Holmes, J Proteomics Bioinform 2017, 10:2
DOI: 10.4172/jpb.1000423

Journal of
Proteomics & Bioinformatics

Research Article Open Access

Comparative and Evolutionary Studies of Vertebrate Extracellular Sulfatase Genes and Proteins: *SULF1* and *SULF2*

Roger S Holmes*

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

Abstract

Extracellular sulfatases (*SULF1; SULF2*) (EC: 3.1.6.-) are members of the sulfatase enzyme family which exhibit endoglucosamine-6-sulfatase activity and carry out essential roles in proteoglycan metabolism. These enzymes regulate a number of critical signalling pathways and the sulfation state of glycoaminoglycans in the extracellular space. *SULF1* and *SULF2* amino acid sequences and structures and *SULF*-like gene locations were examined using bioinformatic data from several genome projects. Sequence alignments and conserved secondary structures and key amino acid residues and domains were studied. Comparative genomic analyses were conducted using the UC Santa Cruz Genome Browser. Phylogeny studies investigated the evolutionary relationships of these genes and proteins. Human and other vertebrate *SULF1* and *SULF2* sequences were conserved, including signal peptides, metal (Ca^{2+}) and substrate binding sequences, active site residues and N-glycosylation sites (sulfatase domain); and a C-terminal positively charged hydrophilic domain. Predicted 2D structures were identified for the sulfatase domain of vertebrate *SULF1* and *SULF2* using a bacterial phosphatase structure (PDB:4UPK). Vertebrate *SULF1* and *SULF2* genes usually contained 18/19 or 20 coding exons, respectively. Transcription factor binding sites and miR-binding sites were identified within the human *SULF1* and *SULF2* gene promoters and 3'-UTR regions, respectively. The Estrogen Receptor Gene (ESR1) was identified in the *SULF2* promoter which may contribute to the higher expression level for this gene in female reproductive tissues. *SULF1* and *SULF2* genes and proteins were present in all vertebrate genomes examined. Phylogenetic analyses suggested that an ancestral invertebrate SUL1 gene underwent a gene duplication event to form two separate lines of vertebrate gene evolution: *SULF1* and *SULF2*.

Keywords: Vertebrates; Invertebrates; Amino acid sequence; Signal peptide; Ca^{2+} binding; N-glycosylation; *SULF*: Extracellular sulfatase; Evolution; Gene duplication; Phylogeny; Sulfate metabolism

Abbreviations: ARS: Arylsulfatase; *SULF*: Sulfatase; HSPG: Heparin Sulfate Proteoglycan; FGF: Fibroblast Growth Factor; WNT: Wingless-related Integration; VEGF: Vascular Endothelial Growth Factor; GDNF: Glial Cell Line-Derived Neurotrophic Factor; PNH: Phosphonate Monoester Hydrolase; BLAST: Basic Local Alignment Search Tool; BLAT: Blast-Like Alignment Tool; NCBI: National Center for Biotechnology Information; UCSC: University of California Santa Cruz; KO: Knock Out; AceView: NCBI Based Representation of Public mRNAs; SWISS-MODEL: Automated Protein Structure Homology-Modeling Server; TFBS: Transcription Factor Binding Sites; UTR: Untranslated Region

Introduction

Extracellular sulfatases 1 and 2 (*SULF1*; *SULF2*; E.C.3.1.6.-; also described as heparan sulfate 6-O-endosulfatase; *SUL1*; *SUL2*) are members of the sulfatase enzyme family, for which seventeen genes have been described on the human genome [1,2]. *SULF1* and *SULF2* are secreted enzymes which carry out essential roles in the extracellular environment by catalysing endoglucosamine-6-sulfatase activity and removing 6-O sulfate groups from Heparin Sulfate Proteoglycans (HSPGs) [3,4]. These perform several key roles: modulating the activity of growth factor receptors and cell signaling pathways, such as FGF, VEGF, GDNF and WNT signaling pathways, which initiate gene transcription signals through cell surface receptors [5-11]; serving essential roles in vertebrate development [12-20]; and in modulating microbial (*Chlamydia muridarum*) infection [21].

Structures for vertebrate *SULF1* and *SULF2* genes and cDNA sequences have been reported, including human (*Homo sapiens)* [3]; mouse (*Mus musculus*) [5,22,23]; rat (*Rattus norvegicus*) [24,25]; frog (*Xenopus laevis*) [13-14]; and zebra fish (*Danio rerio*) [16] *SULF* genes. Human *SULF1*, which spans 194.3 kilobases and comprises 22 exons, is localized on chromosome 8; whereas human *SULF2* spans 128.7 kilobases and comprises 21 exons on chromosome 20 [26,27]. Both of these genes are widely expressed in the body, consistent with their overlapping and essential roles in cell signaling pathways, skeletal muscle regeneration, neonatal development and survival, metastasis and wound repair [9,18,19,23].

This paper reports the predicted gene structures and amino acid sequences for several vertebrate *SULF1* and *SULF2* genes and proteins, the predicted secondary and tertiary structures for human *SULF1* and *SULF2* protein subunits, and the structural, phylogenetic and evolutionary relationships for these genes and enzymes. Evidence is also presented for *SULF2* playing a significant role in female reproductive tissues involving the estrogen receptor localized within the *SULF2* promoter [28,29].

Materials and Methods

Gene and protein identification

BLAST studies were undertaken using web tools from the NCBI

***Corresponding author:** Roger S Holmes, Emeritus Professor, Griffith Institute for Drug Discovery and School of Natural Sciences, Griffith University, Nathan, QLD, 4111 Australia, Tel: 61-410-583-348; E-mail: r.holmes@griffith.edu.au

Received January 10, 2017; **Accepted** February 02, 2017; **Published** February 16, 2017

Citation: Holmes RS (2017) Comparative and Evolutionary Studies of Vertebrate Extracellular Sulfatase Genes and Proteins: *SULF1* and *SULF2*. J Proteomics Bioinform 10: 32-40. doi: 10.4172/jpb.1000423

(http://www.ncbi.nlm.nih.gov) [30]. Protein BLAST analyses used human ARS sequences (Group 1: ARSA, ARSG, GALNS; Group 2: ARSB, ARSI, ARSJ; Group 3: ARSD, ARSE, ARSF, ARSH, STS; Group 4: *SULF*1, *SULF2*, GNS; Group 5: ARSK; Group 6: SGSH; Group 7: IDS); and other vertebrate *SULF1* and *SULF2* amino acid sequences previously described (Tables 1 and 2) [2-4]. Predicted *SULF1* and *SULF2*-like protein sequences were obtained in each case and subjected to protein and gene structure analyses.

