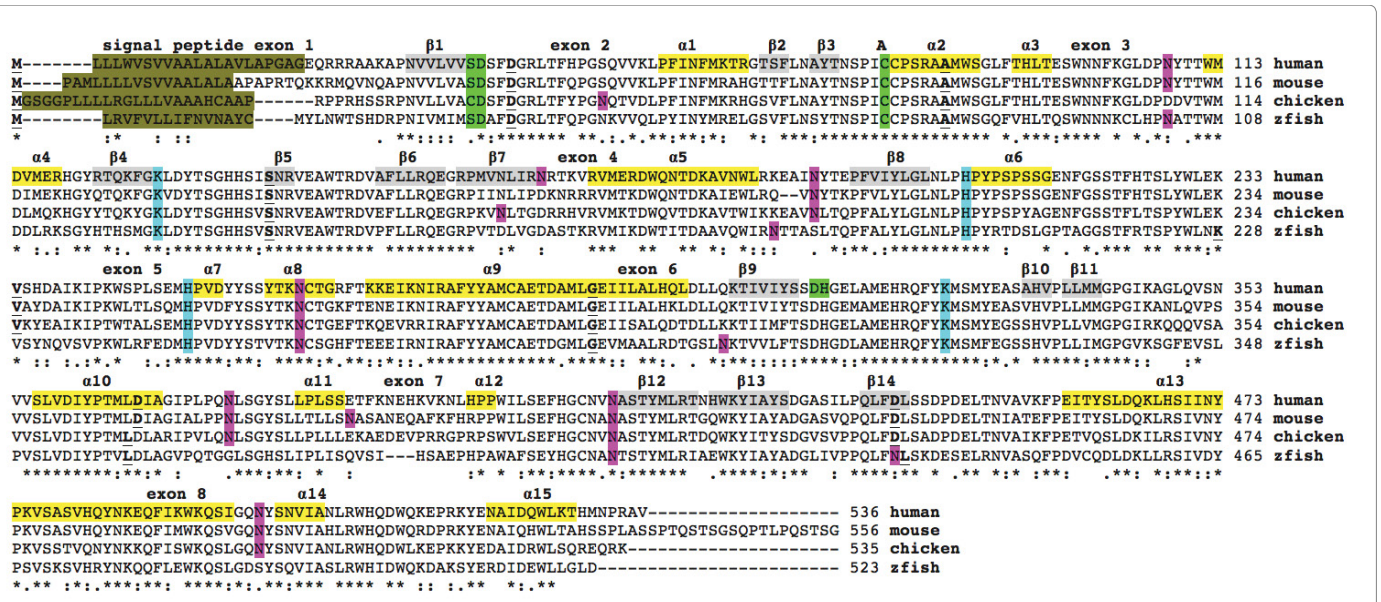

Figure 1: Amino acid sequence alignments for vertebrate ARSK sequences. Sources of ARSK sequence sources are given in Table 1; * shows identical residues for ARSK subunits; : similar alternate residues; . dissimilar alternate residues; predicted N-signal peptides residues are in kkaki; N-glycosylated and potential N-glycosylated Asn sites are in pink; the Ca^{2+} binding residues are shown in green; active site residues are shown in blue; A refers to active site Cys80; predicted α-helices for human ARSK are in shaded yellow and predicted β-sheets are in grey and numbered in sequence from the N-terminus; bold underlined font shows residues corresponding to predicted exon start sites; exon numbers refer to human *ARSK* gene exons.

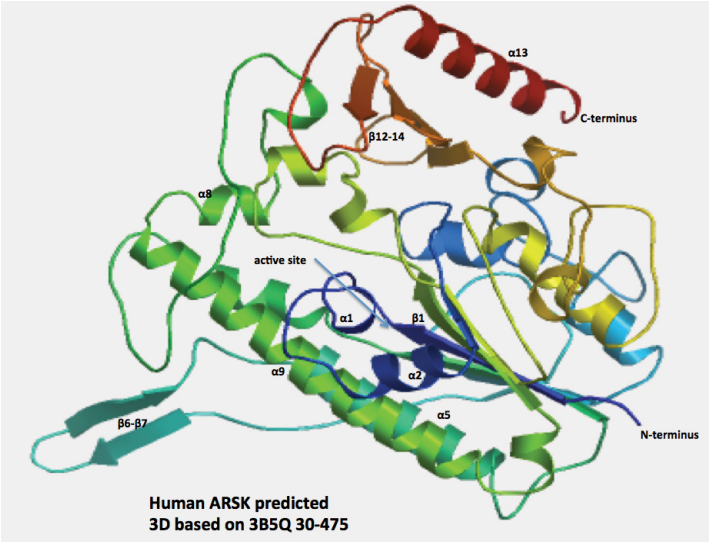

Figure 2: Predicted tertiary structure for human ARSK. Predicted human ARSK subunit structure is based on the structure for a putative sulfatase from *Bacteroides thetaiotaomicron* (PDB:1b5qB), with a modelling residue range of 30-475 and obtained using the SWISS MODEL web site [12]. The rainbow color code describes the 3-D structures from the N- (blue) to C-termini (red color); predicted α-helices, β-sheets and proposed active site cleft are shown.

