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 (ARS I) 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 (Ca²⁺) 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 (ARSH). 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; Ca²⁺ 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; BLAT: Basic Local Alignment Search 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-acetylgalactosamine-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 Moroteaux-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]. Deficiency 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

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


Copyright: © 2016 Holmes RS. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
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 mRNA-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 databases 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 (Ca²⁺) [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 ARSI). 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:
See Table 1 for details of human ARSB, ARSI and ARSJ proteins; **"** shows identical residues for sequences; : **"** shows alternate residues; ", **"** shows 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.

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

<table>
<thead>
<tr>
<th>Organism</th>
<th>Species</th>
<th>Gene</th>
<th>Coding Exons (strand)</th>
<th>Gene Size (bps)</th>
<th>Gene Location</th>
<th>GenBank ID*</th>
<th>UNIPROT ID</th>
<th>Amino acids</th>
<th>Subunit MW (pI)</th>
<th>Signal Peptide</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Homo sapiens</td>
<td>ARSB</td>
<td>8 (-ve)</td>
<td>204,849</td>
<td>5:780,762,223-782,810,071</td>
<td>NM_000046</td>
<td>P14518</td>
<td>533</td>
<td>59.687 (8.4)</td>
<td>1...36</td>
<td>na</td>
</tr>
<tr>
<td>Mouse</td>
<td>Mus musculus</td>
<td>Arsb</td>
<td>8 (+ve)</td>
<td>168,951</td>
<td>13:93,771,778-93,940,728</td>
<td>NM_009712</td>
<td>P50429</td>
<td>534</td>
<td>59.647 (6.8)</td>
<td>1...41</td>
<td>na</td>
</tr>
<tr>
<td>Opossum</td>
<td>Monodelphis</td>
<td>ARSB</td>
<td>8 (+ve)</td>
<td>241,222</td>
<td>3:37,109,032-37,350,253</td>
<td>XP_001381590*</td>
<td>F6X4S3</td>
<td>522</td>
<td>58.845 (6.3)</td>
<td>1...28</td>
<td>na</td>
</tr>
<tr>
<td>Chicken</td>
<td>Gallus gallus</td>
<td>ARSB</td>
<td>8 (+ve)</td>
<td>65,767</td>
<td>2:22,361,300-22,427,066</td>
<td>XP_003642900*</td>
<td>F1P099</td>
<td>528</td>
<td>58.446 (7.0)</td>
<td>1...35</td>
<td>na</td>
</tr>
<tr>
<td>Frog</td>
<td>Xenopus tropicalis</td>
<td>ARSB</td>
<td>8 (+ve)</td>
<td>61,650</td>
<td>^KB021649:16,113,025-16,174,674</td>
<td>XP_002940244*</td>
<td>F7DE4</td>
<td>531</td>
<td>58.373 (5.9)</td>
<td>1...35</td>
<td>na</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Danio rerio</td>
<td>ARSB</td>
<td>8 (+ve)</td>
<td>149,457</td>
<td>21:130,641-280,097</td>
<td>XP_003200848*</td>
<td>na</td>
<td>517</td>
<td>57.496 (6.1)</td>
<td>1...20</td>
<td>na</td>
</tr>
<tr>
<td>Human</td>
<td>Homo sapiens</td>
<td>ARSI</td>
<td>2 (-ve)</td>
<td>5,157</td>
<td>5:150,297,217-150,302,373</td>
<td>NM_001012301</td>
<td>Q5FYB1</td>
<td>569</td>
<td>64.030 (8.8)</td>
<td>1...23</td>
<td>na</td>
</tr>
<tr>
<td>Mouse</td>
<td>Mus musculus</td>
<td>Arsi</td>
<td>2 (+ve)</td>
<td>5,526</td>
<td>18:60,912,240-60,917,765</td>
<td>NM_001038499</td>
<td>Q32KI9</td>
<td>573</td>
<td>63.367 (8.5)</td>
<td>1...23</td>
<td>na</td>
</tr>
<tr>
<td>Opossum</td>
