Article

Gene Structure Evolution of the Short-Chain Dehydrogenase/Reductase (SDR) Family

Franco Gabrielli [1], Marco Antinucci [2] and Sergio Tofanelli [1,*]

[1] Department of Biology, University of Pisa, Via Ghini, 13-56126 Pisa, Italy
[2] Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, 08002 Barcelona, Spain
* Correspondence: sergio.tofanelli@unipi.it

Abstract: SDR (Short-chain Dehydrogenases/Reductases) are one of the oldest and heterogeneous superfamily of proteins, whose classification is problematic because of the low percent identity, even within families. To get clearer insights into SDR molecular evolution, we explored the splicing site organization of the 75 human *SDR* genes across their vertebrate and invertebrate orthologs. We found anomalous gene structures in members of the human SDR7C and SDR42E families that provide clues of retrogene properties and independent evolutionary trajectories from a common invertebrate ancestor. The same analyses revealed that the identity value between human and invertebrate non-allelic variants is not necessarily associated with the homologous gene structure. Accordingly, a revision of the SDR nomenclature is proposed by including the human *SDR40C1* and *SDR7C* gene in the same family.

Keywords: SDR-family; splicing sites; orthologous genes; genetic homology; retrogenes

1. Introduction

Short-chain dehydrogenase/reductase (SDR) is a superfamily of NAD(P)-dependent oxidoreductases and related enzymes, present in all living organisms [1–3]). For almost all members of the family, the structure of the gene, as well as the tertiary structure and the catalytic activity of the protein, have been defined. Despite their low sequence identity (typically from 10% to 30% of conserved sequence identity among families), SDR monomers have very similar tertiary structures and two specific sequences, one responsible for the specific binding of the NAD(P) coenzyme and the other including amino acids directly involved in the catalysis of various SDR substrates.

Among SDR enzymes, the divergence of the monomers increases toward their C-terminals, where the amino acid responsible for the substrate binding-site are localized [1–3]. Sequence variability at the SDR substrate-binding site is associated with a large spectrum of substrates, resulting in unique active sites and substrate specificities [2]. Concerning SDR differences and functions, it is worth mentioning their emerging role in human diseases. Indeed, knowledge has accumulated concerning the role that SNP variants in specific human *SDR* genes plays in a wide variety of cancers [4–7] and metabolic diseases [8–12]. The pattern of low sequence/high structural similarity suggests a fusion of domains specific of a common coenzyme and a wide spectrum of substrates [13]. It has also been argued that the alcohol dehydrogenase (ADH) in *Drosophila melanogaster* model organism is the ancestral "prototype" of the mammalian SDR superfamily and was responsible for the radiation of fruit fly species worldwide some 65 Mya [14]. Since then, many new data have accumulated in the genetic databases. Persson et al. [3], with an accurate analysis of protein sequences in all living species, identified 75 human SDR non-allelic variants classified, upon a clustering approach based on Hidden Markov Models (HMMs), into 48 families out of a total of 200, each having from one to eight members.

The low percent identity upon which *SDR* genes classification is based and the uncertainty about the relationships between structure and function across species make urgent a more comprehensive evolutionary analysis of *SDR* genes.

Splicing site organizations are generally conserved over very long evolutionary distances, as they demonstrated to retain information about gene homology better than other molecular traits, especially when genes belong to a large family [15,16]. Accordingly, splicing sites organization can be useful to identify cryptic homologous variants of *SDR* genes belonging either to the same or to different species [17,18]. In the present paper, we annotated the splicing site organization and the splicing site phase number of each member of the human SDR families, as well as the splicing site organization of their vertebrate and invertebrate orthologs, to achieve a more comprehensive nomenclature.

2. Methods

Computational tools, *SDR* gene and protein data were obtained from public primary databases: European Bioinformatics Institute (EBI), UK; the Genome Browser of the University of California, USA; National Center for Biotechnology Information (NCBI), Maryland, USA; ExPASy, Swiss Institute of Bioinformatics (SIB), Switzerland.

Vertebrate and invertebrate orthologs of human SDR variants were detected using BLASTp (release 2.13.0, NCBI). Protein sequences alignments and their identity values were obtained using Clustal Ω (release 1.2.2, EBI). Exon-intron organizations, intron physical positions, phase numbers, exon-intron boundary sequences and observance of the AG-GT/GC rule were manually checked upon gene sequence/cDNA alignments. The intron phase denotes the position of the intron within a codon: Phase 0, 1 or 2 depending whether it starts before the first base, after the first base, and after the second base, respectively. Any multiple polypeptide alignment was verified by pairwise comparison across vertebrate species; we annotated as orthologs the splicing sites having an identical phase number and identical sequence position. We used the same symbols adopted for human SDR variants by Persson et al. [3]. Symbols include SDR followed by the annotation number and a single letter: A (Atypical), C (Classical), E (Extended), U (Unknown). C and E denote the two major types of SDR family enzymes [3].

We identified the invertebrate orthologs of human SDR variants by the following procedures. First, we detected, by BLAST, the invertebrate protein having the highest sequence similarity with the human SDR protein variant. Then, we verified that the aminoacidic sequences included the *SDR* structure consensus: TGxxxGxG or TGxxGxxG and the catalytic consensus YxxxK, which are diagnostic of SDR superfamily variants [1–3]. SDR cofactor binding site and the catalysis active site consensus of the human SDR family are reported in the supplementary material (Online Resources 1: Tables S1 and S2). Lastly, we verified, in the invertebrate SDR variants, the presence of splicing site orthologs of those for the human variant used as BLAST probe.

Phylogenetic trees were constructed using the binarized matrix of splicing-site phases with the Wagner parsimony method, as implemented in the PARS algorithm of the software package PHYLIP version 3.6 [19]. We performed bootstrap analysis with 10,000 replications to estimate the strength of support for each clade. The same tree topologies were obtained by a Bayesian approach by BEAST 2.7.0.0 [20].

