Holmes, J Proteomics Bioinform 2016, 9:11
DOI: 10.4172/jpb.1000418

Journal of
Proteomics & Bioinformatics

Research Article OMICS International

Comparative and Evolutionary Studies of Vertebrate Arylsulfatase B, Arylsulfatase I and Arylsulfatase J Genes and Proteins: Evidence for an ARSB-like Sub-family

Roger S Holmes*

The Eskitis Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan 4111 QLD, Australia

Abstract

Multiple sulfatase genes have been reported on the human genome, including Arylsulfatase B (ARSB), Arylsulfatase I (ARSI) and Arylsulfatase J (ARSJ). ARSB is localized in lysosomes and catalyses the hydrolysis of chondroitin and dermatan sulfate groups. Bioinformatic analyses of vertebrate genomes were undertaken using known human ARSB, ARSI and ARSJ amino acid sequences to study the relatedness and evolution of these genes and proteins. Several domain regions and key residues were conserved including signal peptides, active site residues, metal (Ca^{2+}) and substrate binding sequences, disulfide linkages and N-glycosylation sites. The genes were widely expressed in human tissues with highest levels in esophagus (ARSB), lung (ARSI) and fibroblast cells (ARSB). Human ARSB was larger in size (>200 kb) and contained 8 coding exons, whereas ARSI and ARSJ contained only 2 coding exons among all vertebrate genomes examined. CpG islands were located within the 5' region of the human ARSB, ARSI and ARSJ genes. In addition, six and seven miR-binding sites were observed within the 3'-UTR of human ARSB and ARSJ genes, respectively. Phylogenetic analyses describe a proposal for a primordial invertebrate SUL-3 gene serving as an ancestor for unequal cross over events generating these three genes in vertebrate genomes.

Keywords: Arylsulfatase B; Arylsulfatase I; Arylsulfatase J; ARSB; ARSI; ARSJ; Vertebrate; Evolution; Phylogeny; Primordial gene; Signal peptide; Transmembranes; Ca^{2+} binding; Active site; N-glycosylation site; Gene duplication

Abbreviations: ARS: Arylsulfatase; STS: Sterylsulfatase; ARSD: Arylsulfatase D; ARSE: Arylsulfatase E, ARSF: Arylsulfatase F; ARSH: Arylsulfatase H; UCSC: University of Santa Cruz California; 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

Arylsulfatase B (ARSB) is localized in mammalian lysosomes and shown to hydrolyze sulfate groups of N-acetyl-D-galactosamine-4-sulfate, chondroitin sulfate and dermatan sulfate [1-2]. Mammalian ARSB has a distinct but related amino acid sequence to other mammalian sulfatases, including Arylsulfatase A (ARSA) [3]; Arylsulfatase G (ARSG) [4]; Arylsulfatase K (ARSK) [5]; Sterylsulfatase (STS) and other members of a closely related group of arylsulfatases encoded on the mammalian X-chromosome (ARSD, ARSE, ARSF and ARSH) [6,7]. Other human sulfatases have been reported with related sequences, including Arylsulfatase I (ARSI) [8], Arylsulfatase J (ARSJ) [5], N-acetylgalactosamine-6-sulfatase (GALNS) [9]; N-acetylglucosamine-6-sulfatase (GNS) [10]; Iduronate-2-sulfatase (IDS) [11]; Heparan N-sulfatase (SGSH) [12]; and extracellular sulfatases (SULF1; SULF2) [13]. Sulfatase Modifying Factor 1 (SUMF1) plays an essential post-translational role by modifying the active site cysteine residue which is required for all of these sulfatases [5].

The structure for the Arylsulfatase B gene (ARSB) has been determined [14] and a lysosomal storage disease (Mucopolysaccharidosis VI, MPS6 or Maroteaux-Lamy syndrome) described with autosomal recessive inheritance associated with ARSB genetic variants [15,16]. Clinical features for MPS6 may include skeletal malformations, corneal clouding, stiff joints, short stature and cardiac abnormalities [17]. In addition, clinical variation of ARSB gene expression regulates colonic epithelial cell migration and cell adhesion [18], consistent with the extra-lysosomal localization of ARSB within nuclear and cell membranes [19]. Moreover, ARSB has been shown to regulate neurite outgrowth and neuronal plasticity in the central nervous system, by way of controlling sulfate glycosaminoglycans and neurocan levels [20]. Defiency of ARSB has been implicated in the restriction of aerobic metabolism during malignancy, given that molecular oxygen is required for the post-translational modification of ARSB by SUMF1 [21]. The 3D structure for human ARSB has been determined showing sequence similarity with other sulfatases, with a common domain like structure supporting an active site involved in stabilizing calcium ion and sulfate substrate binding for catalytic sulfate ester hydrolysis [22].

This study describes the predicted sequences, structures and phylogeny of vertebrate ARSB, ARSI and ARSJ genes and enzymes and compares these results with those previously reported for human and mouse ARSB genes and proteins [1,2]. Evidence is presented on the sequences and properties of ARSB, ARSI and ARSJ from several vertebrate species and for distinct exonic structures and modes of gene regulation and expression, with the identification of CpG Island, miR-

***Corresponding author:** Roger S Holmes, The Eskitis Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan 4111 QLD, Australia, Tel: 61737355077; E-mail: r.holmes@griffith.edu.au

Received October 25, 2016; **Accepted** November 18, 2016; **Published** November 23, 2016

Citation: Holmes RS (2016) Comparative and Evolutionary Studies of Vertebrate Arylsulfatase B, Arylsulfatase I and Arylsulfatase J Genes and Proteins: Evidence for an ARSB-like Sub-family. J Proteomics Bioinform 9: 298-305. doi: 10.4172/jpb.1000418

binding sites and transcription factor binding sites for these genes. Phylogenetic analyses also describe the relationships and potential origins of the ARSB, ARSI and ARSJ genes and enzymes during vertebrate evolution and a proposal for generating these genes from an ancestral invertebrate SUL-3 gene.

Materials and Methods

ARSB, ARSI and ARSJ gene and enzyme identification

Vertebrate ARSB, ARSI and ARSJ amino acid sequences were retrieved from databases (NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and ExPASy (http://www.expasy.org) [23]), using the corresponding human sequences to seed searches [1,5]. An invertebrate ARSB-like (SUL3) sequence was similarly obtained from a search of a worm (*Caenorhabditis elegans*) genome database (NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Identification of these genes and proteins was based on high predictive scores (>850) and sequence coverage (>98%) for ARSB, ARSI and ARSJ proteome sequences in each case (Table 1). BLAT searches were performed using protein sequences to confirm the gene and enzyme sequences among the species examined using the UCSC Genome Browser [24]. Gene locations, predicted gene structures and protein subunit sequences were obtained for each gene and enzyme examined showing identity with the respective sequences (Table 1). Human ARSB, ARSI and ARSJ gene structures were obtained using the AceView web browser (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/) [25]. Identification of potential gene regulatory sites, including transcription factor binding sites (TFBS), CpG islands and miRNA-binding sites within the respective gene regions, was undertaken using the UCSC Human Genome Browser [24].