BLAT analyses were undertaken for each of the predicted *SULF1* and *SULF2* amino acid sequences using the UCSC Genome Browser (http://genome.ucsc.edu) with the default settings to obtain the predicted locations for each of the vertebrate *SULF*-like genes, including exon boundary locations and gene sizes [27]. The structures for the major human *SULF1* and *SULF2* trancripts were obtained using the AceView website (http://www.ncbi.nlm.nih.gov/ieb/research/acembly/) [26]. Alignments of *SULF* sequences with human *SULF1* and *SULF2* protein sequences were assembled using the Clustal Omega multiple sequence alignment program [31]. Predicted micro-RNA binding sites (miR) and CpG islands [32] were examined using the UCSC Genome Browser [27]. Predicted human *SULF1* and *SULF2* transcription factor binding sites (TFBS) were obtained from the PAZAR (OregAnno) dataset [33] (http://www.oreganno.org).

Structures and predicted properties of *SULF1* and *SULF2* proteins

Predicted secondary structures for human and other mammalian *SULF1* and *SULF2* proteins were obtained using the SWISS-MODEL web-server [34] and the reported structures using bacterial phosphonate monoester hydrolase (PNH) from *Silicibacter pomeroyi* (PDB:4UPKA) with modeling residue ranges of 42-416 for human *SULF1* and 43-419 for human *SULF2* (Figure 1). Predicted secondary structures for the hydrophilic zones for both *SULF1* (residues 397-871) and *SULF2* (residues 398-870) were undertaken using the PSIPRED web server [35]. Identification of conserved domains for vertebrate *SULF1* and *SULF2* proteins was made using NCBI web tools [36].

Comparative human *SULF1* and *SULF2* gene expression

RNA-seq gene expression profiles across 53 selected tissues (or tissue segments) were examined from the public database for human *SULF1* and *SULF2*, based on expression levels for 175 individuals [37] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (http://www.gtexportal.org).

Phylogeny studies and sequence divergence

Phylogenetic analyses were undertaken using the http://phylogeny.fr platform [38]. Alignments of *SULF1* and *SULF2* sequences were assembled using MUSCLE (Table 1) [39]. Alignment ambiguous regions were excluded prior to phylogenetic analysis yielding alignments for comparisons of these sequences. The phylogenetic tree was constructed using the maximum likelihood tree estimation program PhyML [40].

Results and Discussion

SULF1, SULF2 and other human sulfatase genes and proteins

Table 2 summarises the comparative genomic and proteomic features for 17 human sulfatase genes and proteins, including *SULF1* and *SULF2*, which are members of the human Group 4 *ARS* genes [2]. These genes were separately located on human chromosomes (chromosomes 8 and 20, respectively. This is in contrast to Group 3 *ARS* genes (*ARSD, ARSE, ARSF, ARSH* and *STS*), which are localized consecutively within a sterylsulfatase gene cluster on the human X-chromosome (Table 2). *SULF1, SULF2* and *GNS* genes have been designated as belonging to ARS Group 4 [2], due to their higher sequence identities (40-67%) than with other human ARS enzymes (12-22% identical), and to similarities in substrate specificities, acting on either endoglucosamine 6-sulfate (*SULF1* and *SULF2*) [3,9,12] or N-acetylglucosamine 6-sulfate (GNS) substrates [41,42], respectively. In addition, these genes have apparently been derived from a common invertebrate ancestral gene, *SUL1* (identified in *C. elegans*) and *SULF1* (identified in *D. melanogaster*) [2,11].

Alignments of *SULF1* and *SULF2* subunits

Alignments of amino acid sequences for human *SULF1* and *SULF2* subunits previously reported [3] are shown in Figure 1. The sequences were 66% sequence identitical (Table 3), suggesting that these are products of two related families of genes and proteins, namely *SULF1* and *SULF2* (Table 2). Studies of the amino acid sequences for other vertebrate SULF subunits have shown that they contained 867-892 residues for *SULF1*, whereas vertebrate *SULF2* subunits contained 870-888 residues (Table 1), with higher levels of sequence identity observed for subunits from the same gene family, in each case (Table 3). Several key amino acid residues or regions for human *SULF1* and *SULF2* were recognized (sequence numbers refer to human *SULF1* (Figure 1). These included the leader peptide (residues 1-22 for *SULF1*; 1-24 for *SULF2*); metal binding residues at the active site (Ca^{2+}) (51Asp, 52Asp, 316Asp and 317His); the active site 87Cys, which functions by forming the 3-oxoalanine residue; and seven N-glycosylation sites, located in the N-terminal region (64Asn, 111Asn, 131Asn, 148Asn, 170Asn, 197Asn and 240Asn) and three N-glycosylation sites in the C-terminal region (623Asn, 773Asn and 783Asn). Comparisons of human *SULF1* and *SULF2* amino acid sequences with other human ARS sequences showed that *SULF1* and *SULF2* subunits contained extended C-terminal sequences (with >300 additional amino acid residues) (Table 2). Moreover, these C-terminal regions contained high basic amino acid content, assisting the formation of ionic linkages between *SULF1* and *SULF2* subunits with the heparan sulfate proteoglycan substrates in the extracellular envoronment, where the enzymes operate to modify the structures of the heparan sulfate chains [10,43]. In addition to these clusters of basic amino acid residues, the *SULF1* C-terminal region contained a poly-Glu (x5) acidic amino acid zone (Glu560-564) which may be involved in the formation of ionic linkages with the highly basic C-terminus (Figure 1).

Predicted secondary structures of *SULF1* and *SULF2* subunits

Analyses of predicted secondary structures for human *SULF1* and *SULF2* sequences were obtained using the SWISS-MODEL web-server [32] and the reported tertiary structures using bacterial Phosphonate Monoester Hydrolase (PNH) from *Silicibacter pomeroyi* (PDB:4UPKA) (Figure 1). Several α-helix and β-sheet structures were observed for the human *SULF1* and *SULF2* subunits examined, with 11 β-sheet and 7 α-helices predicted. Of particular interest was the prediction of β-sheet and α-helix structures at the N-terminal end of the *SULF* subunits, in comparison with extended hydrophilic C-terminal sequence. Secondary structures were readily apparent near key residues or functional domains including the β-sheet and α-helix structures near the substrate binding active site (87Cys) and the metal binding residues at the active site (Ca^{2+}) (51Asp, 52Asp, 316Asp and 317His) [3,10].