3. Results and Discussion

3.1. Gene Structure of Vertebrate and Invertebrate Variants of SDR Families

The human SDR families, classified by the relative identity values of their protein variants [3], may include members having either identical (SDR7C1 and SDR7C2), similar (SDR7C1 and SDR7C3), or completely different (SDR7C1, SDR7C4 and SDR7C5) splicing patterns (Figure 1; Tables 1–3; Online Resources 2: Figures S1–S12, Tables S3–S14; Online Resources 3: Figures S13–S24, Tables S15–S26).

```
SDR7C1    ----------------MVELMFPLLLLLLPFLLYMAAPQIRKMLSSGVCTSTVQLPGKVV    1p.1
SDR7C2    -------------------MLVTLGLLTSFFSFLYMVAPSIRKFFAGGVCRTNVQLPGKVV    1p.1
SDR7C3    ------------------MSRYLLPLSALGTVAGAAVLLKDYVTGGACPSKATIPGKTV    1p.1
SDR7C4    ------------MAVATAAAVLAALGGALWLAARRFVGPRVQRLRRGGDPGLMHGKTV
SDR7C5    --------------------MEALLLGAGLLLGAYVLVYYNLVKAPPCGGMGNLRGRTA
SDR40C1   ----------------MSLYRSVVWFAKGLREYTKSGYESACKDFVPHDLEVQIPGRVF    1p.1
                                                                  *

                   x    x x
SDR7C1    VVTGANTGIGKETAKELAQRGARVYLACRDVEKGELVAKEIQTT-------------TGN    2p.2
SDR7C2    VITGANTGIGKETARELASRGARVYIACRDVLKGESAASEIRVD-------------TKN    2p.2
SDR7C3    IVTGANTGIGKQTALELARRGGNIILACRDMEKCEAAAKDIRGE-------------TLN    2p.2
SDR7C4    LITGANSGLGRATAAELLRLGARVIMGCRDRARAEEAAGQLRRELRQAAECGPEPGVSGV
SDR7C5    VVTGANSGIGKMTALELARRGARVVLACRSQERGEAAAFDLRQE-------------SGN    1p.2 2p.0
SDR40C1   LVTGGNSGIGKATALEIAKRGGTVHLVCRDQAPAEDARGEIIRE-------------SGN    2p.2
                 **  *  *  *   **  *     *        **      *

SDR7C1    QQVLVRKLDLSDTKSIRAFAKGFLAEEKHLHVLINNAGVMMCPYSKTADGFEMHIGVNHL    3p.2        4p.2
SDR7C2    SQVLVRKLDLSDTKSIRAFAEGFLAEEKQLHILINNAGVMMCPYSKTADGFETHLGVNHL    3p.2        4p.2
SDR7C3    HHVNARHLDLASLKSIREFAAKIIEEEERVDILINNAGVMRCPHWTTEDGFEMQFGVNHL    3p.2        4p.2
SDR7C4    GELIVRELDLASLRSVRAFCQEMLQEEPRLDVLINNAGIFQCPYMKTEDGFEMQFGVNHL    1p.0
SDR7C5    NEVIFMALDLASLASVRAFATAFLSSEPRLDILIHNAGISSCG--RTREAFNLLLRVNHI    3p.2
SDR40C1   QNIFLHIVDLSDPKQIWKFVENFKQ-EHKLHVLINNAGCMVNKRELTEDGLEKNFAANTL    3p.0 4p.0 5p.2
                 **        *        *     **  ***        *           *

                                                            +    +
SDR7C1    GHFLLTHLLLEKLKESAPSRIVNVSSLAHHLGRIHFHNLQG-EKFYNA-GLAYCHSKLAN
SDR7C2    GHFLLTYLLLERLKVSAPARVVNVSSVAHHIGKIPFHDLQS-EKRYSR-GFAYCHSKLAN
SDR7C3    GHFLLTNLLLDKLKASAPSRIINLSSLAHVAGHIDFDDLNWQTRKYNT-KAAYCQSKLAI
SDR7C4    GHFLLTNLLLGLLKSSAPSRIVVVSSKLYKYGDINFDDLNS-EQSYNK-SFCYSRSKLAN
SDR7C5    GPFLLTHLLLPCLKACAPSRVVVVASAAHCRGRLDFKRLDRPVVGWRQELRAYADTKLAN
SDR40C1   GVYILTTGLIPVLEKEHDPRVITVSSGGMLVQKLNTNDLQSERTPFDG-TMVYAQNKRQQ    6p.0 7p.0
          *    **   *     *        *        *              *        *   *

SDR7C1    ILFTQELARRLKGSGVTTYSVHPGTVQSELVRHSSFMRWM------WWLFSFFIKTPQQG    5p.2
SDR7C2    VLFTRELAKRLQGTGVTTYAVHPGVVRSELVRHSSLLCLL------WRLFSPFVKTAREG    5p.2
SDR7C3    VLFTKELSRRLQGSGVTVNALHPGVARTELGRHTGIHGSTFSSTTLGPIFWLLVKSPELA    5p.2 6p.2
SDR7C4    ILFTRELARRLEGTNVTVNVLHPGIVRTNLGRHIHIPLLV--KPLFNLVSWAFFKTPVEG
SDR7C5    VLFARELANQLEATGVTCYAAHPGPVNSELFL-RHVPGWL--RPLLRPLAWLVLRAPRGG    4p.2
SDR40C1   VVLTERWAQGHP--AIHFSSMHPGWADTPGVRQ--------AMPGFHARFGDRLRSEAQG    8p.2
                            ***

SDR7C1    AQTSLHCALTEG-LEILSGNHFSDCHVAWVS-AQARN---ETIARRLWDVSCDLLGLPID    6p.1
SDR7C2    AQTSLHCALAEG-LEPLSGKYFSDCKRTWVS-PRARN---NKTAERLWNVSCELLGIRWE    6p.1
SDR7C3    AQPSTYLAVAEE-LADVSGKYFDGLKQKAPA-PEAED---EEVARRLWAESARLVGLEAP
SDR7C4    AQTSIYLASSPE-VEGVSGRYFGDCKEEELL-PKAMD---ESVARKLWDISEVMVGLLK-
SDR7C5    AQTPLYCALQEG-IEPLSGRYFANCHVEEVP-PAARD---DRAAHRLWEASKRLAGLGPG
SDR40C1   ADTMLWLALSSAAAAQPSGRFFQDRKPVSTHLPLATASSSPAEEEKLIEILEQLAQTFK-    9p.2
          *        *        **   *              *           *

SDR7C1    ------------------------------------------------------------
SDR7C2    ------------------------------------------------------------
SDR7C3    SVREQPLPR---------------------------------------------------
SDR7C4    ------------------------------------------------------------
SDR7C5    EDAEPDEDPQSEDSEAPSSLSTPHPEEPTVSQPYPSPQSSPDLSKMTHRIQAKVEPEIQLS
SDR40C1   ------------------------------------------------------------
```

Figure 1. Alignment of the human SDR7C family and SDR40C1 protein variants. × and + symbols mark the structure consensus and the catalysis consensus respectively. The couples of amino acid symbols in red mark the splicing-site positions. Splicing sites are progressively numbered, and phase (p.) type is indicated after the splicing-site number. * symbol marks the position of identical amino acid residues of the aligned SDR protein sequences.

Table 1. Genetic and molecular data of the human SDR7C family and human SDR40C1 protein variants. Chr, chromosome; the variant phase formula includes phase type symbols aligned according to the sequence of their relative splicing sites; aa n., number of the variant amino acids; standard amino acids of the consensuses are in red.