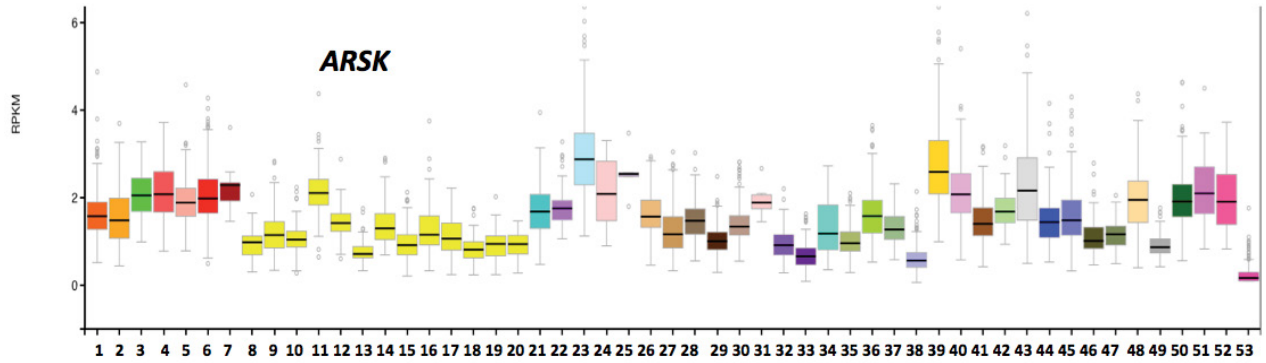

Figure 3: Tissue expression for human *ARSK* gene. RNA-seq gene expression profiles across 53 selected tissues (or tissue segments) were examined from the public database for human *ARSK*, based on expression levels for 175 individuals [14] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (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.

Gene locations, exonic structures and regulatory sequences for vertebrate ARSK genes

The predicted chromosome and strand locations and exonic structures for vertebrate *ARSK* genes are presented in Table 1, together with a proposed ancestral invertebrate *SUL1* gene. The predicted vertebrate *ARSK* genes were transcribed on the negative DNA strand (mouse, tasmanian devil (marsupial genome), chicken and frog genomes) or the positive DNA strand (human, baboon, cow and zebra fish genomes). Predicted exonic start sites for human, mouse, chicken and zebra fish *ARSK* genes (8 coding exons in each case) are shown in Figure 1, in identical or similar positions to those predicted for the human *ARSK* gene.

The human *ARSK* transcript was ~50 kbps in length with CpG61 and several Transcription Factor Binding Sites (TFBS) located in the 5'-untranslated promoter region of human *ARSK* on chromosome 5, and a 3'-untranslated region (UTR) lacking any detectable microRNA target sites (Figure 4). CpG61 contained 762 bps with a C plus G count of 463 bps, a C or G content of 61% and showed a ratio of observed to expect CpG of 0.87. This *ARSK* CpG Island may play a key role in gene regulation and may contribute to the broad gene expression observed in human tissues (Figure 3) [7,8,21,22]. Ten TFBS sites were colocated with CpG61 in the human *ARSK* promoter region which may contribute to the ubiquitous tissue expression. Of special interest among these identified *ARSK* TFBS were the following: *SMARCA4* (2 sites), encoding transcriptional

activator BRG which controls transcription of genes important for the G1/S phase transition of the cell cycle [23]; *STAT1*, a signal transducer, transcription activator and modulator of retinoblastoma protein phosphorylation; *RB1* (2 sites), retinoblastoma associated protein, and *RBL2* (3 sites), are key regulators of entry into cell division [24]; *ETS1*, protein C-ets-1 transcription factor, which controls the expression of cytokine and chemokine genes [25]; and *CTCF*, CCCTC-binding factor, which organizes higher order chromatin structure, regulates gene expression and plays a role in learning and memory [26].