The predicted secondary structures for human *SULF1* and *SULF2* showed similarities to structures previously reported for other ARS proteins, including human ARSA [44], ARSB [45], STS [46] and SGSH

Gene	Organism	Species	Chromosome location	Coding Exons (strand)	Gene Size bps	GenBank ID*	UNIPROT ID	Amino acids	Subunit MW (pI)	Leader Peptide
SULF1	Human	*Homo sapiens*	8:69,563,976-69,638,860	18 (+ve)	74,885	NM_001128204	Q8IWU6	871	101,027 (9.2)	1..22
SULF1	Baboon	*Papio anubis*	8:65,467,074-65,541,661	19 (+ve)	74,588	*XP_003902891	A0A096MPY6	869	100,761 (9.2)	1..22
Sulf1	Mouse	*Mus musculus*	1:12,786,527-12,848,462	18 (+ve)	61,936	NM_001198565	Q8K007	870	100,923 (9.2)	1..22
SULF1	Opossum	*Mondelphis domestica*	3:167,373,907-167,459,171	18 (-ve)	85,265	*XP_007487069	F7DW81	872	100,915 (9.2)	1..22
SULF1	Chicken	*Gallus gallus*	2:115,993,905-116,052,482	19 (+ve)	58,578	*XP_015138388	E1BRF7	867	100,410 (9.2)	1..22
SULF1	Lizard	*Anolis carolinensis*	4:31,950,638-31,992,762	19 (+ve)	42,125	*XP_016848361	G1KQZ3	878	101,578 (8.9)	1..22
SULF1	Frog	*Xenopus tropicalis*	^KB021656:29,757,270-29,789,632	18 (-ve)	32,363	NM_001097379	F6X5B1	884	102,523 (8.5)	1..22
SULF1	Zebra fish	*Danio rerio*	24:19,374,157-19,443,110	19 (+ve)	42,892	NM_001003846	Q6EF99	892	103,540 (9.2)	1..21
SULF2	Human	*Homo sapiens*	20:47,659,398-47,757,363	20 (-ve)	97,966	NM_001161841	Q8IWU5	870	100,455 (9.3)	1..24
SULF2	Baboon	*Papio anubis*	10:16,419,479-16,547,200	20 (+ve)	#######	*XP_003902891	A0A096NSD4	870	100,488 (9.3)	1..24
Sulf2	Mouse	*Mus musculus*	2:166,075,494-166,132,762	20 (-ve)	57,269	NM_001252578	Q8CFG0	875	100,497 (9.2)	1..24
SULF2	Opossum	*Mondelphis domestica*	1:498,521,847-498,631,016	20 (+ve)	#######	*XP_001379302	F7C2B7	878	101,667 (9.2)	1..24
SULF2	Chicken	*Gallus gallus*	20:6,263,766-6,319,814	20 (-ve)	56,049	*XP_004947107	E1BZH8	877	102,208 (9.3)	1..24
SULF2	Lizard	*Anolis carolinensis*	4:145,321,427-145,439,481	20 (-ve)	#######	*XP_003220666	G1KSG0	888	103,272 (9.0)	1..30
SULF2	Frog	*Xenopus tropicalis*	^KB021662:12,053,932-12,080,096	20 (-ve)	26,165	NM_001005661	Q6GL29	875	101,598 (9.3)	1..19
SULF2	Zebra fish	*Danio rerio*	11:24,728,567-24,752,009	20 (+ve)	23,443	NM_200936	Q7ZVU8	873	100,578 (9.5)	1..24
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

*=Predicted sequence; ^=Gene scaffold ID; pI=Isoelectric point; bps=Base pairs of nucleotide sequence.

Table 1: Vertebrate *SULF1* and *SULF2* and *Caenorhabditis elegans SUL-1* genes and proteins.

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

Note the proposed classification of human arylsulfatase genes and proteins into 7 groups; *SULF1* and *SULF2* are highlighted in red; pI=Isoelectric point; bps=Base pairs of nucleotide sequence.

Table 2: Proposed classification of human arylsulfatase genes and proteins.

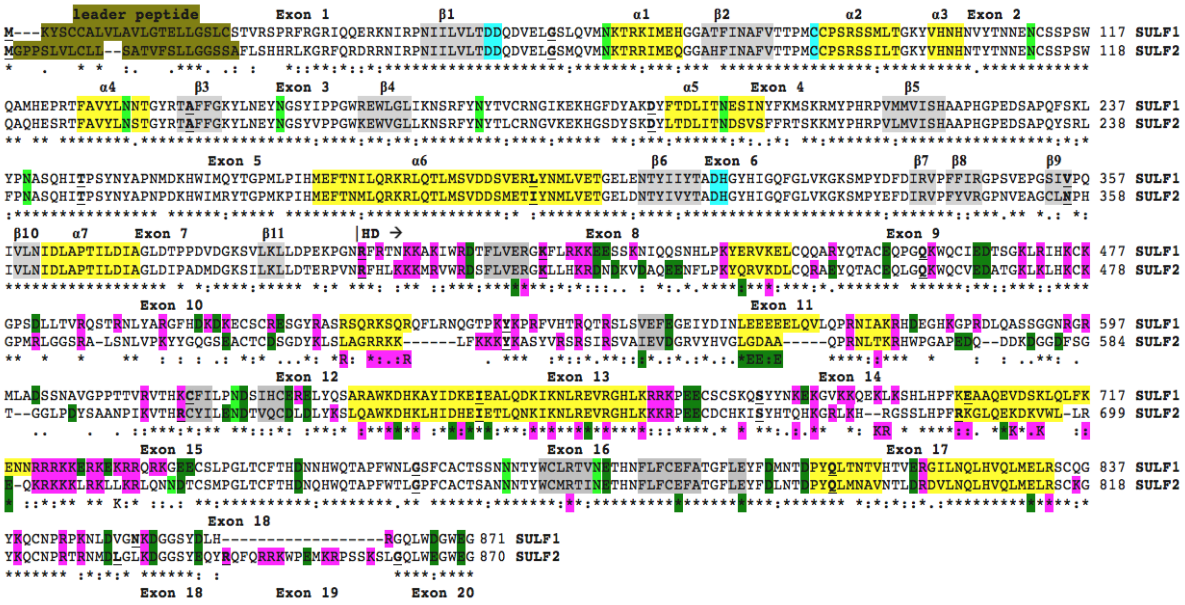

Figure 1: Amino acid sequence alignments for human *SULF1* and *SULF2* subunits. See Table 1 for sources of SULF1 and SULF2 sequences; *Shows identical residues for SULF subunits; similar alternate residues; .dissimilar alternate residues; leader peptide residues are in dark yellow;predicted helix is shown in yellow; predicted sheet is shown in grey; active site residues shown in blue; N-glycosylated Asn residues are in light green; HD refers to hydrophilic C-terminal sequence; acidic amino acids in HD zone are in dark green; basic amino acid residues in HD zone are in pink; bold font shows known or predicted exon junctions; exon numbers refer to human *SULF1* gene.