Family Symbol and Name	Enzyme Symbol	Gene Symbol	Gene ID	Chr	Exon Number	Phase Formula	aa n.	Structure Consensus	Catalysis Consensus
SDR7C	SDR7C1	RDH11	51109	14	7	122221	318	GANTGIG	YCHSK
	SDR7C2	RDH12	145226				316		
	SDR7C3	RDH13	112724	19	7	122222	331		
	SDR7C4	RDH14	57665	12	2	0	336	GANSGLG	YSRSK
Retinol dehydrogenase	SDR7C5	DHRS13	147015	17	5	2022	377	GANSGIG	YADTK
SDR40C Dehydrogenase/ reductase SDR family	SDR40C1	DHRS12	79758	13	10	120020022	317	GGNSGIG	YAQNK

Table 2. The percent identity of the human SDR7C family and human SDR40C1 protein variants.

	% Identity				
	SDR7C1	**SDR7C2**	**SDR7C3**	**SDR7C4**	**SDR7C5**
SDR7C2	71.66				
SDR7C3	49.68	48.87			
SDR7C4	46.15	46.47	48.88		
SDR7C5	45.78	46.41	42.32	44.48	
SDR40C1	32.54	33.22	30.64	31.21	28.23

Variants of seven human SDR families retain an active site made of an insertion of two amino acids: Asparagine (N) and Serine (S). The four aminoacids (N-S-Y-K) are called the catalytic tetrad [3]. We found that the human SDR families with the catalytic tetrad are SDR9C, SDR12C, SDR16C, SDR25C, SDR26C SDR28C, SDR32C (respectively Figures S2, S5, S6 and S8–S11 in Online Resources 2). The members of these human families have invertebrate orthologues carrying identical catalytic tetrads (Online Resources 4: Figures S26–S36), suggesting a pre-vertebrate acquisition of these sites.

Table 3. Splicing-site organization of the human SDR7C family and SDR40C1 protein variants and their respective invertebrate orthologs. Phase symbols in a same column and highlighted in green or pink, mark the orthologous splicing sites. For other details, see Table 1.

Species	Variants	Phase Formula							Splicing-Site Phases						
Homo sapiens	**SDR7C1**	122221		1	2			2		2		2				1
	SDR7C2	122221		1	2			2		2		2				1
	SDR7C3	122222		1	2			2		2		2			2	
	SDR7C4	0						0								
	SDR7C5	2022			2		0			2		2				
	SDR40C1	120020022	1		2		0		0	2		0	0		2	2
Ciona intestinalis	SDR7C-1C1	22212			2	2				2		1			2	
	SDR7C-1C2	02021					0			2		0			2	1
	SDR7C-3C2	100022	1					0				0			2	2
	SDR7C-3C3	202021			2				0	2		0			2	1
	SDR7C-3C4	02021						0		2		0			2	1
Strongylocentrotus purpuratus	SDR7C-1C1	2221				2				2					2	1
	SDR7C-1C3	122221		1	2			2		2					2	1
	SDR7C-2C2	222			2					2		2				
Musca domestica	SDR7C-1C3	10102		1			0		1			0			2	
Brugia malayi	SDR7C-1C1	2022001				2		0		2		2		0	0	1
Aplysia californica	SDR7C-1C2	212222			2		1				2	2			2	2
	SDR7C-1C3	122221	1		2					2		2			2	1

Splicing site organizations and splicing phase numbers of human SDR7C (Online Resources 3: Figure S13, Table S15) and SDR21C (Online Resources 3: Figure S21, Table S23) variants are identical to those of their respective orthologs in each vertebrate class. Moreover, the other SDR families present a highly conserved splicing site organization, which is identical from fishes to humans. However, a species may carry a gene, belonging to a given family, which differs with shorter or longer exons, and/or lack one splicing site, and/or bears one or two extra splicing sites with respect to its human ortholog (Online Resources 3: Figures S13–S18 and S20–S24, Tables S15–S20 and S22–S26). Additionally, invertebrate SDR variants may have either identical, similar, or completely different gene structure with respect to the human SDR orthologs, identified by the sequence similarity of their codified polypeptides (Online Resources 4: Figures S25–S36, Tables S27–S50). Among invertebrate proteins, we only detected a polypeptide variant (in the sea urchin *S. purpuratus*), which has an identical gene structure with human SDR11E1 and SDR11E2 variants (Online Resources 4: Table S42), whereas variants of the other human SDR families show splicing patterns differing for the length of one or more exons, and/or by a different number of splicing sites (Online Resources 4: Figures S25–S36, Tables S27–S50).

The time of acquisition of the vertebrate gene structure can be inferred by the evolutionary position of *C. intestinalis*, a deuterostome considered a good approximation of the ancestral chordates living about 540 Mya. After the formation of the phylum, splicing patterns were conserved from fishes up to humans while genetic and protein sequences diverged (Online Resources 3: Figures S13–S24, Tables S15–S26; Online Resources 5: Figures S37–S48).

Conversely, four human variants, SDR7C4, SDR42E1 (Online Resources 4: Tables S39 and S50), SDR21C, and SDR21C2 (Online Resources 4: Table S45) have high identity values and completely different splicing patterns when compared to their respective invertebrate orthologs. These data show that, in SDR families, protein sequences and gene structures are phylogenetically uncoupled.

Interestingly, *Saccharomyces cerevisiae* genome has few introns, usually limited to one per gene [21]. We analyzed *S. cerevisiae* proteins having structure and catalysis consensuses diagnostic of the SDR family, but none of them had introns. Even though little evidence supports increased intron loss for paralogous gene families in plant and animal evolution [22], specific surveys on yeast suggest increased intron-loss over intron-gain events, with genes involved in metabolism, molecular transportation and enzyme activity regulation being more prone to introns loss [23]. Indeed, a recent study conducted on 263 fungal species, highlighted how the major evolutionary trend for intron changes in this kingdom, involves the loss of such sequences [24].

It is important to bear in mind, though, that analyses of Zrt-, Irt-like protein (ZIP) gene family, deemed as ancestral genes related to prion gene family evolution, revealed how intron conservation can be high when considering their relative positions in comparison of multiple sequences [25]. On the other hand, intron length conservation seems more diluted over evolutionary time scales, resulting in a heterogeneous set of intronic sequences length in ZIP genes [25].

3.2. Gene Structure of Human SDR7C and SDR42E Family Variants

Human SDR7C family has five non-allelic variants (Table 1; Online Resources 1: Figure S1). SDR7C1 and SDR7C2 variants have six splicing sites and an identical gene structure (Table 3). SDR7C3 variant has a very similar gene structure but only five splicing sites out of six are homologous to those of SDR7C1-2 variants (Table 3). SDR7C4 variant has only one splicing site, which has the same protein sequence position of the third splicing sites of SDR7C1, SDR7C2 and SDR7C3 variants, but a different phase number (Table 3). The SDR7C5 variant has four splicing sites and a gene structure completely different from those of all other human SDR7C family members (Table 3).