<td>Monodelphis</td>
<td>ARSI</td>
<td>2 (-ve)</td>
<td>7,998</td>
<td>1:343,247,971-343,255,968</td>
<td>XP_003642900*</td>
<td>F1P099</td>
<td>528</td>
<td>58.446 (7.0)</td>
<td>1...35</td>
<td>na</td>
</tr>
<tr>
<td>Chicken</td>
<td>Gallus gallus</td>
<td>ARSI</td>
<td>2 (+ve)</td>
<td>3,008</td>
<td>13:12,503,423-12,506,430</td>
<td>XP_00495003*</td>
<td>F1NQP9</td>
<td>574</td>
<td>65.069 (9.0)</td>
<td>1...18</td>
<td>na</td>
</tr>
<tr>
<td>Frog</td>
<td>Xenopus tropicalis</td>
<td>ARSI</td>
<td>2 (-ve)</td>
<td>13,694</td>
<td>^KB021649:45,092,627-45,106,320</td>
<td>XP_002940432*</td>
<td>F6ZRO1</td>
<td>573</td>
<td>64.490 (8.9)</td>
<td>1...18</td>
<td>na</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Danio rerio</td>
<td>ARSI</td>
<td>2 (+ve)</td>
<td>3,756</td>
<td>21:44,004563-44,008,291</td>
<td>XP_692237*</td>
<td>E7F908</td>
<td>568</td>
<td>64.875 (9.4)</td>
<td>1...16</td>
<td>na</td>
</tr>
<tr>
<td>Human</td>
<td>Homo sapiens</td>
<td>ARSJ</td>
<td>2 (-ve)</td>
<td>76,558</td>
<td>4:113,902,277-113,978,834</td>
<td>NM_024590</td>
<td>Q5FYB0</td>
<td>599</td>
<td>67.235 (9.2)</td>
<td>24...44</td>
<td>na</td>
</tr>
<tr>
<td>Mouse</td>
<td>Mus musculus</td>
<td>Arsj</td>
<td>2 (+ve)</td>
<td>74,627</td>
<td>3:126,364,774-126,439,400</td>
<td>NM_173451</td>
<td>Q8BM89</td>
<td>598</td>
<td>67.354 (9.3)</td>
<td>24...44</td>
<td>na</td>
</tr>
<tr>
<td>Opossum</td>
<td>Monodelphis</td>
<td>ARSJ</td>
<td>2 (-ve)</td>
<td>133,705</td>
<td>5:67,618,341-67,752,045</td>
<td>XP_00365999*</td>
<td>F7DM73</td>
<td>607</td>
<td>68.543 (9.3)</td>
<td>24...44</td>
<td>na</td>
</tr>
<tr>
<td>Chicken</td>
<td>Gallus gallus</td>
<td>ARSJ</td>
<td>2 (+ve)</td>
<td>39,513</td>
<td>4:56,052,213-56,091,725</td>
<td>XP_420639*</td>
<td>F1NH07</td>
<td>578</td>
<td>64.975 (9.2)</td>
<td>1...19</td>
<td>na</td>
</tr>
<tr>
<td>Frog</td>
<td>Xenopus tropicalis</td>
<td>ARSJ</td>
<td>2 (+ve)</td>
<td>30,946</td>
<td>^KB021649:153,510,397-153,541,342</td>
<td>XP_002940299*</td>
<td>F6XVC7</td>
<td>564</td>
<td>63.962 (9.0)</td>
<td>1...17</td>
<td>na</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Danio rerio</td>
<td>ARSJ</td>
<td>2 (+ve)</td>
<td>11,114</td>
<td>7:57,375,146-57,386,259</td>
<td>XP_688265*</td>
<td>F1REG3</td>
<td>568</td>
<td>64.733 (9.2)</td>
<td>1...32</td>
<td>na</td>
</tr>
<tr>
<td>Worm</td>
<td>C. elegans</td>
<td>SUL-3</td>
<td>14 (-ve)</td>
<td>6,839</td>
<td>X:7,827,197-7,834,035</td>
<td>NM_00147767</td>
<td>H2ZKF6</td>
<td>484</td>
<td>55.541 (9.1)</td>
<td>1...24</td>
<td>na</td>
</tr>
</tbody>
</table>

Na: Not Available; ^: Predicted ID from NCBI; ~: Scaffold ID; TM: Predicted Transmembrane Sequence
Cys117←→Cys521; and Cys405←→Cys447 (Figure 1). A poly-Glu (acidic) region within the human ARS) N-terminal sequence was observed (residues 51Glu-55Glu) (Figure 1), which was shared with other mammalian ARS) sequences (results not shown). A poly-basic amino acid sequence (Asp526-Glu527-Glu528-Glu529-Glu530-) was also found at the C-terminal end of other mammalian ARSJ sequences (results not shown). A poly-Glu (acidic) region within the human ARSJ N-terminal sequence was located within an enzyme cavity containing a metal (Ca2+) 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 ARS) [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 shows 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 comparable 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 proteome.