Comparative human ARSB, ARSI and ARSJ gene expression

RNA-seq gene expression profiles across 53 selected tissues (or tissue segments) were examined from the public database for human ARSB, ARSI and ARSJ, based on expression levels for 175 individuals [26] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (http://www.gtex.org).

Predicted structures and properties of human ARSB, ARSI and ARSJ subunits

Predicted secondary and tertiary structures for human sequences were obtained using SWISS MODEL web tools [27]. The human ARSB tertiary structure (PDB:1FSU) [22] served as a reference for obtaining these structures, with modelled residue ranges of 42-533 for human ARSB; 44-524 for human ARSI; and 73-555 for human ARSJ. Predicted transmembrane structures for vertebrate ARSJ subunits were obtained using a web server (http://www.cbs.dtu.dk/services/TMHMM-2.0) provided by the Center for Biological Sequence Analysis of the Technical University of Denmark [28]. SignalP 3.0 web tools were used to predict the presence and location of signal peptide cleavage sites (http://www.cbs.dtu.dk/services/SignalP); and the NetNGlyc 1.0 server was used to predict potential N-glycosylation sites for vertebrate ARSB, ARSI and ARSJ subunits [29] (http://www.cbs.dtu.dk/services/NetNGlyc).

Amino acid sequence alignments and phylogenetic analyses

Alignments of human ARSB, ARSI, ARSJ, GNS, SULF1, IDS, ARSK, SGSH, ARSA, ARSG, STS, GALNS and *C. elegans* SUL3 sequences were undertaken using Clustal Omega, a multiple sequence alignment program [30] (Table 1 and Supplementary Table 1). Percentage identities were derived from the results of these alignments (Table 2). 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 [31]. Sequences were identified as vertebrate ARSB, ARSI and ARSJ members, as well as a proposed primordial *C. elegans* SUL3 gene and protein (Table 1).

Results and Discussion

Percentage identities of human arylsulfatase amino acid sequences

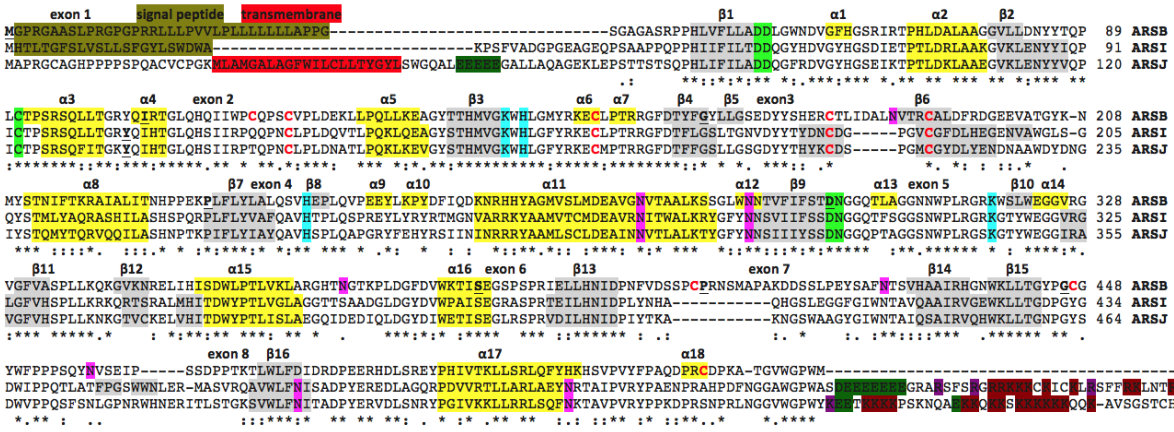
Percentages of amino acid sequence identities for 12 human ARS enzyme subunits are presented in Table 2. The sulfatase genes examined are separately localized on the human genome, encoding enzyme subunits with distinct MWs, pI values and amino acid sequence lengths (Table 3). The human ARSB, ARSI and ARSK genes were located on human chromosome 5; the human SGSH and ARSG genes on human chromosome 17; and others on separate chromosomes, in each case. This is in contrast to multiple human STS-like genes, which are located in a tandem fashion on the X-chromosome: ARSD-ARSE-ARSH-ARSF, within a 200 kb gene cluster, encoding enzymes with ≥50% sequence identities (data not shown). Of particular interest to this study were the higher levels of sequence identities observed for the human ARSB, ARSI and ARSJ enzyme subunits, which showed ≥54% sequence identities, suggesting that these genes and proteins are members of a closely related ARSB-like sub-family of human sulfatases.

Alignments of human ARSB, ARSI and ARSJ amino acid sequences

Amino acid sequence alignments for human ARSB, ARSI and ARSJ sequences (Table 1) are shown in Figure 1. Comparisons of these sequences with the human ARSB sequence, for which the tertiary structure has been described (template pdb: 1FSU) [22], enabled prediction of secondary structures and likely key residues contributing to catalysis, structure and function for the ARSI and ARSJ proteins. Active site residues (human ARSB numbers used) binding calcium ions (Ca^{2+}) [53Asp, 54Asp, 300Asp, 301Asn) or substrate (91Cys; 145Lys; 147His; 242His; 318Lys) were conserved. One of the conserved active site residues (75Cys) has been shown to undergo post-translational modification by sulfatase modifying factor 1 (SUMF1) to form C(alpha)-formylglycine (Fgly), which is required at the active site for all of these sulfatases [5]. Genetic deficiency of SUMF1 results in multiple sulfatase deficiency (MSD) [32].