We found that human SDR40C1 variant, which is not included in the SDR7C family [3], has two splicing sites homologous to those of SDR7C1, SDR7C2 and SDR7C3 variants

(Table 3), despite relatively low levels of identity with the other human SDR7C variants (26.3–31%, Table 4). Presumably, the low sequence homology prevented a correct assignment of the SDR40C1 variant into the SDR7C family by HMM-based clustering models [3]. Conversely, parsimony analyses based on the splicing-phase structure (Online resources 5: Figure S37) cluster the SDR40C1 variant within vertebrate and invertebrate orthologs of the SDR7C family with bootstrap support up to 92%. Proteins belonging to different superfamilies, which may have identities as low as 4% and whose homology can only be inferred from a similar 3D structure and function, may share splicing patterns [26].

Table 4. Percent identity values of the human SDR7C family and SDR40C1 protein variants and of their invertebrate orthologs.

Species	Variants	% Identity					
		SDR7C1	SDR7C2	SDR7C3	SDR7C4	SDR7C5	SDR40C1
Homo sapiens	SDR7C2	71.66					
	SDR7C3	49.20	48.89				
	SDR7C4	45.31	46.75	48.72			
	SDR7C5H	45.87	47.21	42.01	45.39		
	SDR40C1	30.20	31.00	28.05	28.72	26.33	
Ciona intestinalis	SDR7C-1C1	37.26	38.78	34.46	36.31	35.62	26.40
	SDR7C-1C2	46.05	48.11	47.68	46.08	36.79	24.64
	SDR7C-3C2	26.00	27.48	25.25	26.51	25.83	54.57
	SDR7C-3C3	36.69	35.48	33.02	32.57	34.50	25.33
	SDR7C-3C4	46.25	46.93	44.73	46.91	39.54	23.67
Strongylocentrotus purpuratus	SDR7C-1C1	54.12	53.76	53.36	50.53	44.04	30.71
	SDR7C-1C3	47.17	47.78	59.09	48.90	40.30	29.61
	SDR7C-2C2	42.49	42.77	42.86	42.44	45.02	27.63
Musca domestica	SDR7C-1C3	49.37	51.91	53.92	47.78	44.52	31.25
Brugia malayi	SDR7C-1C1	36.77	39.48	38.54	41.40	33.66	27.67
Aplysia californica	SDR7C-1C2	39.81	41.04	41.38	36.86	33.13	29.24
	SDR7C-1C3	49.05	46.47	52.05	47.34	40.26	28.57

3.3. Invertebrate Orthologs of Human SDR Family Variants

The splicing patterns of the human SDR7C family is the same in all vertebrate orthologs. Human SDR7C1, SDR7C2 and SDR7C3 variants have orthologs in several invertebrate species. *S. purpuratus* SDR7C-1C3 variant has a splicing phase identical to that of human SDR7C1 and SDR7C2 variants (Table 3). Therefore, it can be confidently considered the ancestral form. Human SDR7C4 variant has a single splicing site with zero phase number, the same protein sequence position of one splicing site of *S. purpuratus* SDR7C-1C3, *B. malayi* SDR7C-1C1 and *A. californica* SDR7C-1C3 variants. However, the invertebrate splicing sites have a different phase number (Table 3). Human SDR7C5 variant has diagnostic splicing sites and phases completely different from those of all other human SDR7C family members but shares orthologous splicing sites with invertebrate SDR7C variants. Some variants (SDR7C-1C1 and SDR7C-3C3 in *C. intestinalis*, SDR7C-2C2 in *S. purpuratus*, SDR7C-1C2 in *A. californica*) have orthologous splicing sites only with human SDR7C5 variant (Table 3), while others (SDR7C-1C2 and SDR7C-3C4 in *C. intestinalis*, SDR7C-1C1 in *S. purpuratus*) have orthologous splicing sites with human SDR7C5 as well as with human SDR7C1, SDR7C2 and SDR7C3 variants (Table 3, Online resources 5: Figure S37). We speculate that human SDR7C1, SDR7C2, SDR7C3 and SDR7C5 variants originated from

a common ancestral gene which differentiated early in invertebrate evolutionary history (Online Resources 4: Figure S25).

Four splicing sites, diagnostic of human SDR40C1 variant, have orthologous splicing sites in *C. intestinalis* SDR7C-3C2 variant, which has orthologous splicing sites only with human SDR40C1 variant (Table 3). However, *S. purpuratus* SDR7C-1C3 and *A. californica* SDR7C-1C3 variants have orthologous splicing sites with human SDR40C1 as well as SDR7C1, SDR7C2 and SDR7C3 variants (Table 3). Human SDR40C1 variant shares with *C. intestinalis* SDR7C-3C2 variant a 54.57% of the sequence and below 30% with the other invertebrate orthologous variants (Table 4). Parsimony analyses confirm they are closely related (Online resources 5: Figure S37). Thus, we assume that *C. intestinalis* SDR7C-3C2 is the closest variant to the most recent common ancestor of the human SDR40C1 variant. This gene may have evolved from an invertebrate gene, an ancestor of the human gene clade SDR40C1-SDR7C1-SDR7C2-SDR7C3. These data support our hypothesis that the SDR40C1 variant is a member of the SDR7C family.

Human SDR42E family has only two non-allelic variants: SDR42E1 and SDR42E2 (Table 5), which have one and ten splicing sites, respectively (Table 6). However, SDR42E1 and SDR42E2 variants have relatively high sequence similarity (Figure 2, Table 7), with consensus regions spread in distinct blocks throughout the polypeptide molecule (Online Resources 2, Figure S12), and their gene structures are identical to those of all their respective vertebrate orthologs (Online Resources 3, Figure S24). Only the human SDR42E2 variant has orthologous splicing sites with the invertebrate SDR variants (Table 6). In particular, the ten human SDR42E2 splicing sites have orthologs in several invertebrate variants and all the ten splicing sites in *C. intestinalis* SDR42E-1E1 variant (Table 6). Such splicing pattern is also phylogenetically supported (Online resources 5: Figure S48). Thus, the *C. intestinalis* SDR42E-1E1 gene may be the closest proxy of the more recent common ancestor of invertebrate and human SDR42E variants.

Table 5. Genetic and molecular data of the human SDR42E family variants. For further details, see Table 1.

Family Symbol and Name	Enzyme Symbol	Gene Symbol	Gene ID	Chr	Exons	Phase Formula	aa n.	% Identity	Structure Consensus	Catalysis Consensus
SDR42E 3-β-HSD family	SDR42E1	SDR42E1	93517	16	2	1	393	47.18	GGSGYFG	YSRTK
	SDR42E2	SDR42E2	100288072		12	20020200020	626		GGGGYLG	

Table 6. Splicing-site organization of the human SDR42E family variants and their invertebrate orthologs. For further details see Table 1.