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

<table>
<thead>
<tr>
<th>Human Gene</th>
<th>Coding Exons (strand)</th>
<th>Gene Size (bps)</th>
<th>Gene Location</th>
<th>GenBank ID*</th>
<th>UNIPROT ID</th>
<th>Amino acids</th>
<th>Subunit MW (pI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARSB</td>
<td>8 (-ve)</td>
<td>2,04,849</td>
<td>5,780,762,223-782,810,071</td>
<td>NM_000046</td>
<td>P14518</td>
<td>533</td>
<td>59,687 (8.4)</td>
</tr>
<tr>
<td>ARSI</td>
<td>2 (-ve)</td>
<td>5,157</td>
<td>5,150,297,217-150,302,373</td>
<td>NM_00102301</td>
<td>Q5FYB1</td>
<td>569</td>
<td>64,030 (8.8)</td>
</tr>
<tr>
<td>ARSJ</td>
<td>2 (-ve)</td>
<td>76,558</td>
<td>4,113,902,277-113,978,834</td>
<td>NM_024590</td>
<td>Q5FYB0</td>
<td>599</td>
<td>67,235 (9.2)</td>
</tr>
<tr>
<td>GNS</td>
<td>14 (+ve)</td>
<td>42,533</td>
<td>12,64,716,744-64,759,276</td>
<td>NM_020754</td>
<td>P15586</td>
<td>552</td>
<td>62,082 (8.6)</td>
</tr>
<tr>
<td>SULF1</td>
<td>18 (+ve)</td>
<td>74,885</td>
<td>8,69,563,976-69,638,860</td>
<td>NM_00128200</td>
<td>Q8VU6</td>
<td>871</td>
<td>101,027 (9.2)</td>
</tr>
<tr>
<td>IDS</td>
<td>9 (-ve)</td>
<td>22,389</td>
<td>X:149,842,749-149,505,137</td>
<td>NM_000207</td>
<td>P22304</td>
<td>550</td>
<td>61,873 (5.2)</td>
</tr>
<tr>
<td>ARSK</td>
<td>8 (+ve)</td>
<td>48,245</td>
<td>5,95,555,279-95,603,523</td>
<td>NM_198150</td>
<td>Q6IUY0</td>
<td>536</td>
<td>61,450 (9.0)</td>
</tr>
<tr>
<td>SGSH</td>
<td>8 (-ve)</td>
<td>9,859</td>
<td>17,80,210,455-80,220,313</td>
<td>NM_001999</td>
<td>P5I688</td>
<td>502</td>
<td>56,695 (6.5)</td>
</tr>
<tr>
<td>ARSA</td>
<td>8 (-ve)</td>
<td>2,626</td>
<td>22,50,625,148-50,627,773</td>
<td>NM_000487</td>
<td>P15289</td>
<td>507</td>
<td>53,588 (5.7)</td>
</tr>
<tr>
<td>ARSG</td>
<td>11 (+ve)</td>
<td>1,12,967</td>
<td>17,68,307,494-68,420,460</td>
<td>NM_00126772</td>
<td>Q96EG1</td>
<td>525</td>
<td>57,061 (6.2)</td>
</tr>
<tr>
<td>STS</td>
<td>10 (+ve)</td>
<td>97,065</td>
<td>X:7,253,194-7,350,258</td>
<td>NM_000351</td>
<td>P08842</td>
<td>583</td>
<td>65,492 (7.6)</td>
</tr>
<tr>
<td>GALNS</td>
<td>14 (+ve)</td>
<td>42,536</td>
<td>16,88,880,850-88,923,285</td>
<td>NM_000512</td>
<td>P34059</td>
<td>522</td>
<td>58,026 (6.3)</td>
</tr>
</tbody>
</table>

*NCBI sequence; pI=plietodic point
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 nasopharyngeal 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 ARSI 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 ARSI 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.

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

**Figure 4:** 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:

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