[47]. The active site for *SULF1* was centrally located with two β-sheet structures (β1, β6) and the metal binding residues at the active site (Ca^{+}) (51Asp, 52Asp, 316Asp and 317His). The hydrophilic C-terminal region was absent in the ARSA, ARSB, STS and SGSH proteins previously reported [44-47]. The positively charged Hydrophilic Domain (HD) domain has been previously characterized as having high affinity with heparan/heparan sulfate, with specific regions influencing different aspects of heparan sulfate binding, cellular localization and enzyme function [4].

Predicted gene locations and exonic structures for vertebrate and invertebrate *SULF* genes

Table 1 summarizes the predicted locations for vertebrate *SULF1* and *SULF2* genes based upon BLAT interrogations of genomes using the reported sequences for human, mouse and frog *SULF1* and *SULF2* [3-5,14] and the predicted sequences for other *SULF1* and *SULF2* proteins and the UCSC Genome Browser [27]. Human *SULF1* and *SULF2* genes were located on different chromosomes (chromosomes 18 and 20, respectively), which is the case for all vertebrate genomes examined (Table 1). Of particular interest to the evolution of *SULF*-like genes in invertebrate genomes, the worm (*Caeborhabditis elegans*) showed evidence of having only one gene which was similar to the vertebrate *SULF1* gene which encoded a *SULF*-like gene (designated as *sul1*). This amino acid sequence also encoded a leader peptide, similar to that for human *SULF1* and *SULF2* proteins.

Figure 1 summarizes the predicted exonic start sites for human *SULF1* and *SULF2* genes which contained 18 or 20 exons, respectively, in identical or similar positions, with the exception of 2 additional exons (exons 19 and 20) encoded at the C-terminus end of SULF2. In each case, exon 1 encoded the leader peptide and the double aspartate (Asp51-Asp52) Ca^{2+} binding site; exon 2 encoded the active site 87Cys, which functions by forming the 3-oxoalanine residue; exon 6 encoded two other active site residues, 316Asp and 317His; and exons 9-18 (or 20, in the case of SULF2) encoded the hydrophilic C-terminus region.

Figure 2 illustrates the predicted structures of mRNAs for human *SULF1* and *SULF2* transcripts for the major transcript isoforms in each case [26]. The genes cover 194.3 and 130.5 kilobases in length, respectively, with 18 introns and 20 exons present for the mRNA transcripts. The human *SULF1* gene promoter contained six predicted TFBS (Figure 2 and Table 4), including 3 binding sites for *FOXA1*, encoding hepatocyte nuclear factor 3-alpha, which participates in embryonic development and directs tissue-specific gene expression [48]; 2 binding sites for *TFA2PC,* encoding transcription factor AP-2 gamma, which is involved in eye, face, body wall, limb and neural tube development [49]; and a binding site for *EGR1*, encoding early growth response protein 1, a gene regulator which regulates the transcription of several genes involved in early vertebrate development [50]. Three of these TFBS were also observed for the *SULF2* promoter, including *FOXA1, TFA2PC* and *EGR1*, although three others were found in this region, including *ESR1*, encoding the estrogen receptor [51]; *HNF4A*, encoding hepatocyte nuclear factor 4-alpha, controlling several genes essential for the development of liver, intestine and kidney [52]; and *CTCF*, encoding CCCTC-binding factor, which is necessary for memory formation and for basal and experience-dependent gene regulation [53].

Many microRNA binding sites were located in the 3'-UTR of human *SULF1*, which are potentially of major significance for the regulation of this gene (Supplementary Table 1 and Figure 3). A recent study of miR-19 has shown that it contributes to the regulation of newborn neuronal cell migration and is enriched in neural progenitor cells [54]. Several other miR binding sites within the 3'-UTR of human *SULF1* have been reported with significant roles in regulating cell proliferation during carcinogenesis, including miR-26, miR-205, miR-130, miR-148, miR-26, miR-1, miR-200, miR-140, miR-145, miR-17, miR-202, miR-433