Species	Variants	Phase Formula	Splicing-Site Phases											
H. sapiens	**SDR42E1**	**1**	1											
	SDR42E2	0020200020		0	0	2	0	2	0	0	0	2	0	
C. intestinalis	SDR42-1E1	100202000202	1	0	0	2	0	2	0	0	0	2	0	2
S. purpuratus	SDR42-1E1	02020020			0	2	0	2	0	0		2	0	
Caenorhabditis elegans	SDR42-1E1	20000	2	0	0		0		0					0
	SDR42-2E1	000000		0	0		0		0		0			0
Caenorhabditis remanei	SDR42-1E2	00000			0		0		0		0			0
A. californica	SDR42-1E1	2000020		2	0		0	0			0	2	0	

```
SDR42E1    --------------------------------------------------------
SDR42E2    MKSNPPRSSLEACKAAGQGEKSCPVCQACGEVSGPRSGSGSESRPAPKPGAIPGPGLGPK


SDR42E1    --------------------------------------------------------
SDR42E2    AIPGPQAGSGTVPRPGAISGTGPGLGPGPGAGSVPGPGAGSVPGLGARSVPGPGAGSVPG


SDR42E1    --------------------------------------------------------
SDR42E2    PGAGSVPGPGAGSVPGPGAGSGPGLGGGLGPGVGAGPGAGSVPGPGAGSVPGPGAGSVPG


SDR42E1    --------------------------------------------------MDPKRSQKESV
SDR42E2    PGAGSVPGAGAGSTPEPELGPGLRQGSGTGPRPSESTTTPTPAPQQKTQAKPTKAARQKV
                                                             *        *

                       ×   ×   ×
SDR42E1    LITGGSGYFGFr1GCALNQNGVHVILFDISSPAQTIPEGIKFIQGDIRHLSDVEKAFQDA    1p.1
SDR42E2    LVTGGGGYLGFSLGSHLAKSGTSVILLDRRRPQWELSPETKFIqaDVRDEEALYRAFE--    1p.2
            * *** ** ** **   *    *   *** *    *    **** * *      **

SDR42E1    DVTCVFHIASYGMSGREQLNRNLIKEVNVRGTDNILQVCQRRRVPRLVYTSTFNVIFGGQ    3p.0  4p.2
SDR42E2    GVDCVFHVASYGMSGAEk1QKEQIESINVGGTKLVIdvCVRRRVPRLIYTSTVNVAFGGK
             * **** ******* * *    *    ** **    ** ******* **** ** ***

                         +       +
SDR42E1    VIRNGD-ESLPYLPLHLHPDHYSRTKSIAEQKVLEANATPLDRGDGVLRTCALRPAGIYG    5p.0  6p.2
SDR42E2    PIEQGDEDSVPYFPLDehVDHYSRTKAIADQLTLMANGMPL-pgGGTLRTCVLRPPGIYG
             *   **   * **  **   * ******* **  *   * **   * * **** *** ****

SDR42E1    PGEQRHLPRIVSYIEKGLFKFVYGDPRSLVEFVHVDNLVQAHILASEALRADKGHIASGQ    7p.0  8p.0
SDR42E2    PEEQRHLPRVagHIKKRLFMFRFGDHKARMNWVHVHNLVQAHVLAAEALTTAKGYVasGQ
            * *******   * * *   **    *** ****** ** ***    ** ****

SDR42E1    PYFISDGRPVNNFEFFRPLVEGLGYTFPSTRLPLTLVYCFAFLTEMVHFILGRLYNFQPF    9p.0 10p.2
SDR42E2    AYYINDGESVNLFEWMAP1fEKLGYSQPWIQVPTSWVYLtaAVMERLHLALRPICSLPPL
            * * **  ** **    ** * ***   *      **  *    *   *   *     *

SDR42E1    LTRTEVYKTGVTHYFSLEKAKKELGYKAQPFDLQEAVEWFKAHGHGRSSGSRDSECFVW-    11p.0
SDR42E2    LTRSevRSVAVTHTFQIAKARAQLGYAPDKFRFADAVELYVQSTTRRPRGSTARTLLRLL
            *** **   *** *   **   ***    *   ***      *   * **

SDR42E1    DGLLVFLLIIAVLM-WLPSSVILS----L
SDR42E2    LRLLLFLGLLALALHFLGLQPLHAAVERL
            ** **    *    *           *
```

Figure 2. Alignment of the human SDR42E protein variants. × and + symbols mark the structure consensus and the catalysis consensus respectively. For further details, see Figure 1.

Table 7. Percent identity values of the human SDR42E family variants and of their invertebrate orthologs.

Species	Variants	% Identity	
		SDR42E1	SDR42E2
H sapiens	SDR42E2	47.18	
C. intestinalis	SDR42-1E1	41.58	40.69
S. purpuratus	SDR42-1E1	49.55	43.88
C. elegans	SDR42-1E1	27.01	27.54
	SDR42-2E1	24.72	29.46
C. remanei	SDR42-1E2	28.33	31.58
A. californica	SDR42-1E1	48.97	39.69

3.4. Human SDR7C4 and SDR42E1 Are Possibly Active Retrogenes

We speculate that human *SDR7C4* and *SDR42E1* genes are active retrogenes. Retrogenes are generated from processed mRNA, do not have splicing sites and are not transcribed. However, a certain number of retrogenes are functionally active [27,28]. They

may have inherited the promoter bound to the coding sequence of the parental gene or accidentally acquired a new promoter from another gene when the retrogene sequence is inserted in the DNA of a given chromosome. To be heritable, the retrotransposition needs to occur in a germline or during early embryonic stages [28]. Many active retrogenes have been discovered in mammals, and many of them developed their functional role in the germ line. After retrotransposition, active retrogenes may acquire new introns [29].

Human *SDR7C4* gene has the basic characteristics of an active retrogene. It is highly expressed in the testis (NCBI, gene ID: 57665), its chromosome localization is different from any other human SDR7C variant (Online Resources 2, Table S3) and it has a single splicing site, which does not have orthologs with identical phase number in the invertebrate variants (unlike other SDR7C variants). Moreover, the human SDR7C4 variant has high identity values with several invertebrate variants, orthologs of other human SDR7C variants (Table 4). We interpreted these data assuming that the human SDR7C4 variant was formed by retroposition of a metazoan ancestor gene before the chordate radiation.

Likewise, the human *SDR42E1* gene has the basic characteristics of an active retrogene. *SDR42E1* gene is highly synthetized in human testis, (NCBI, gene ID: 93517) and is localized on a chromosome different from those of *SDR42E2* in vertebrate species except Catarrhines, where *SDR42E1* and *SDR42E2* are localized on the same chromosome. *SDR42E1* has a single splicing site that is shared across all vertebrate species (Online Resources 3, Table S26a) but does not have orthologs in the analyzed invertebrate variants (Table 6). Moreover, the human SDR42E1 variant shares slightly higher percent identity values with the invertebrate SDR42E2 ortholog variants than those of the human SDR42E2 variant (Table 7). These data suggest that human SDR42E1 variant has been generated by retrotransposition of an invertebrate ortholog of human SDR42E2 variant before the formation of the chordate phylum.