Signal peptides of varying lengths were predicted for the vertebrate ARSB, ARSI and ARSJ sequences, which were consistent with the reported N-linked glycosylation and membrane associations for ARSB within lysosomal membranes (Table 1) [1]. In contrast, mammalian ARSJ sequences did not contain a predicted signal peptide, although a transmembrane structure was observed for the extended N-terminal sequence (residues 24-44 for human ARSJ). Human ARSB contained 6 predicted N-glycosylation sites (Asn188, Asn279, Asn291, Asn366, Asn426 and Asn458) for which Asn279 and Asn291 were also shared with the human ARSI and ARSJ sequences (Figure 1). In contrast, human ARSI and ARSJ sequences exhibited two other predicted N-glycosylation sites (human ARSI sequence numbers used): Asn466 and Asn496 (Figure 1). Four Cys residues involved in disulfide bond formation for human ARSB [20] were also conserved in the ARSI and ARSJ sequences (human ARSB numbers used): Cys121←→Cys155; Cys181←→Cys192; whereas four other Cys residues within the ARSB structure were not conserved for the human ARSI and ARSJ sequences:

Citation: Holmes RS (2016) Comparative and Evolutionary Studies of Vertebrate Arylsulfatase B, Arylsulfatase I and Arylsulfatase J Genes and Proteins: Evidence for an ARSB-like Sub-family. J Proteomics Bioinform 9: 298-305. doi: 10.4172/jpb.1000418



See Table 1 for details of human ARSB, ARSI and ARSJ proteins; "*" shows identical residues for sequences; ":" similar alternate residues; "." less similar alternate residues; bold font shows known or predicted exon junctions; exon numbers refer to the human ARSB gene; predicted α-helices and β-sheets are numbered as described for human ARSB [20]; active site residues are shown consistent with their roles in Ca2+ and substrate binding; predicted transmembrane, signal peptide, poly-acidic , poly-basic and N-glycosylation Asn (N) amino acids are shown; Cys residues identified as –S-S- bonds for human ARSB [22] are also shown as C

Figure 1: Amino acid sequence alignments for human ARSB, ARSI and ARSJ sequences.

Organism	Species	Gene	Coding Exons (strand)	Gene Size bps	Gene Location	GenBank ID*	UNIPROT ID	Amino acids	Subunit MW (pI)	Signal Peptide	TM
Human	Homo sapiens	ARSB	8 (-ve)	204,849	5:780,762,223-782,810,071	NM_000046	P14518	533	59,687 (8.4)	1…36	na
Mouse	Mus musculus	Arsb	8 (+ve)	168,951	13:93,771,778-93,940,728	NM_009712	P50429	534	59,647 (6.8)	1…41	na
Opossum	Monodelphis domestica	ARSB	8 (+ve)	241,222	3:37,109,032-37,350,253	XP_001381590*	F6X4S3	522	58,845 (6.3)	1…28	na
Chicken	Gallus gallus	ARSB	8 (+ve)	65,767	Z:22,361,300-22,427,066	XP_003642960*	F1P099	528	58,446 (7.0)	1…35	na
Frog	Xenopus tropicalis	ARSB	8 (+ve)	61,650	^KB021649:16,113,025-16,174,674	XP_002940244*	F7DJE4	531	58,373 (5.9)	1…35	na
Zebrafish	Danio rerio	ARSB	8 (+ve)	149,457	21:130,641-280,097	XP_003200848*	na	517	57,495 (6.1)	1…20	na
Human	Homo sapiens	ARSI	2 (-ve)	5,157	5:150,297,217-150,302,373	NM_001012301	Q5FYB1	569	64,030 (8.8)	1…23	na
Mouse	Mus musculus	Arsi	2 (+ve)	5,526	18:60,912,240-60,917,765	NM_001038499	Q32KI9	573	64,367 (8.5)	1…23	na
Opossum	Monodelphis domestica	ARSI	2 (-ve)	7,998	1:343,247,971-343,255,968	XP_001378869*	F6RZZ5	584	65,872 (8.0)	1…21	na
Chicken	Gallus gallus	ARSI	2 (+ve)	3,008	13:12,503,423-12,506,430	XP_004945003*	F1NQP9	574	65,069 (9.0)	1…18	na
Frog	Xenopus tropicalis	ARSI	2 (-ve)	13,694	^KB021651:45,092,627-45,106,320	XP_002940132*	F6ZR01	573	64,490 (8.9)	1…18	na
Zebrafish	Danio rerio	ARSI	2 (-ve)	3,756	21:44,004536-44,008,291	XP_692237*	E7F908	568	64,875 (9.4)	1…16	na
Human	Homo sapiens	ARSJ	2 (-ve)	76,558	4:113,902,277-113,978,834	NM_024590	Q5FYB0	599	67,235 (9.2)	na	24..44
Mouse	Mus musculus	Arsj	2 (+ve)	74,627	3:126,364,774-126,439,400	NM_173451	Q8BM89	598	67,354 (9.3)	na	24..44
Opossum	Monodelphis domestica	ARSJ	2 (-ve)	133,705	5:67,618,341-67,752,045	XP_001365999*	F7DM73	607	68,543 (9.3)	na	24..44
Chicken	Gallus gallus	ARSJ	2 (+ve)	39,513	4:56,052,213-56,091,725	XP_420639*	F1NH07	578	64,975 (9.2)	1…19	na
Frog	Xenopus tropicalis	ARSJ	2 (-ve)	30,946	^KB021649:153,510,397-153,541,342	XP_002934299*	F6XVC7	564	63,962 (9.0)	1…17	na
Zebrafish	Danio rerio	ARSJ	2 (+ve)	11,114	7:57,375,146-57,386,259	XP_688265*	F1REG3	568	64,733 (9.2)	1…32	na
Worm	C. elegans	SUL-3	14 (-ve)	6,839	X:7,827,197-7,834,035	NM_001047767	H2KZF6	484	55,541 (9.1)	1…24	na

Na: Not Available; *: Predicted ID from NCBI; ^: Scaffold ID; TM: Predicted Transmembrane Sequence

Table 1: Human ARSB, ARSI and ARSJ genes and enzymes.

	GNS	SULF1	IDS	ARSK	SGSH	ARSB	ARSI	ARSJ	ARSA	ARSG	STS	GALNS
GNS	100	38	22	22	22	20	22	22	22	19	22	22
SULF1	38	100	22	20	19	18	21	20	20	17	20	19
IDS	22	22	100	24	24	27	24	24	27	20	25	25
ARSK	22	20	24	100	23	21	20	21	24	20	24	21
SGSH	22	19	24	23	100	27	25	23	28	27	26	27
ARSB	20	18	27	21	27	**100**	**55**	**54**	30	26	31	30
ARSI	22	21	24	20	25	**55**	**100**	**55**	28	26	27	27
ARSJ	22	20	24	21	23	**54**	**59**	**100**	25	25	25	26
ARSA	22	20	27	24	28	30	28	25	100	39	36	35
ARSG	19	17	20	20	27	26	26	25	39	100	32	35
STS	22	20	25	24	26	31	27	26	36	32	100	34
GALNS	22	19	25	21	27	30	27	26	35	35	34	100

Higher percentages for human ARSB, ARSI and ARSJ sequences are highlighted as Bold

Table 2: Percentage sequence identities for human ARS-like proteins.