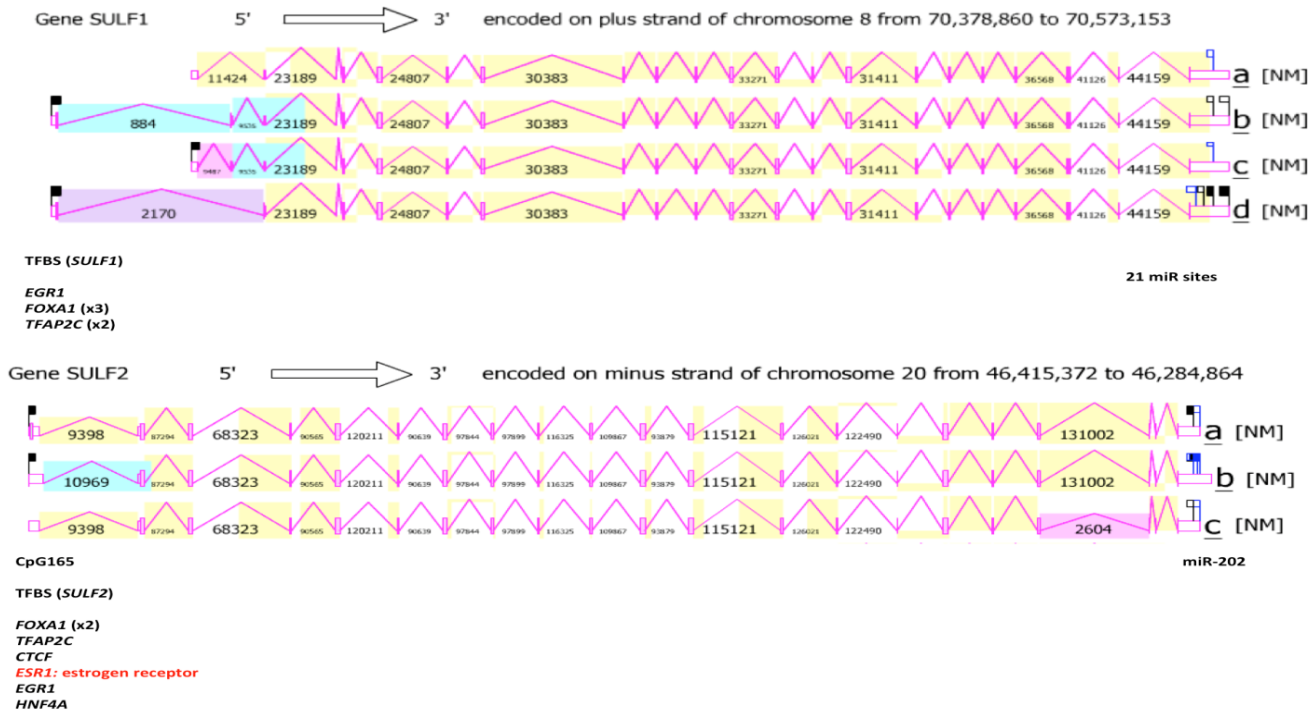

Figure 2: Gene Structures for the Human *SULF1* and *SULF2* genes. Derived from the AceView website http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/ [26]; the major isoform variants are shown with capped 5'- and 3'- ends for the predicted mRNA sequences; introns (pink lines) and exons (pink boxes) are shown; the length of the mRNAs (as kilobases or kb) are shown; a CpG island (CpG165) is shown for the *SULF2* promoter; 21 miR-binding sites were observed in for the 3'UTR of the human *SULF1* gene (Supplementary Table 1); a miRNA-202 binding site was identified for the 3'UTR of the human *SULF2* gene; the direction for transcription is shown; TFBS refers to transcription factor binding sites located within the *SULF1* and *SULF2* gene promoters (Table 4); individual TFBS were identified within these promoter regions, including the estrogen receptor (*ESR1*) within the *SULF2* promoter.

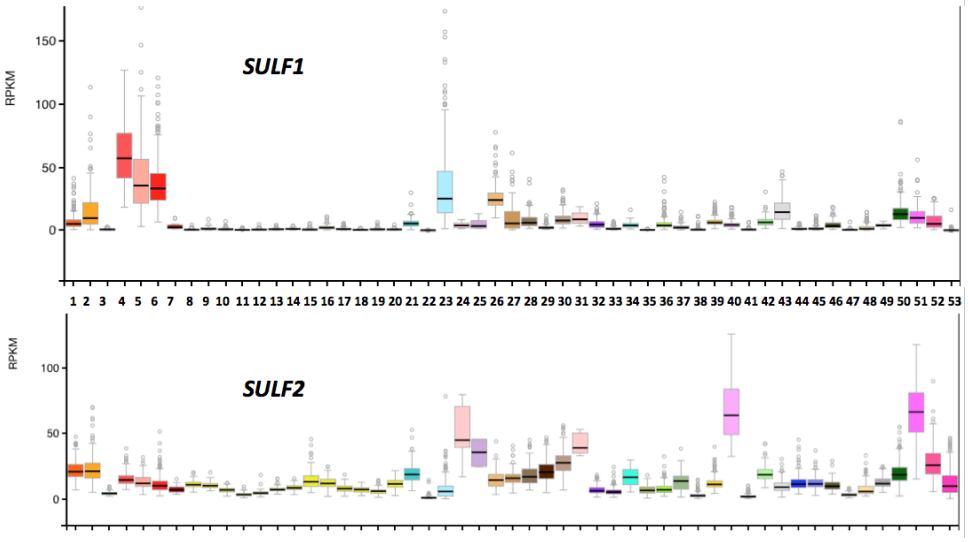

Figure 3: Comparative Tissue Expression for Human *SULF1* and *SULF2* genes. RNA-seq gene expression profiles across 53 selected tissues (or tissue segments) were examined from the public database for human *SULF1* and *SULF2*, based on expression levels for 175 individuals [37] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (http://www.gtexportal.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.

SULF	Human	Mouse	Zebra fish	Human	Mouse	Zebra fish	Worm
Protein	*SULF1*	*SULF1*	*SULF1*	*SULF2*	*SULF2*	*SULF2*	*SUL-1*
Human *SULF1*	100	93	73	66	65	59	45
Mouse *SULF1*	93	100	72	65	64	59	44
Zebra fish *SULF1*	73	72	100	65	64	59	44
Human *SULF2*	66	65	65	100	95	69	44
Mouse *SULF2*	64	64	64	95	100	67	44
Zebra fish *SULF2*	59	59	59	69	67	100	44
Worm *SUL1*	45	44	44	44	44	44	100

Table 3: Percentage identity matrix for vertebrate and *Caenorhabditis elegans SULF* amino acid sequences.

PAZAR Data Set	ORegAnno ID	Location	Strand	Genomic Size	Transcription Factor	UNIPROT ID	Genomic Role
TFBS *SULF1*	OREG1488705	Chr8:70383809-70384340	(+)	532	EGR1	P18146	Early growth factor response protein
	OREG1573708	Chr8:70399986-70400646	(+)	661	FOXA1	P55317	Embryonic development tissue specific expression
	OREG1168635	Chr8:70400026-70401446	(+)	1421	TFAP2C	Q92754	Eye, face, body wall, limb and neural tube development
	OREG1090824	Chr8:70401122-70401136	(-)	15	TFAP2C	Q92754	Eye, face, body wall, limb and neural tube development
	OREG1573711	Chr8:70401386-70401986	(+)	601	FOXA1	P55317	Embryonic development tissue specific expression
	OREG1632397	Chr8:70401396-70402016	(+)	621	FOXA1	P55317	Embryonic development tissue specific expression
TFBS *SULF2*	OREG1646208	Chr20:46412693-46414183	(+)	1491	FOXA1	P55317	Embryonic development tissue specific expression
	OREG1587772	Chr20:46412693-46414123	(+)	1431	FOXA1	P55317	Embryonic development tissue specific expression
	OREG1181298	Chr20:46413243-46415853	(+)	2611	TFAP2C	Q92754	Eye, face, body wall, limb and neural tube development
	OREG1375314	Chr20:46413443-46413630	(+)	188	CTCF	P49711	Transcriptional regulation by binding to chromatin insulators
	OREG1532513	Chr20:46413719-46413779	(+)	61	ESR1	P03372	Estrogen receptor nuclear hormone receptor
	OREG1502386	Chr20:46414183-46415359	(+)	1177	EGR1	P18146	Early growth factor response protein
	OREG1718320	Chr20:46414739-46414847	(+)	109	HNF4A	P41235	Hepatocyte nuclear factor 4-alpha essential for liver development
TFBS were identified using the PAZAR data set [31]; UNIPROT refers to Universal Protein Resource (uniprot.org); PAZAR identifies TFBS by OregAnno IDs.							