We could not find reported data about *SDR42E1* and *SDR42E2* chromosomal localizations in fish, amphibian, and reptilian species. However, vertebrate SDR42E1 and SDR42E2 variants have different chromosomal localization from birds up to early primate species (*Callithrix jacchus* and *Microcebus murinus*) and the same chromosomal localization in Catarrhini (Online Resources 3, Table S26a; Online Resources 6, Figures S49 and S50, Table S51a). We interpreted these data assuming that, during evolution, a genomic rearrangement brought *SDR42E1* and *SDR42E2* genes on the same chromosome. This hypothesis is supported by the observation that, in Catarrhines, *SDR42E1* and *SDR42E2* gene loci have identical relative positions and comparable distances whereas SDR42E1 and SDR42E2 proteins have the same molecular characteristics of those of early primates and other vertebrate species which have *SDR42E1* and *SDR42E2* genes localized on different chromosomes (Online Resources 3, Table S26; Online Resources 6, Figures S49 and S50, Table S51a,b). These data suggest that human *SDR42E1* is an active retrogene and not a duplicated form of *SDR42E2* gene, thus adding a new case-study to stress the importance of retroposition in gene evolution [30–34].

4. Conclusions

A deeper insight into the molecular evolution of *SDR* gene families allowed us to resolve classification schemes and evolutionary patterns, regardless of the low sequence homology. The sequences of one member of the human *SDR7C* and *SDR42E* gene families retain traces of a very deep divergence time, at the root of chordate clade. The human *SDR7C4* and *SDR42E1* genes show the properties of an active retrogene, while the human *SDR40C1* gene shows a conservative splicing formula which suggests its inclusion in the same protein family of the SDRC7 variants.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14010110/s1, Figure S1: Alignment of the human SDR7C family and of SDR40C1 protein variants; Figure S2: Alignment of the human SDR9C family protein variants; Figure S3: Alignment of the human SDR10E protein variants; Figure S4: Alignment of the human SDR11E protein variants; Figure S5: Alignment of the human the SDR12C1 protein

variants; Figure S6: Alignment of the human SDR16C protein variants; Figure S7: Alignment of the human SDR21C protein variants; Figure S8: Alignment of the human SDR25C protein variants; Figure S9: Alignment of the human SDR26C1 protein variants; Figure S10: Alignment of the human SDR28C protein variants; Figure S11: Alignment of the human SDR32C protein variants; Figure S12: Alignment of the human SDR42E protein variants; Figure S13: Alignment of the vertebrate SDR7C family and of SDR40C1 variants; Figure S14: Alignment of the vertebrate SDR9C family variants; Figure S15: Alignment of the vertebrate SDR10E family variants; Figure S16: Alignment of the vertebrate SDR11E family variants; Figure S17: Alignment of the vertebrate SDR12C family variants; Figure S18: Alignment of the vertebrate SDR16C family variants; Figure S19: Alignment of the vertebrate SDR21C family variants; Figure S20: Alignment of the vertebrate SDR25C family variants; Figure S21: Alignment of the vertebrate SDR26C family variants; Figure S22: Alignment of the vertebrate SDR28C family variants; Figure S23: Alignment of the vertebrate SDR32C family variants; Figure S24: Alignment of the vertebrate SDR42E family variants; Figure S25: Alignment of the invertebrate orthologs of the human SDR7C family and of the human SDR40C1 variants; Figure S26: Alignment of the invertebrate orthologs of the human SDR9C family variants; Figure S27: Alignment of the invertebrate orthologs of the human SDR10E1 family variants; Figure S28: Alignment of the invertebrate orthologs of the human SDR11E family variants; Figure S29: Alignment of the invertebrate orthologs of the human SDR12C family variants; Figure S30: Alignment of the invertebrate orthologs of the human SDR16C family variants; Figure S31: Alignment of the invertebrate orthologs of the human SDR21C family variants; Figure S32: Alignment of the invertebrate orthologs of the human SDR25C family variants; Figure S33: Alignment of the invertebrate orthologs of the human SDR26C family variants; Figure S34: Alignment of the invertebrate orthologs of the human SDR28C family variants; Figure S35: Alignment of the invertebrate orthologs of the human SDR32C family variants; Figure S36: Alignment of the invertebrate orthologs of the human SDR42E family variants; Figure S37: Alignment of the primate SDR42E protein variants; Figure S38: Alignment of the primate SDR42E2 proteins variants; Figure S39: Phylogram of SDR7C gene structures; Figure S40: Phylogram of SDR9C gene structures; Figure S41: Phylogram of SDR10E gene structures; Figure S42: Phylogram of SDR11E gene structures; Figure S43: Phylogram of SDR12C gene structures; Figure S44: Phylogram of SDR16C gene structures; Figure S45: Phylogram of SDR21C gene structures; Figure S46: Phylogram of SDR25C gene structures; Figure S47: Phylogram of SDR26C gene structures; Figure S48: Phylogram of SDR28C gene structures; Figure S49: Phylogram of SDR32C gene structures; Figure S50: Phylogram of SDR42E gene structures; Table S1: Structure and catalysis consensuses of the human SDR enzymes belonging to the Classical SDR families; Table S2: Structure and catalysis consensuses of the human SDR enzymes belonging to the Extended, Atypical and Unknown SDR families; Table S3a: Genetic and molecular data of the human SDR7C family and of the human SDR40C1 protein variants; Table S3b: The relative percent identity of the human SDR7C family and of the human SDR40C1 protein variants; Table S4a: Genetic and molecular data of the human SDR11E family protein variants; Table S4b: Relative percent identity of the SDR11E family protein variants; Table S5: Genetic and molecular data of the human SDR10E family protein variants; Table S6a: Genetic and molecular data of the human SDR9C family protein variants; Table S6b: Relative percent identity of the SDR9C family protein variants; Table S7a: Genetic and molecular data of the human SDR12C family protein variants; Table S7b: Relative percent identity of the SDR12C family protein variants; Table S8a: Genetic and molecular data of the human SDR17C family protein variants; Table S8b: Relative percent identity of the SDR16C family protein variants; Table S9: Genetic and molecular data of the human SDR21C family protein variants; Table S10: Genetic and molecular data of the human SDR25C family protein variants; Table S11: Genetic and molecular data of the human SDR26C family protein variants; Table S12: Genetic and molecular data of the human SDR28C family protein variants; Table S13: Genetic and molecular data of the human SDR32C family protein variants; Table S14: Genetic and molecular data of the human SDR42E family protein variants; Table S15: Genetic and molecular data of the vertebrate SDR7C family and of the SDR40C variants; Table S16: Genetic and molecular data of the vertebrate SDR9C family variants; Table S17: Genetic and molecular data of the vertebrate SDR10E family variants; Table S18: Genetic and molecular data of the vertebrate SDR11E family variants; Table S19: Genetic and molecular data of the vertebrate SDR12C family variants; Table S20: Genetic and molecular data of the vertebrate SDR16C family variants; Table S21: Genetic and molecular data of the vertebrate SDR21C family variants; Table S22: Genetic and molecular data of the vertebrate SDR25C family variants; Table S23: Genetic and molecular data of the vertebrate SDR21C family variants; Table S24: Genetic and molecular data