Human Gene	Coding Exons (strand)	Gene Size bps	Gene Location	GenBank ID*	UNIPROT ID	Amino acids	Subunit MW (pI)
ARSB	8 (-ve)	2,04,849	5:780,762,223-782,810,071	NM_000046	P14518	533	59,687 (8.4)
ARSI	2 (-ve)	5,157	5:150,297,217-150,302,373	NM_001012301	Q5FYB1	569	64,030 (8.8)
ARSJ	2 (-ve)	76,558	4:113,902,277-113,978,834	NM_024590	Q5FYB0	599	67,235 (9.2)
GNS	14 (-ve)	42,533	12:64,716,744-64,759,276	NM_002076	P15586	552	62,082 (8.6)
SULF1	18 (+ve)	74,885	8:69,563,976-69,638,860	NM_001128204	Q8IWU6	871	101,027 (9.2)
IDS	9 (-ve)	22,389	X:149,482,749-149,505,137	NM_000207	P22304	550	61,873 (5.2)
ARSK	8 (+ve)	48,245	5:95,555,279-95,603,523	NM_198150	Q6UWY0	536	61,450 (9.0)
SGSH	8 (-ve)	9,859	17:80,210,455-80,220,313	NM_000199	P51688	502	56,695 (6.5)
ARSA	8 (-ve)	2,626	22:50,625,148-50,627,773	NM_000487	P15289	507	53,588 (5.7)
ARSG	11 (+ve)	1,12,967	17:68,307,494-68,420,460	NM_001267727	Q96EG1	525	57,061 (6.2)
STS	10 (+ve)	97,065	X:7,253,194-7,350,258	NM_000351	P08842	583	65,492 (7.6)
GALNS	14 (-ve)	42,536	16:88,880,850-88,923,285	NM_000512	P34059	522	58,026 (6.3)

*NCBI sequence; pI-isoelectric point

Table 3: Human ARS-like genes and proteins.

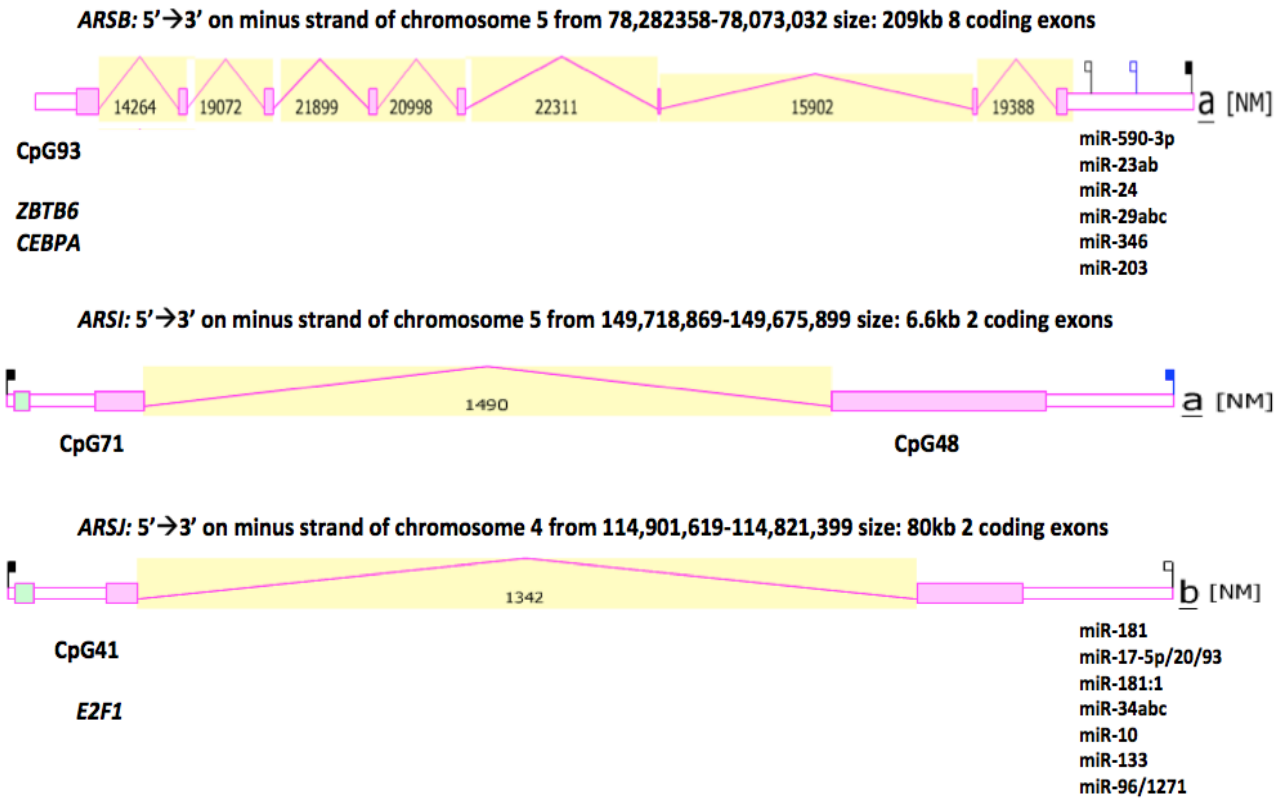
Cys117←→Cys521; and Cys405←→Cys447 (Figure 1). A poly-Glu (acidic) region within the human ARSJ N-terminal sequence was observed (residues 51Glu-55Glu) (Figure 1), which was shared with other mammalian ARSJ sequences (results not shown). A poly-acidic amino acid sequence (**Asp526-Glu527-Glu528-Glu529-Glu530-Glu531-Glu532-Glu533**) was also found at the C-terminus end of the human ARSI sequence. In contrast, a region of poly-basic amino acid residues was observed for the human ARSJ C-terminus sequence: **Lys571-Lys572**-Gln573-**Lys574-Lys575**-Ser576-**Lys575-Lys578-Lys579-Lys580-Lys581-Lys582**-Gln583-Gln584-**Lys585**- (Figure 1). The roles for these regions of negative and positive charges for the C-termini of the ARSI and ARSJ sequences remain to be determined, however it is likely that specific protein-protein binding may be assisted by these regions, in a similar way to that previously reported for glycosylphosphatidylinositol-anchored high density lipoprotein-binding protein 1, which binds lipoprotein lipase via a poly-acidic amino acid sequence [33]. Alternatively, these C-terminal negative and positive regions may contribute to specific microlocalization properties for these enzymes or to ARSI-ARSJ associations in forming dimeric hybrid enzymes, *in vivo*.