Evolution of ARSK and other ARS sequences

A phylogenetic tree (Figure 5) was calculated by aligning vertebrate ARSK amino acid sequences, which was 'rooted' with a worm (*Caenorhabditis elegans*) arylsulfatase (SUL1) sequence (Table 1). The phylogram showed clustering of the vertebrate ARSK sequences which were consistent with their evolutionary relatedness and separation of the ARSK protein group, which was distinct from the worm SUL1 sequence. It is apparent that ARSK is an ancient protein in vertebrate evolution, being encoded among all of the vertebrate genomes examined. In addition, a proposed common ancestor for *ARSK* and other ARS genes may have arisen during early vertebrate or invertebrate evolution, which predated the appearance of fish during vertebrate evolution. Moreover, multiple vertebrate sulfatase genes may have arisen independently of the *ARSK* gene duplication step, generating many arylsulfatase genes, consistent with the evolutionary appearance of seven distinct groups and 17 human sulfatase gene families (Table 2).

Conclusion

Vertebrate *ARSK* genes and encoded proteins represent a distinct gene and protein arylsulfatase family which share key conserved sequences and domains reported for other arylsulfatase proteins previously studied [2-5,17-20]. The metabolic role for ARSK has recently been established and the natural substrate for this enzyme described as lysosomal 2-sulfoglucuronate, a key enzyme in the catabolism of heparin sulfate and dermatan sulfate [8]. A single gene (*ARSK/Arsk*) encodes this enzyme among the vertebrate genomes studied (Table 1), which is moderately expressed in a wide range of human tissues, but with highest levels of expression in fibroblasts (Figure 4). *ARSK* usually contained 8 coding exons on the negative or positive strands, depending on the vertebrate genome (Table 1). The human *ARSK* gene contained a large CpG island within the promoter region, with several transcription factor binding sites collocated within the *ARSK* gene promoter region (Figure 4). Predicted secondary and tertiary structures for human ARSK showed similarities with reported 3D structures for other arylsulfatases [17-20]. Structural domains reported for human ARSK, included the N-terminal signal peptide; the active site (including a Ca^{2+} binding site), which is responsible for arylsulfatase activity; and seven predominantly conserved N-glycosylation sites, which ensured that ARSK was suitably micro localized within the cell as a glycoprotein. Phylogenetic studies suggested that the *ARSK* gene appeared early in evolution, prior to the appearance of bony fish.

Acknowledgements

I thank Dr. Laura Cox from the Texas Biomedical Research Institute for useful discussions.

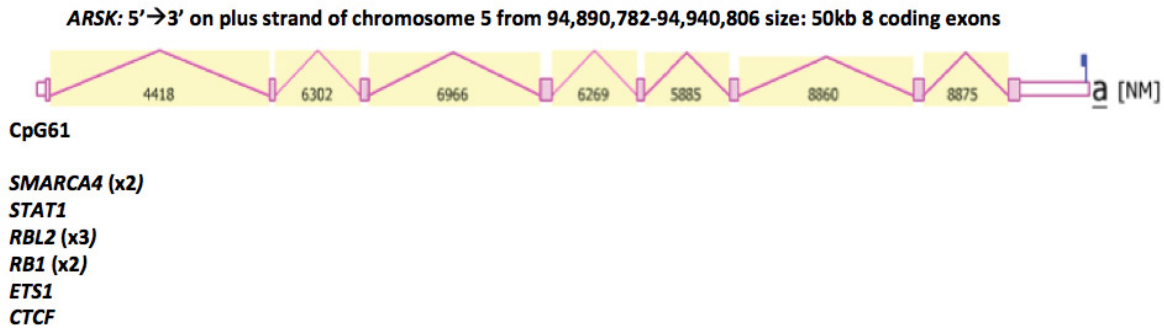

Figure 4: Human *ARSK* gene: structure and major transcript. Derived from the AceView website http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/ [11]; shown with capped 5'- and 3'- ends for the predicted mRNA sequences; NM refers to the NCBI reference sequence; coding exons are in pink; the direction for transcription is shown as 5' → 3'; a large CpG61 island at the gene promoter is shown; predicted Transcription Factor Binding Sites (TFBS) for the human *ARSK* gene promoter.

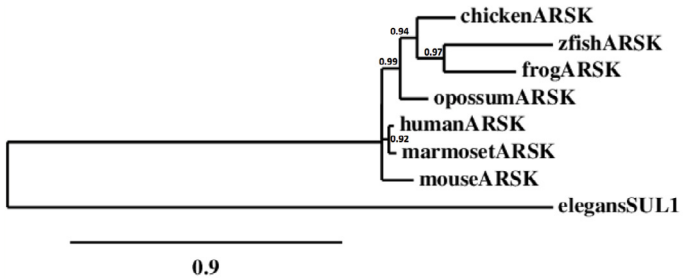

Figure 5: Vertebrate ARSK phylogenetic tree with the *C. elegans* SUL1 sequence. The tree is labeled with the ARSK and sulfatase-like (SUL1) name and the name of the organism with the worm (*Caenorhabditis elegans)* SUL1 sequence, which was used to 'root' the tree. Note the major cluster corresponding to the vertebrate *ARSK* gene family. 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. Replicate values of 0.9 or more, which are highly significant, are shown with 100 bootstrap replicates performed in each case.