Table 4: Transcription factor binding sites (TFBS) identified for human *SULF1* and *SULF2* gene promoters.

and miR-137 (Supplementary Table 1). In addition, miR-202, located in the 3'-UTR of human *SULF2,* has been shown to inhibit the progression of human cervical cancer [55].

Comparative *SULF1* and *SULF2* human tissue gene expression

Figure 3 shows comparative gene expression for various human tissues obtained from RNA-seq gene expression profiles for human *SULF1* and *SULF2* genes obtained for 53 selected tissues or tissue segments for 175 individuals [37] (Data Source: GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1) (http://www.gtex.org). These data supported a much higher level of tissue expression for human *SULF1* in arterial and fibroblast cells, and for *SULF2* in female reproductive tissues, including cervix, ovary, uterus and vagina. The presence of multiple TFBS within the *SULF1* gene promoter (*EGR1, FOXA1* and *TFA2PC*) and the *SULF2* (*ESR1, EGR1, HNF4A, CTCF* and *FOXA1*) gene promoter may contribute to this high level in expression level for these genes. In addition, the presence of the binding site for the estrogen receptor (*ESR1*) within the *SULF2* promoter, which is highly expressed in female reproductive tissues, is potentially of major significance for this enhanced *SULF2* expression profile.

Phylogeny and Divergence of Vertebrate *SULF1* and *SULF2*

A phylogenetic tree (Figure 4) was calculated by the progressive alignment of human and other vertebrate *SULF1* and *SULF2* amino acid sequences with an invertebrate (worm: *Caeborhabditis elegans)* sequence (*SUL1*). The phylogram was 'rooted' with this *C. elegans SUL1* sequence and showed clustering of the *SULF*-like sequences into two groups: vertebrate *SULF1* and *SULF2* sequences. Overall, these data suggest that the vertebrate *SULF1* and *SULF2* genes arose from a gene duplication event of an ancestral invertebrate *SULF*-like gene, resulting in two separate lines of vertebrate gene evolution for *SULF1*-like and *SULF2*-like genes. This is supported by the comparative biochemical and genomic evidence for vertebrate *SULF1* and *SULF2*-like genes and encoded proteins, which shared several key features of protein and gene

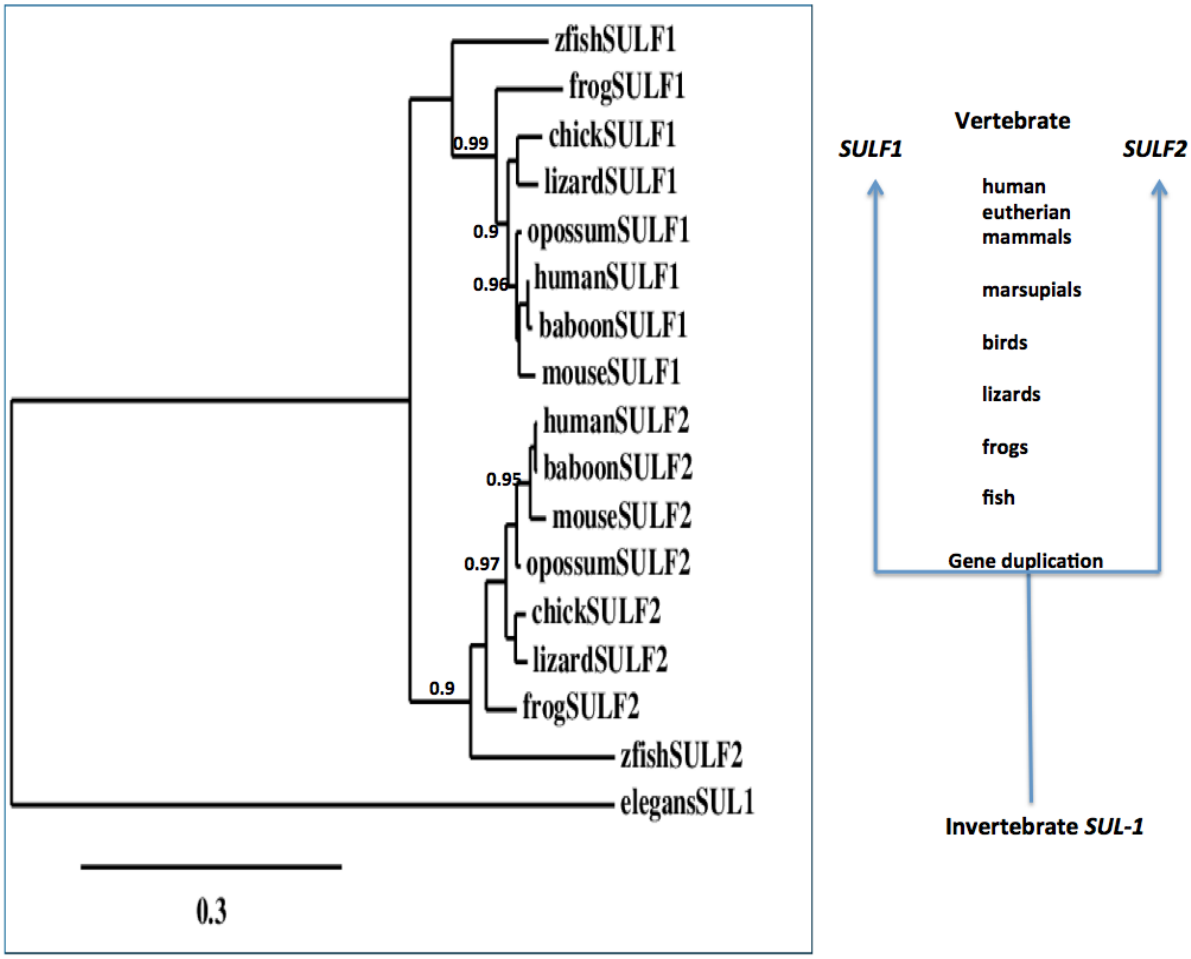

Figure 4: Phylogenetic tree of vertebrate *SULF1* and *SULF2* amino acid sequences. The tree is labeled with the gene name and the name of the animal. Note 2 major clusters for the vertebrate *SULF1* and vertebrate *SULF2* sequences. The tree is 'rooted' with the worm (*Caenorhabditis elegans*) *SUL1* sequence. Table 1 for details of sequences and gene locations. A genetic distance scale is shown (% amino acid substitutions). The number of times a clade (sequences common to a node or branch) occurred in the bootstrap replicates is shown. Only replicate values of 0.9 or more which are highly significant are shown with 100 bootstrap replicates performed in each case. An evolutionary model for a proposed gene duplication event of an ancestral invertebrate *SULF (SUL-1)* gene is shown.