Predicted secondary structures for human ARSB, ARSI and ARSJ subunits are compared in Figure 1. Similar structures were observed for each of the enzymes examined, with the exception of the number of the signal peptide and transmembrane structures observed, previously discussed. Supplementary Figure 1 provides a 3-dimensional model for human ARSB which is based on the structure previously described by Bond and coworkers [22]. This shows a cleft, with the active site region located within an enzyme cavity containing a metal (Ca^{2+}) ion on the ARSB surface at the carboxyl end of the central parallel portion of the β sheet. The enzyme has 2 domains, with the active site at the base of a cleft on the larger domain. Predicted 3D structures for human ARSI and ARSJ sequences were also undertaken (results not shown) which showed similar results for each of these enzymes.

Predicted gene locations and exonic structures for ARSB, ARSI and ARSJ genes and proteins

Table 1; Figures 1 and 2 summarize the predicted locations, sizes and number of coding exons for vertebrate ARSB, ARSI and ARSJ genes examined, and of the encoded human ARSB, ARSI and ARSJ subunit amino acid sequences. These were based on BLAST interrogations of vertebrate gene databases (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using the reported sequences for human ARSB [1], ARSI [8] and ARSJ [5], and BLAT analyses of vertebrate genomes using the UC Santa Cruz Genome Browser (http://genome.ucsc.edu/cgi-bin/hgBlat) [24]. Vertebrate ARSB genes contained 8 coding exons and were larger in size (61-241 kb), whereas vertebrate ARSI and ARSJ genes contained only 2 coding exons and were smaller in size (3-14 kb and 11-133 kb, respectively), in each case (Table 1).

Table 1 show comparative locations, gene sizes and coding exon compositions for vertebrate ARSB, ARSI and ARSJ genes and a worm (*C. elegans*) ARSB-like gene (SUL-3), as well as comparative protein structures for the enzyme subunits. As can be seen, the three ARSB-like genes are widely separately on chromosomes for all species examined which may reflect the antiquity of these genes among the vertebrate



Derived from the AceView website http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/ [25]; 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'; nucleotide numbers for introns are shown; CpG93, CpG71 and CpG41 islands were identified within the gene promoters for the human ARSB, ARSI and ARSJ, respectively; TFBS were also located within the gene promoters for ARSB (ZBTB6; CEBPA) and ARSJ (E2F1); predicted miRNA target site were identified within the 3'-UTR regions of the human ARSB and ARSJ genes

Figure 2: Gene structures and major gene transcripts for the human ARSB, ARSI and ARSJ genes.

genomes examined. Figure 1 summarizes the predicted exonic start sites for human ARSB, ARSI and ARSJ genes with ARSB having 8 exons, whereas ARSI and ARSJ contained only 2 coding exons. Of particular interest to this comparison was the similar positioning for the exon 2 start site for each of these genes, which may reflect a common evolutionary origin for this site within these genes.

Figure 2 presents diagrams for the major isoforms for human ARSB, ARSI and ARSJ genes, showing comparative locations and sizes for introns and exons, and for 5'- and 3'-UTR regions. As can be seen, the human ARSB gene is >30 times larger than the human ARSI gene and 2.6 times larger than the ARSJ gene, predominantly due to fewer exons being present for the latter genes (2 exons compared with 8 exons for ARSB). The human ARSB gene promoter contained a CpG island (CpG93) [34] and two predicted TFBS: ZBTB6, which encodes a Zinc finger and BTB domain-containing protein 6 which mediates transcriptional repression [35]; and CEBPA or CCAAT/enhancer-binding protein alpha, a transcription factor that coordinates differentiation of hepatocytes, adipocytes, myeloid progenitors and cells of the placenta and lung [36]. Six microRNA sites were also located in the 3'-UTR of human ARSB, which are potentially of major significance for the regulation of this gene (Figure 2). A recent study of miR-590 has shown that it regulates osteogenic differentiation in developing human mesenchymal cells [37]. In addition, miR-24 functions as a tumor suppressor in nasopharangeal carcinoma [38]; miR-29 promotes Type II cell differentiation in the developing lung [39]; miR-346 regulates osteogenic differentiation of human bone marrow-derived mesenchymal stem cells [40]; and miR-203 suppresses cell proliferation, migration and invasion in colorectal cancer [41].

The ARSI gene promoter contained a CpG71 island although no predicted TFBS were detected in this region, and no miR-binding sites were observed in the ARSI 3'-UTR. The ARSJ gene promoter contained a CpG41 island and a predicted TFBS (E2F1), which represses transcriptional activity and may block adipocyte differentiation [41]. Seven miRNA binding regions were predicted in the 3'-UTR of the human ARSJ gene: miR-181 which functions as a tumor suppressor in non-small cell lung cancer [42]; miR-17-5p, which is strongly expressed in embryonic stem cells and has essential roles in cell cycle regulation, proliferation and apoptosis [43]; miR-181:1, which may act as a tumor suppressor in the pathogenesis of acute myeloid leukemia [44]; miR-34a, which inhibits breast cancer proliferation [45]; miR-10, which participates in the regulation of Hox gene developmental regulators [46]; miR-133, recognized as a biomarker for lung cancer [47]; and miR-96, which promotes the growth of prostate carcinoma cells [48].

Comparative ARSB, ARSI and ARSJ human tissue expression

Figure 3 shows comparative gene expression for various human tissues obtained from RNA-seq gene expression profiles for human ARSB, ARSI and ARSJ genes obtained for 53 selected tissues or tissue segments for 175 individuals [26] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (http://www.gtex.org). These data supported a wide tissue expression profile for the 3 genes,