References

1. Sardiello M, Annunziata I, Roma G, Ballabio A (2005) Sulfatases and sulfatase modifying factors: an exclusive and promiscuous relationship. Hum Mol Genet 14: 3203-3217.
2. Holmes RS (2016) Comparative and evolutionary studies of vertebrate arylsulfatase B, arylsulfatase I and arylsulfatase J genes and proteins. J Proteomics Bioinform 9: 11.
3. Holmes RS (2017) Comparative and evolutionary studies of mammalian arylsulfatase and sterylsulfatase genes and proteins encoded on the X-chromosome. Comp Biol Chem 68: 71-77.
4. Holmes RS (2017) Comparative studies of vertebrate iduronate 2 sulfatase (*IDS*) genes and proteins: evolution of a mammalian X-linked gene. 3 Biotech 7: 22.
5. Holmes RS (2017) Comparative and evolutionary studies of vertebrate extracellular sulfatase genes and proteins: SULF1 and SULF2. J Proteomics Bioinform 10: 2.
6. Obaya AJ (2006) Molecular cloning and initial characterization of three novel human sulfatases. Gene 372: 110-177.
7. Wiegmann EM, Westendorf E, Kalus I, Pringle TH, Lübke T, et al. (2013) Arylsulfatase K, a novel lysosomal sulfatase. J Biol Chem 288: 30019-3028.
8. Dhamale OP, Lawrence R, Wiegmann EM, Shah BA, Al-Mafraji K, et al. (2017) Arylsulfatase K is the lysosomal 2-sulfoglucuronate sulfatase. ACS Chem Biol 12: 367-373.
9. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinform 10: 421.
10. Karolchik D, Bejerano G, Hinrichs AS, Kuhn RM, Miller W, et al. (2009) Comparative genomic analysis using the UCSC genome browser. Methods Mol Biol 395: 17-34.
11. Thierry-Mieg D, Thierry-Mieg J (2006) AceView: A comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 7: 1-14.
12. Schwede T, Kopp J, Guex N, Pietsch MC (2003) SWISS-MODEL: An automated protein homology-modelling server. Nucleic Acids Res 31: 3381-3385.
13. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acid Res 39: D225-D229.
14. GTEx Consortium (2015) Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660.
15. Sievers F, Higgins DG (2014) Clustal omega. Curr Protoc Bioinformatics 48: 1-16.
16. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, et al. (2008) Phylogeny. fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36: W465- W469.
17. Bond CS, Clements PR, Ashby SJ, Collyer CA, Harrop SJ, et al. (1997) Structure of a human lysosomal sulfatase. Structure 5: 277-289.
18. Chruszcz M, Laidler P, Monkiewicz M, Ortlund E, Lebioda L, et al. (2003) Crystal structure of a covalent intermediate of endogenous human arylsulfatase A. J Inorg Biochem 96: 386-392.
19. Hernandez-Guzman FG, Higashiyama T, Pangborn W, Osawa Y, Ghosh D, et al. (2003) Structure of human estrone sulfatase suggests functional roles of membrane association. J Biol Chem 278: 22989-22997.
20. Rivera-Colón Y, Schutsky EK, Kita AZ, Garman SC (2012) The structure of human GALNS reveals the molecular basis for mucopolysaccharidosis IV A. J Mol Biol 423: 736-751.
21. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 994-1006.
22. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA 103: 1412-1417.
23. Leung JY, Nevins JR (2012) E2F6 associates with BRG1 in transcriptional regulation. PLoS One 7: e47967.
24. Kulkarni A, Scully TJ, O'Donnell LA (2016) The antiviral cytokine interferon-gamma restricts neural stem/progenitor cell proliferation through activation of STAT1 and modulation of retinoblastoma protein phosphorylation. J Neurosci Res 95: 1582-1601.
25. Zook EC, Ramirez K, Guo X, van der Voort G, Sigvardsson M, et al. (2016) The ETS1 transcription factor is required for the development and cytokine-induced expansion of ILC2. J Exp Med 213: 687-696.
26. Sams DS, Nardone S, Getselter D, Raz D, Tal M, et al. (2016) Neuronal CTCF is necessary for basal and experience-dependent gene regulation, memory formation and genomic structure of BDNF and Arc. Cell Rep 17: 2418-2430.