structure, including having similar alpha-beta secondary structures (Figure 1). In addition, the locations of vertebrate *SULF1* and *SULF2* genes on separate chromosomes (Table 1) may reflect on a possible mechanism for ancestral vertebrate *SULF* gene duplication by whole-genome duplication rather than by an unequal crossover event of a single ancestral chromosome, as exemplified by studies supporting at least two rounds of whole genome duplication during early vertebrate evolution [56].

Conclusion

In conclusion, the results of the present study suggested that vertebrate *SULF1* and *SULF2* genes and encoded *SULF1* and *SULF2* enzymes represented a distinct arylsulfatase enzyme and gene family which share key conserved sequences and structures with those reported for other arylsulfatase gene families [1,2]. *SULF1* has been recognized as a major extracellular sulfatase expressed in many tissues of the body, particularly in arterial and fibroblast cells, which plays a specific role in removing sulfate from heparan sulfate proteoglycan extracellular substrates, catalysing the hydrolysis of endoglucosamine-6-sulfate residues. *SULF2* has also been described as a second major extracellular sulfatase expressed in many tissues of the body, particularly in female reproductive tissues, also with a specific endoglucosamine-6-sulfatase role [3]. Bioinformatic methods were used to predict the amino acid sequences, secondary and tertiary structures and gene locations for *SULF1* and *SULF2* genes and encoded proteins using data from several vertebrate genome projects. Vertebrate *SULF* protein subunits shared 59-93% sequence identities and exhibited sequence alignments and identities for key *SULF* amino acid residues as well as conservation of predicted secondary structures with those previously reported for a bacterial phosphonate monoester hydrolase from *Silicibacter pomeroyi* (PDB:4upk). Phylogenetic analyses demonstrated the relationships and potential evolutionary origins of the vertebrate *SULF1* and *SULF* gene families which were related to a worm (*Caeborhabditis elegans)* extracellular sulfatase (*SUL1*) gene and protein. These studies indicated that *SULF1* and *SULF2* genes may have appeared early in vertebrate evolution following gene duplication of an ancestral *SUL*-like gene, following whole-genome duplication in the vertebrate ancestor.