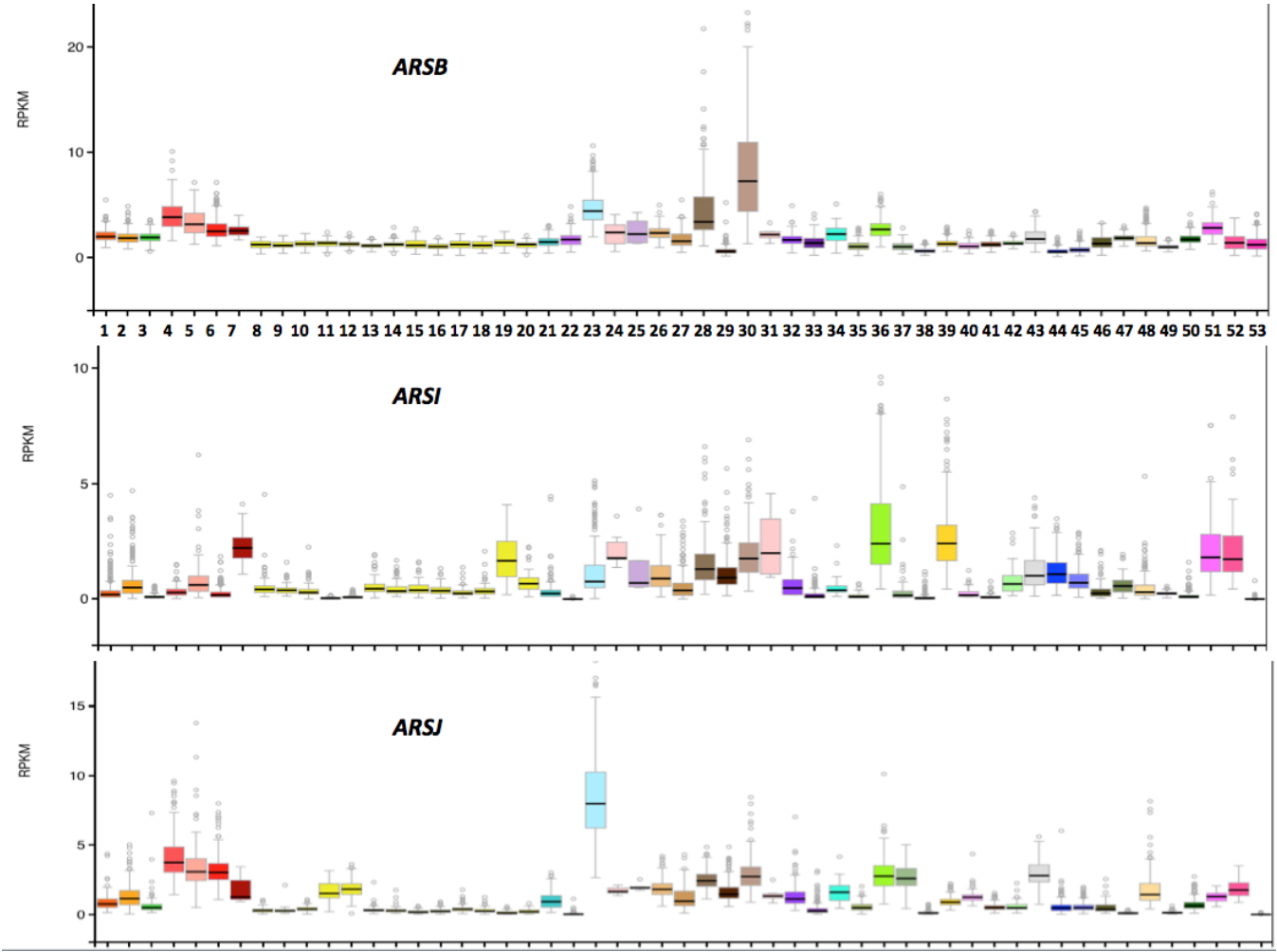
with highest levels for human ARSB observed in the esophagus and transformed fibroblasts; the highest ARSI gene expression level was observed in lung and the tibial nerve; whereas highest ARSJ expression was seen in transformed fibroblasts.

Phylogeny and evolution of vertebrate ARSB, ARSI and ARSJ sequences

A phylogram (Figure 4) was calculated by the progressive alignment of vertebrate ARSB, ARSI, and ARSJ amino acid sequences, using a worm SUL-3 sequence (from *C. elegans*) (Table 1) to 'root' the tree. Homolog sequences were identified for all vertebrate genomes examined. The phylogram demonstrates separation of these sequences into three distinct groups consistent with their relatedness during vertebrate evolution, and suggests that these genes have been derived from an ancestral invertebrate SUL-3 gene.

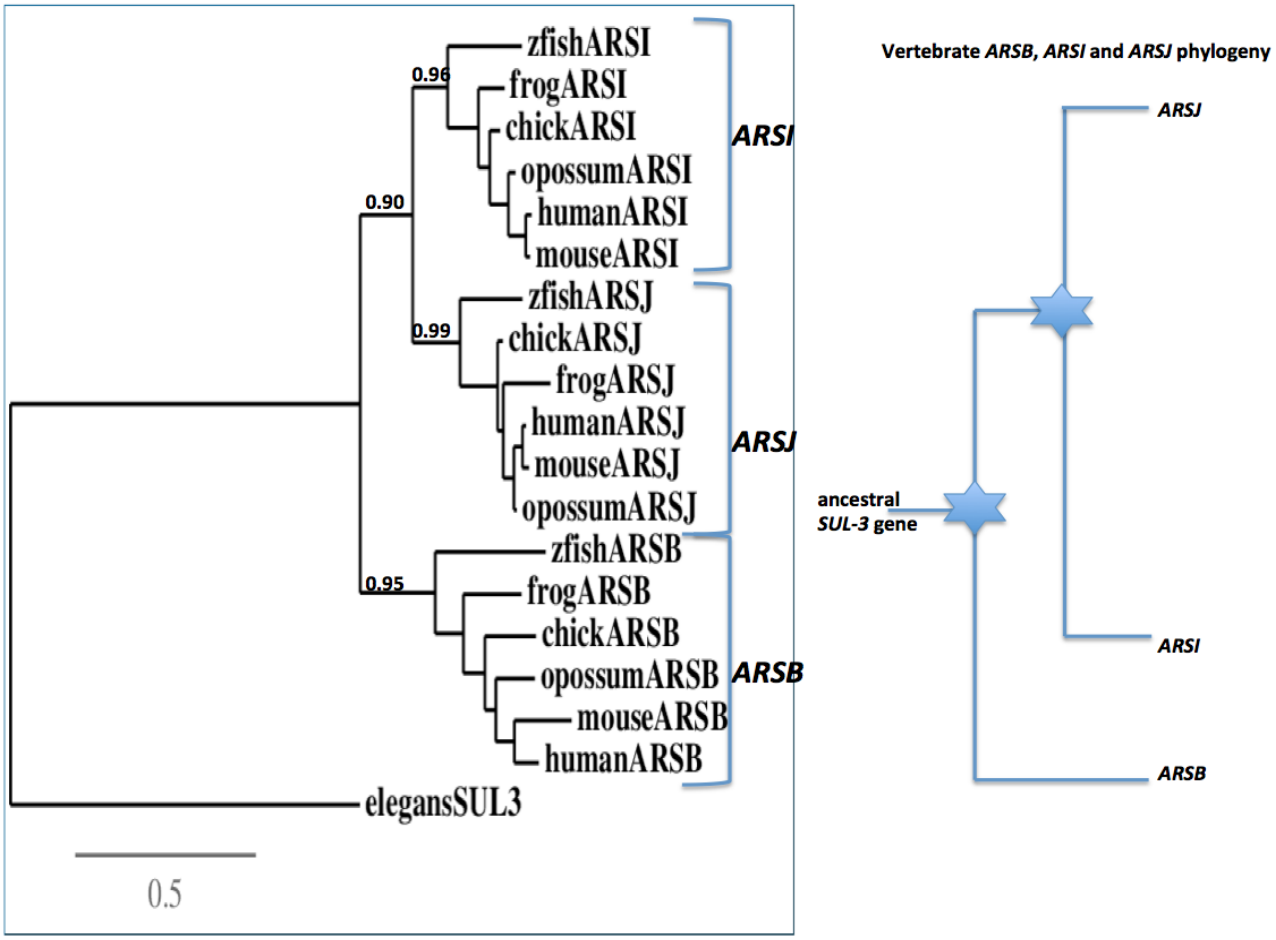
Figure 4 also summarizes a working hypothesis for the evolution of vertebrate ARSB-like gene genes:

1. A proposed primordial invertebrate ARSB-like gene (SUL-3) was derived from a bacterial ancestor.
2. A proposed vertebrate ARSB ancestral gene containing 8 coding exons was inherited from a primordial vertebrate ancestor.
3. Following cDNA formation, a 2 exon transcript was retrointegrated into an ancestral vertebrate genome, forming an ancestral ARSI-ARSJ primordial gene.
4. A gene duplication event generated 2 separate lines of evolution: ARSI and ARSJ genes, which underwent sequence divergence and separate integration into the vertebrate genome.



RNA-seq gene expression profiles across 53 selected tissues (or tissue segments) were examined from the public database for human ARSB, ARSI and ARSJ, based on expression levels for 175 individuals [26] (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. Brain-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-Gastroesophageal 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

Figure 3: Comparative tissue expression for human ARSB, ARSI and ARSJ genes.



The tree is labelled with the gene name and the name of the mammal or invertebrate. The tree is 'rooted' with the C. elegans SUL3 sequence. See Table 1 for details of ARSB, ARSI and ARSJ 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. A diagram of the proposed evolutionary and gene duplication events (represented as a star) are shown

Figure 4: Phylogenetic tree for vertebrate and invertebrate ARSB, ARSI and ARSJ sequences.

Conclusion

BLAST and BLAT analyses of veretebrate genome databases were undertaken using amino acid sequences reported for human ARSB for interrogation of vertebrate genome sequences and an invertebrate genome (Sul-3) sequence (*C. elegans*). Predicted amino acid sequences for these vertebrate ARSB, ARSI and ARSJ subunits showed a high degree of sequence identity (>54% identical). Secondary structure and key residue identification were undertaken using a previous report for human ARSB 3D structure [22], which enabled identification of putative secondary structures and likely key residues contributing to catalysis, structure and function.

Bioinformatic analyses enabled the identification of putative gene regulation sites, including CpG islands, TFBS and miR-binding sites, for the promoter and 3'-UTR regions for the human ARSB, ARSI and ARSJ genes examined. These included CpG93, CpG71 and CpG41 localized within the human ARSB, ARSI and ARSJ gene promoter region; and 13 miR-binding sites, including miR-590 (for ARSB 3'-UTR) which regulates osteogenic differentiation in developing human mesenchymal cells [37]; and miR-10, which participates in the regulation of Hox gene developmental regulators [46].

Phylogenetic analyses suggested that vertebrate ARSB, ARSI and ARSJ genes were derived from an initial gene duplication event of a primordial invertebrate Sul-3 gene, generating 2 sub-families: ARSB and ARSI/ARSJ genes, with the latter containing only 2 coding exons, in comparison with the vertebrate ancestral ARSB gene, containing 8 coding exons.

Acknowledgements

I thank Dr Laura Cox from the Texas Biomedical Research Institute for useful discussions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Peters C, Schmidt B, Rommerskirch W, Rupp K, Zühlsdorf M, et al. (1990) Phylogenetic conservation of arylsulfatases. cDNA cloning and expression of human arylsulfatase B. J Biol Chem 265: 3374-3381.
2. Schuchman EH, Jackson CE, Desnick RJ (1990) Human arylsulfatase B: MOPAC cloning, nucleotide sequence of a full-length cDNA, and regions of amino acid identity with arylsulfatases A and C. Genomics 6: 149-158.
3. Stein C, Hille A, Seidel J, Rijnbout S, Waheed A, et al. (1989) Cloning and expression of human steroid-sulfatase. Membrane topology, glycosylation, and subcellular distribution in BHK-21 cells. J Biol Chem 264: 13865-13872.

4. Ferrante P, Messali S, Meroni G, Ballabio A (2002) Molecular and biochemical characterisation of a novel sulphatase gene: Arylsulfatase G (ARSG). Eur J Hum Genet 10: 813-818.

5. 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.

6. Basler E, Grompe M, Parenti G, Yates J, Ballabio A (1992) Identification of point mutations in the steroid sulfatase gene of three patients with X-linked ichthyosis. Am J Hum Genet 50: 483-491.

7. Franco B, Meroni G, Parenti G, Levilliers J, Bernard L, et al. (1995) A cluster of sulfatase genes on Xp22.3: mutations in chondrodysplasia punctata (CDPX) and implications for warfarin embryopathy. Cell 81: 15-25.

8. Oshikawa M, Usami R, Kato S (2009) Characterization of the arylsulfatase I (ARSI) gene preferentially expressed in the human retinal pigment epithelium cell line ARPE-19. Mol Vis 15: 482-494.

9. Tomatsu S, Fukuda S, Masue M, Sukegawa K, Fukao T, et al. (1991) Morquio disease: isolation, characterization and expression of full-length cDNA for human N-acetylgalactosamine-6-sulfate sulfatase. Biochem Biophys Res Commun 181: 677-683.

10. Robertson DA, Freeman C, Morris CP, Hopwood JJ (1992) A cDNA clone for human glucosamine-6-sulphatase reveals differences between arylsulphatases and non-arylsulphatases. Biochem J 288: 539-544.

11. Bielicki J, Freeman C, Clements PR, Hopwood JJ (1990) Human liver iduronate-2-sulphatase. Purification, characterization and catalytic properties. Biochem J 271: 75-86.

12. Scott HS, Blanch L, Guo XH, Freeman C, Orsborn A, et al. (1995) Cloning of the sulphamidase gene and identification of mutations in Sanfilippo A syndrome. Nat Genet 11: 465-467.

13. Takashima Y, Keino-Masu K, Yashiro H, Hara S, Suzuki T, et al. (2016) Heparan sulfate 6-O-endosulfatases, Sulf1 and Sulf2, regulate glomerular integrity by modulating growth factor signaling. Am J Physiol Renal Physiol 310: F395-408.