of the vertebrate SDR28C family variants; Table S25: Genetic and molecular data of the vertebrate SDR32C family variants; Table S26a: Genetic and molecular data of the vertebrate SDR42E family variants; Table S26b: Percent identity of the vertebrate SDR42E1 and SDR42E2 variants; Table S27: Genetic and molecular data of the invertebrate orthologs of the human SDR7C family and of the SDR40C1 variants; Table S28: Genetic and molecular data of the invertebrate orthologs of the human SDR9C family variants; Table S29: Genetic and molecular data of the invertebrate orthologs of the human SDR10E family variants; Table S30: Genetic and molecular data of the invertebrate orthologs of the human SDR11E family variants; Table S31: Genetic and molecular data of the invertebrate orthologs of the human SDR12C family variants; Table S32: Genetic and molecular data of the invertebrate orthologs of the human SDR16C family variants; Table S33: Genetic and molecular data of the invertebrate orthologs of the human SDR21C family variants; Table S34: Genetic and molecular data of the invertebrate orthologs of the human SDR25C family variants; Table S35: Genetic and molecular data of the invertebrate orthologs of the human SDR26C family variants; Table S36: Genetic and molecular data of the invertebrate orthologs of the human SDR28C family variants; Table S37: Genetic and molecular data of the invertebrate orthologs of the human SDR32C family variants; Table S38: Genetic and molecular data of the invertebrate orthologs of the human SDR42E family variants; Table S39a: Splicing site organization of the human SDR7C family, of the SDR40C1 variants, and of their respective invertebrate orthologs; Table S39b: Percent identity values of the human SDR7C family, of the SDR40C1 variants and of their invertebrate orthologs; Table S40a: Splicing site organization of the human SDR9C family variants and of their invertebrate orthologs; Table S40b: Percent identity values of the human SDR9C family variants and of their invertebrate orthologs; Table S41a: Splicing site organization of the human SDR10E family variants and of their invertebrate orthologs; Table S41b: Percent identity values of the human SDR10E family variants and of their invertebrate orthologs; Table S42a: Splicing site organization of the human SDR11E family variants and of their invertebrate orthologs; Table S42b: Percent identity values of the human SDR11E family variants and of their invertebrate orthologs; Table S43a: Splicing site organization of the human SDR12C family variants and of their invertebrate orthologs; Table S43b: Percent identity values of the human SDR12C family variants and of their invertebrate orthologs; Table S44a: Splicing site organization of the human SDR16C family variants and of their invertebrate orthologs; Table S44b: Percent identity values of the human SDR16C family variants and of their invertebrate orthologs; Table S45a: Splicing site organization of the human SDR21C family variants and of their invertebrate orthologs; Table S45b: Percent identity values of the human SDR21C family variants and of their invertebrate orthologs; Table S46a: Splicing site organization of the human SDR25C family variants and of their invertebrate orthologs; Table S46b: Percent identity values of the human SDR25C family variants and of their invertebrate orthologs; Table S47a: Splicing site organization of the human SDR26C family variants and of their invertebrate orthologs; Table S47b: Percent identity values of the human SDR26C family variants and of their invertebrate orthologs; Table S48a: Splicing site organization of the human SDR28C family variants and of their invertebrate orthologs; Table S48b: Percent identity values of the human SDR28C family variants and of their invertebrate orthologs; Table S49a: Splicing site organization of the human SDR32C family variants and of their invertebrate orthologs; Table S49b: Percent identity values of the human SDR32C family variants and of their invertebrate orthologs; Table S50a: Splicing site organization of the human SDR42E family variants and of their invertebrate orthologs; Table S50b: Percent identity values of the human SDR42E family variants and of their invertebrate orthologs; Table S51a: Genetic and molecular data of primate SDR42E family variants; Table S51b: Relative percent identity of primate SDR42E family protein variants. References [35,36] are cited in Supplementary Material.