References

1. Ratzka A, Mundlos S, Vortkamp A (2010) Expression patterns of sulfatase genes in the developing mouse. Dev Dyn 239: 1779-1788.
2. 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 Prot Bioinform 9: 298-305.
3. Morimoto-Tomita M, Uchimura K, Werb Z, Hemmerich S, Rosen SD (2002) Cloning and characterization of two extracellular heparin-degrading endosulfatases in mice and humans. J Biol Chem 277: 49175-49185.
4. Frese MA, Milz F, Dick M, Lamanna WC, Dierks T (2009) Characterization of the human sulfatase Sulf1 and its high affinity heparin/heparan sulfate interaction domain. J Biol Chem 284: 28033-28044.
5. Lamanna WC, Baldwin RJ, Padva M, Kalus I, Ten Dam G, et al. (2006) Heparan sulfate 6-O-endosulfatases: discrete in vivo activities and functional co-operativity. Biochem J 400: 63-73.
6. Sahota AP, Dhoot GK (2009) A novel SULF1 splice variant inhibits Wnt signalling but enhances angiogenesis by opposing SULF1 activity. Exp Cell Res 315: 2752-2764.
7. Wojcinski A, Nakato H, Soula C, Glise B (2011) DSulfatase-1 fine-tunes Hedgehog patterning activity through a novel regulatory feedback loop. Dev Biol 358: 168-180.
8. Langsdorf A, Schumacher V, Shi X, Tran T, Zaia J, et al. (2011) Expression regulation and function of heparan sulfate 6-O-endosulfatases in the spermatogonial stem cell niche. Glycobiology 21: 152-161.
9. Tran TH, Shi X, Zaia J, Ai X (2012) Heparan sulfate 6-O-endosulfatases (Sulfs) coordinate the Wnt signaling pathways to regulate myoblast fusion during skeletal muscle regeneration. J Biol Chem 287: 32651-32664.
10. Fellgett SW, Maguire RJ, Pownall ME (2015) Sulf1 has ligand-dependent effects on canonical and non-canonical Wnt signalling. J Cell Sci 128: 1408-1421.
11. Nakato H, Li JP (2016) Functions of heparansulfate proteoglycans in development: insights from Drosopila models. Int Rev Cell Mol Biol 325: 275-293.
12. Holst CR, Bou-Reslan H, Gore BB, Wong K, Grant D, et al. (2007) Secreted sulfatases Sulf1 and Sulf2 have overlapping yet essential roles in mouse neonatal survival. PLoS One 2: e575.
13. Freeman SD, Moore WM, Guiral EC, Holme AD, Turnbull JE, et al. (2008) Extracellular regulation of developmental cell signaling by XtSulf1. Dev Biol 320: 436-445.
14. Winterbottom EF, Pownall ME (2009) Complementary expression of HSPG 6-O-endosulfatases and 6-O-sulfotransferase in the hindbrain of Xenopus laevis. Gene Expr Patterns 9: 166-172.
15. Guiral EC, Faas L, Pownall ME (2010) Neural crest migration requires the activity of the extracellular sulphatases XtSulf1 and XtSulf2. Dev Biol 341: 375-388.
16. Gorsi B, Liu F, Ma X, Chico TJ, Kramer KL, et al. (2014) The heparansulfate editing enzyme Sulf1 plays a novel role in zebrafish VegfA mediated arterial venous identity. Angiogenesis 17: 77-91.
17. Wang YH, Beck C (2015) Distinct patterns of endosulfatase gene expression during Xenopus laevis limb development and regeneration. Regeneration 2: 19-25.
18. Kalus I, Rohn S, Puvirajesinghe TM, Guimond SE, Eyckerman-Kölln PJ, et al. (2015) Sulf1 and Sulf2 Differentially Modulate Heparan Sulfate Proteoglycan Sulfation during Postnatal Cerebellum Development: Evidence for Neuroprotective and Neurite Outgrowth Promoting Functions. PLoS One 10: e0139853.
19. Zaman G, Staines KA, Farquharson C, Newton PT, Dudhia J, et al. (2016) Expression of Sulf1 and Sulf2 in cartilage, bone and endochondral fracture healing. Histochem Cell Biol 145: 67-79.
20. Freeman SD, Keino-Masu K, Masu M, Ladher RK (2015) Expression of the heparansulfate 6-O-endosulfatases, Sulf1 and Sulf2, in the avain and mammalian inner ear suggests a role for sulfatation during inner ear development. Dev Dyn 244: 168-180.
21. Kim JH, Chan C, Elwell C, Singer MS, Dierks T, et al. (2013) Endosulfatases SULF1 and SULF2 limit Chlamydia muridarum infection. Cell Microbiol 15: 1560-1571.
22. Lum DH, Tan J, Rosen SD, Werb Z (2007) Gene trap disruption of the mouse heparan sulfate 6-O-endosulfatase gene, Sulf2. Mol Cell Biol 27: 678-688.
23. Maltseva I, Chan M, Kalus I, Dierks T, Rosen SD (2013) The SULFs, extracellular sulfatases for heparansulfate, promote the migration of corneal epithelial cells during wound repair. PLoS One 8: e69642.
24. Nagamine S, Koike S, Keino-Masu K, Masu M (2005) Expression of a heparin sulphate remodeling enzyme, heparin sulfate 6-O-endosulfatase sulfatase FP2, in the rat nervous system. Brain Res Dev Brain Res 159: 135-143.
25. Uppalapati D, Ohta N, Zhang Y, Kawabata A, Pyle MM, et al. (2011) Identification and characterization of unique tumoricidal genes in rat umbilical cord matrix stem cells. Mol Pharm 8: 1549-1558.
26. Thierry-Mieg D, Thierry-Mieg J (2006) AceView: A comprehensive cDNA-supported gene and transcripts annotation. Genome Biology 7: S12
27. 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.
28. Hur K, Han TS, Jung EJ, Yu J, Lee HJ, et al. (2012) Up-regulated expression of sulfatases (SULF1 and SULF2) as prognostic and metastasis predictive markers in human gastric cancer. J Pathol 228: 88-98.
29. Joy MT, Vrbova G, Dhoot GK, Anderson PN (2015) Sulf1 and Sulf2 expression in the nervous system and its role in limiting neurite outgrowth in vitro. Exp Neurol 263: 150-60.
30. Altschul F, Vyas V, Cornfield A, Goodin S, Ravikumar TS, et al. (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
31. Sievers F, Higgins DG (2014) Clustal omega. Curr Protoc Bioinformatics 48: 3.13.1-3.13.16.
32. 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.
33. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, et al. (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36: 107-113.
34. Kopp J, Schwede T (2004) The SWISS-MODEL repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res 32: D230-D234.
35. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matches. J Mol Biol 292: 195-202.
36. Marchler-Bauer A, Panchenko AR, Ariel N, Bryant SH (2002) Comparison of sequence and structure alignments for protein domains. Proteins 48: 439-446.
37. GTEx Consortium (2015) Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660.
38. 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.
39. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
40. Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol 537: 113-137.
41. 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.
42. Valstar MJ, Bertoli-Avella AM, Wessels MW, Ruijter GJ, de Graaf B, et al. (2010) Mucopolysaccharidosis type IIID: 12 new patients and 15 novel mutations. Hum Mutat 31: E1348-E1360.
43. Higginson JR, Thompson SM, Santos-Silva A, Guimond SE, Turnbull JE, et al. (2012) Differential sulfation remodelling of heparan sulfate by extracellular 6-O-sulfatases regulates fibroblast growth factor-induced boundary formation by glial cells: implications for glial cell transplantation. J Neurosci 32: 15902-15912.
44. Lukatela G, Krauss N, Theis K, Selmer T, Gieselmann V, et al. (1998) Crystal structure of human arylsulfatase A: the aldehyde function and the metal ion at the active site suggest a novel mechanism for sulfate ester hydrolysis. Biochemistry 37: 3654-3664.

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

46. Hernandez-Guzman FG, Higashiyama T, Pangborn W, Osawa Y, Ghosh D (2003) Structure of human estrone sulfatase suggests functional roles of membrane association. J Biol Chem 278: 22989-22997.

47. Sidhu NS, Schreiber K, Pröpper K, Becker S, Usón I, et al. (2014) Structure of sulfamidase provides insight into the molecular pathology of mucopolysaccharidosis IIIA. Acta Crystallogr D Biol Crystallogr 70: 1321-1335.

48. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, et al. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132: 958-970.

49. Bamforth SD, Bragança J, Eloranta JJ, Murdoch JN, Marques FI, et al. (2001) Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. Nat Genet 29: 469-474.

50. Hu CT, Chang TY, Cheng CC, Liu CS, Wu JR, et al. (2010) Snail associates with EGR-1 and SP-1 to upregulate transcriptional activation of p15INK4b. FEBS J 277: 1202-1218.

51. Stein B, Yang MX (1995) Repression of the interleukin-6 promoter by estrogen receptor is mediated by NF-kappa B and C/EBP beta. Mol Cell Biol 15: 4971-4979.

52. Eeckhoute J, Formstecher P, Laine B (2004) Hepatocyte nuclear factor 4alpha enhances the hepatocyte nuclear factor 1alpha-mediated activation of transcription. Nucleic Acids Res 32: 2586-2593.

53. 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.

54. Han J, Kim HJ, Schafer ST, Paquola A, Clemenson GD, et al. (2016) Functional Implications of miR-19 in the Migration of Newborn Neurons in the Adult Brain. Neuron 91: 79-89.

55. Yi Y, Li H, Lv Q, Wu K, Zhang W, et al. (2016) miR-202 inhibits the progression of human cervical cancer through inhibition of cyclin D1. Oncotarget 7: 72067-72075.

56. Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3: e314.