14. Modaressi S, Rupp K, von Figura K, Peters C (1993) Structure of the human arylsulfatase B gene. Biol Chem Hoppe Seyler 374: 327-335.

15. Wicker G, Prill V, Brooks D, Gibson G, Hopwood J, et al. (1991) Mucopolysaccharidosis VI (Maroteaux-Lamy syndrome). An intermediate clinical phenotype caused by substitution of valine for glycine at position 137 of arylsulfatase B. J Biol Chem 266: 21386-21391.

16. Garrido E, Cormand B, Hopwood JJ, Chabás A, Grinberg D, et al. (2008) Maroteaux-Lamy syndrome: functional characterization of pathogenic mutations and polymorphisms in the arylsulfatase B gene. Mol Genet Metab 94: 305-312.

17. Azevedo AC, Schwartz IV, Kalakun L, Brustolin S, Burin MG, et al. (2004) Clinical and biochemical study of 28 patients with mucopolysaccharidosis type VI. Clin Genet 66: 208-213.

18. Bhattacharyya S, Tobacman JK (2009) Arylsulfatase B regulates colonic epithelial cell migration by effects on MMP9 expression and RhoA activation. Clin Exp Metastasis 26: 535-545.

19. Prabhu SV, Bhattacharyya S, Guzman-Hartman G, Macias V, Kajdacsy-Balla A, et al. (2011) Extra-lysosomal localization of Arylsulfatase B in human colonic epithelium. J Histochem Cytocham 59: 328-335.

20. Zhang X, Bhattacharyya S, Kusumo H, Goodlett CR, Tobacman JK, et al. (2014) Arylsulfatase B modulates neurite outgrowth via astrocyte chondroitin-4-sulfate: dysregulation by ethanol. Glia 62: 259-271.

21. Bhattacharyya S, Feferman L, Tobacman JK (2016) Restriction of Aerobic Metabolism by Acquired or Innate Arylsulfatase B Deficiency: A New Approach to the Warburg Effect. Sci Rep 6: 32885.

22. Bond CS, Clements PR, Ashby SJ, Collyer CA, Harrop SJ, et al. (1997) Structure of a human lysosomal sulfatase. Structure 5: 277-289.

23. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, et al. (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40: W597-W603.

24. Karolchik D, Bejerano G, Hinrichs AS, Kuhn RM, Miller W, et al. (2007) Comparative genomic analysis using the UCSC genome browser. Methods Mol Biol 395: 17-34.

25. Thierry-Mieg D, Thierry-Mieg J (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 7: S12.

26. GTEx Consortium (2015) Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660.

27. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31: 3381-3385.

28. Krogh A, Larsson B (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.

29. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953-971.

30. Sievers F, Higgins DG (2014) Clustal omega. Curr Protoc Bioinformatics 48: 3.

31. 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.

32. Prasad C, Rupar CA, Campbell C, Napier M, Ramsay D, et al. (2014) Case of multiple sulfatase deficiency and ocular albinism: a diagnostic odyssey. Can J Neurol Sci 41: 626-631.

33. Holmes RS, Cox LA (2012) Comparative studies of glycosylphosphatidylinositol-anchored high-density lipoprotein-binding protein 1: evidence for a eutherian mammalian origin for the GPIHBP1 gene from an LY6-like gene. 3 Biotech 2: 37-52.

34. Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. Nat Rev Genet 14: 204-220.

35. Deweindt C, Albagli O, Bernardin F, Dhordain P, Quief S, et al. (1995) The LAZ3/BCL6 oncogene encodes a sequence-specific transcriptional inhibitor: a novel function for the BTB/POZ domain as an autonomous repressing domain. Cell Growth Differ 6: 1495-1503.

36. Jules J, Chen W, Feng X, Li YP (2016) CCAAT/Enhancer-binding Protein α (C/EBPα) Is Important for Osteoclast Differentiation and Activity. J Biol Chem 291: 16390-16403.

37. Wu S, Liu W, Zhou L (2016) MiR-590-3p regulates osteogenic differentiation of human mesenchymal stem cells by regulating APC gene. Biochem Biophys Res Commun 478: 1582-1587.

38. Li YQ, Lu JH, Bao XM, Wang XF, Wu JH, et al. (2015) MiR-24 functions as a tumor suppressor in nasopharyngeal carcinoma through targeting FSCN1. J Exp Clin Cancer Res 34: 130.

39. Guo W, Benlhabib H, Mendelson CR (2016) The MicroRNA 29 Family Promotes Type II Cell Differentiation in Developing Lung. Mol Cell Biol 36: 2141.

40. Dellago H, Bobbili MR, Grillari J (2016) MicroRNA-17-5p: At the Crossroads of Cancer and Aging - A Mini-Review. Gerontology.

41. Deng B, Wang B, Fang J, Zhu X, Cao Z, et al. (2016) MiRNA-203 suppresses cell proliferation, migration and invasion in colorectal cancer via targeting of EIF5A2. Sci Rep 6: 28301.

42. Huang P, Ye B, Yang Y, Shi J, Zhao H (2015) MicroRNA-181 functions as a tumor suppressor in non-small cell lung cancer (NSCLC) by targeting Bcl-2. Tumour Biol 36: 3381-3387.

43. Zaragoza K, Bégay V, Schuetz A, Heinemann U, Leutz A (2010) Repression of transcriptional activity of C/EBPalpha by E2F-dimerization partner complexes. Mol Cell Biol 30: 2293-2304.

44. Weng H, Lal K, Yang FF, Chen J (2015) The pathological role and prognostic impact of miR-181 in acute myeloid leukemia. Cancer Genet 208: 225-229.

45. Si W, Li Y, Shao H, Hu R, Wang W, et al. (2016) MiR-34a Inhibits Breast Cancer Proliferation and Progression by Targeting Wnt1 in Wnt/β-Catenin Signaling Pathway. Am J Med Sci 352: 191-199.

46. Tehler D, Høyland-Kroghsbo NM, Lund AH (2011) The miR-10 microRNA precursor family. RNA Biol 8: 728-734.

47. Xiao B, Liu H, Gu Z, Ji C (2016) Expression of microRNA-133 inhibits epithelial-mesenchymal transition in lung cancer cells by directly targeting FOXQ1. Arch Bronconeumol 52: 505-511.

48. Xu L, Zhong J, Guo B, Zhu Q, Liang H, et al. (2016) miR-96 promotes the growth of prostate carcinoma cells by suppressing MTSS1. Tumour Biol 37: 12023-12032.