Author Contributions: F.G. contributed to the study conception and design, to data collection and analysis, to the preparation of figures and tables. S.T. handled the phylogenetic analyses. The first draft of the manuscript was written by F.G., S.T. and M.A. contributed to write down the final version. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kallberg, Y.; Oppermann, U.; Jörnvall, H.; Persson, B. Short-chain dehydrogenases/reductases (SDRs). Coenzyme-based functional assignments in completed genomes. *Eur. J. Biochem.* **2002**, *269*, 4409–4417. [CrossRef] [PubMed]
2. Lukacik, P.; Kavanagh, K.L.; Oppermann, U. SDR-type human hydroxysteroid dehydrogenases involved in steroid hormone activation. *Mol. Cell. Endocrinol.* **2007**, *71*, 1265–1266.
3. Persson, B.; Kallberg, Y.; Bray, J.E.; Bruford, E.; Dellaporta, S.L.; Favia, A.D.; Duarte, R.G.; Jörnvall, H.; Kavanagh, K.L.; Kedishvili, N.; et al. The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Interact.* **2009**, *178*, 94–98. [CrossRef] [PubMed]
4. Zhou, Y.; Wang, L.; Ban, X.; Zeng, T.; Li, M.; Guan, X.-Y.; Li, Y. DHRS2 inhibits cell growth and motility in esophageal squamous cell carcinoma. *Oncogene* **2018**, *37*, 1086–1094. [CrossRef] [PubMed]
5. Han, Y.; Song, C.; Wang, J.; Tang, H.; Peng, Z.; Lu, S. HOXA13 contributes to gastric carcinogenesis through DHRS2 interacting with MDM2 and confers 5-FU resistance by a p53-dependent pathway. *Mol. Carc.* **2018**, *57*, 722–734. [CrossRef]
6. Li, J.M.; Jiang, G.M.; Zhao, L.; Yang, F.; Yuan, W.Q.; Wang, H.; Luo, Y.Q. Dehydrogenase/reductase SDR family member 2 silencing sensitizes an oxaliplatin resistant cell line to oxaliplatin by inhibiting excision repair cross complementing group 1 protein ex-pression. *Oncology Rep.* **2019**, *42*, 1725–1734. [CrossRef]
7. Luo, X.; Li, N.; Zhao, X.; Liao, C.; Ye, R.; Cheng, C.; Xu, Z.; Quan, J.; Liu, J.; Cao, Y. DHRS2 mediates cell growth inhibition induced by Trichothecin in nasopharyngeal carcinoma. *J. Exp. Clin. Cancer Res.* **2019**, *38*, 300. [CrossRef]
8. Peltoketo, H.; Luu-The, V.; Simard, J.; Adamski, J. 17b-hydroxysteroid dehydrogenase (HSD)/17-ketosteroid reductase (KSR) family; nomenclature and main characteristics of the 17HSD/KSR enzymes. *J. Mol. Endocrinol.* **1999**, *23*, 1–11. [CrossRef]
9. Oppermann, U.; Filling, C.; Jornvall, H. Forms and functions of human SDR enzymes. *Chem. Biol. Interact.* **2001**, *130–132*, 699–705. [CrossRef]
10. Heinz, S.; Krause, S.W.; Gabrielli, F.; Wagner, H.M.; Andreesen, R.; Rehli, M. Genomic organization of the human gene HEP27: Al-ternative promoter usage in HepG2 cells and monocyte-derived dendritic cells. *Genomics* **2002**, *79*, 608–615. [CrossRef]
11. Wu, X.; Lukacik, P.; Kavanagh, K.L.; Oppermann, U. SDR-type human hydroxysteroid dehydrogenases involved in steroid hormone activation. *Rev. Mol. Cell. Endocrinol.* **2007**, *265–266*, 71–76. [CrossRef]
12. Crean, D.; Felice, L.; Taylor, C.T.; Rabb, H.; Jennings, P.; Leonard, M.O. Glucose reintroduction triggers the activation of Nrf2 during experimental ischemia reperfusion. *Mol. Cell. Biochem.* **2012**, *366*, 231–238. [CrossRef] [PubMed]
13. Benyajati, C.; Place, A.R.; Powers, D.A.; Sofer, W. Alcohol dehydrogenase gene of Drosophila melanogaster: Relationship of inter-vening sequences to functional domains in the protein". *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 2717–2721. [CrossRef] [PubMed]
14. Ladenstein, R.; Winberg, J.-O.; Benach, J. Medium- and short-chain dehydrogenase/reductase gene and protein families. *Cell. Mol. Life Sci.* **2008**, *65*, 3918–3935. [CrossRef]
15. Irimia, M.; Roy, S.W. Spliceosomal introns as tools for genomic and evolutionary analysis. *Nucleic. Acids Res.* **2008**, *36*, 1703–1712. [CrossRef]
16. Tress, M.L.; Wesselink, J.-J.; Frankish, A.; Lopez, G.; Goldman, N.; Löytynoja, A.; Massingham, T.; Pardi, F.; Whelan, S.; Harrow, J.; et al. Determination and validation of principal gene products. *Bioinformatics* **2007**, *24*, 11–17. [CrossRef] [PubMed]
17. Fedorov, A.; Merican, A.F.; Gilbert, W. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci USA* **2002**, *99*, 16128–16133. [CrossRef] [PubMed]
18. Sonnhammer, E.L.L.; Gabaldón, T.; da Silva, A.W.S.; Martin, M.-J.; Robinson-Rechavi, M.; Boeckmann, B.; Thomas, P.; Dessimoz, C. The Quest for Orthologs consortium Big data and other challenges in the quest for orthologs. *Bioinformatics* **2014**, *30*, 2993–2998. [CrossRef] [PubMed]
19. Felsenstein, J. *PHYLIP (Phylogeny Inference Package) Version 3.6*; Department of Genome Science, University of Washington: Seattle, WA, USA, 2015. Available online: https://evolution.genetics.washington.edu/phylip/ (accessed on 1 January 2021).
20. Suchard, M.A.; Lemey, P.; Baele, G.; Ayres, D.L.; Drummond, A.J.; Rambaut, A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **2018**, *4*, vey016. [CrossRef]
21. Spingola, M.; Grate, L.; Haussler, D.; Ares, M., Jr. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **1999**, *5*, 221–234. [CrossRef]
22. Babenko, V.N.; Rogozin, I.B.; Mekhedov, S.L.; Koonin, E.V. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* **2004**, *32*, 3724–3733. [CrossRef] [PubMed]
23. Zhu, T.; Niu, D.K. Mechanisms of intron loss and gain in the fission yeast Schizosaccharomyces. *PLoS One* **2013**, *8*, e61683. [CrossRef] [PubMed]
24. Lim, C.S.; Weinstein, B.N.; Roy, S.W.; Brown, C.M. Analysis of Fungal Genomes Reveals Commonalities of Intron Gain or Loss and Functions in Intron-Poor Species. *Mol. Biol. Evol.* **2021**, *38*, 4166–4186. [CrossRef]
25. Ehsani, S.; Huo, H.; Salehzadeh, A.; Pocanschi, C.L.; Watts, J.C.; Wille, H.; Westaway, D.; Rogaeva, E.; George-Hyslop, P.; Schmitt-Ulms, G. Family reunion–the ZIP/prion gene family. *Progr. Nurobiol.* **2011**, *93*, 405–420. [CrossRef] [PubMed]
26. Betts, M.J.; Guigó, R.; Agarwal, P.; Russell, R.B. Exon structure conservation despite low sequence similarity: A relic of dramatic events in evolution? *EMBO J.* **2001**, *20*, 5354–5360. [CrossRef]

27. Vinckenbosch, N.; Dupanloup, I.; Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 3220–3225. [CrossRef]

28. Kaessmann, H.; Vinckenbosch, N.; Long, M. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* **2009**, *10*, 19–31. [CrossRef]

29. Szczésniak, M.W.; Ciomborowska, J.; Nowak, W.; Rogozin, I.B.; Makałowska, I. Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants. *Mol. Biol. Evol.* **2011**, *28*, 33–37. [CrossRef]

30. Dai, H.; Yoshimatsu, T.F.; Long, M. Retrogene movement within-and between-chromosomes in the evolution of Drosophila ge-nomes. *Gene* **2006**, *385*, 96–102. [CrossRef]

31. Chen, M.; Zou, M.; Fu, B.; Li, X.; Vibranovski, M.D.; Gan, X.; Wang, D.; Wang, W.; Long, M.; He, S. Evolutionary Patterns of RNA-Based Duplication in Non-Mammalian Chordates. *PLoS ONE* **2011**, *6*, e21466. [CrossRef]

32. Navarro, F.C.; Galante, P.A. A genome-wide landscape of retrocopies in primate genomes. *Genome Biol. Evol.* **2015**, *7*, 2265–2275. [CrossRef] [PubMed]

33. Carelli, F.N.; Hayakawa, T.; Go, Y.; Imai, H.; Warnefors, M.; Kaessmann, H. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **2016**, *26*, 301–314. [CrossRef] [PubMed]

34. Casola, C.; Betrán, E. The genomic impact of gene retrocopies: What have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome Biol. Evol.* **2017**, *9*, 1351–1373. [CrossRef] [PubMed]

35. Gabrielli, F.; Tofanelli, S. Molecular and functional evolution of human DHRS2 and DHRS4 duplicated genes. *Gene* **2012**, *511*, 461–469.

36. Meier, M.; Tokarz, J.; Haller, F.; Mindnich, R.; Adamski, J. Human and zebrafish hydroxysteroid dehydrogenase like 1 (HSDL1) proteins are inactive enzymes but conserved among species. *Chem. Biol. Interact.* **2009**, *178*, 197–205.