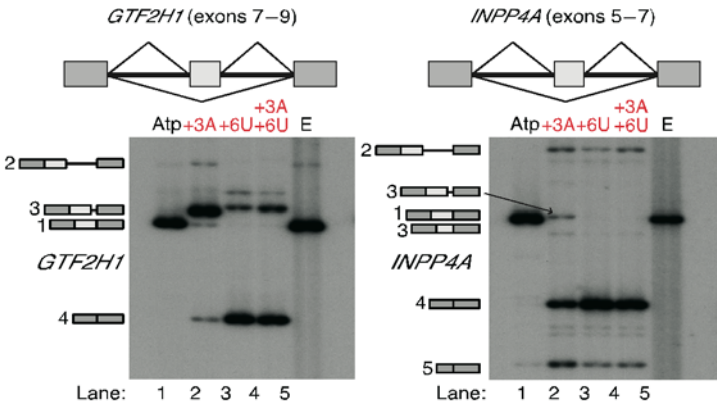
SUPPLEMENTARY MATERIAL

TITLE: Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA

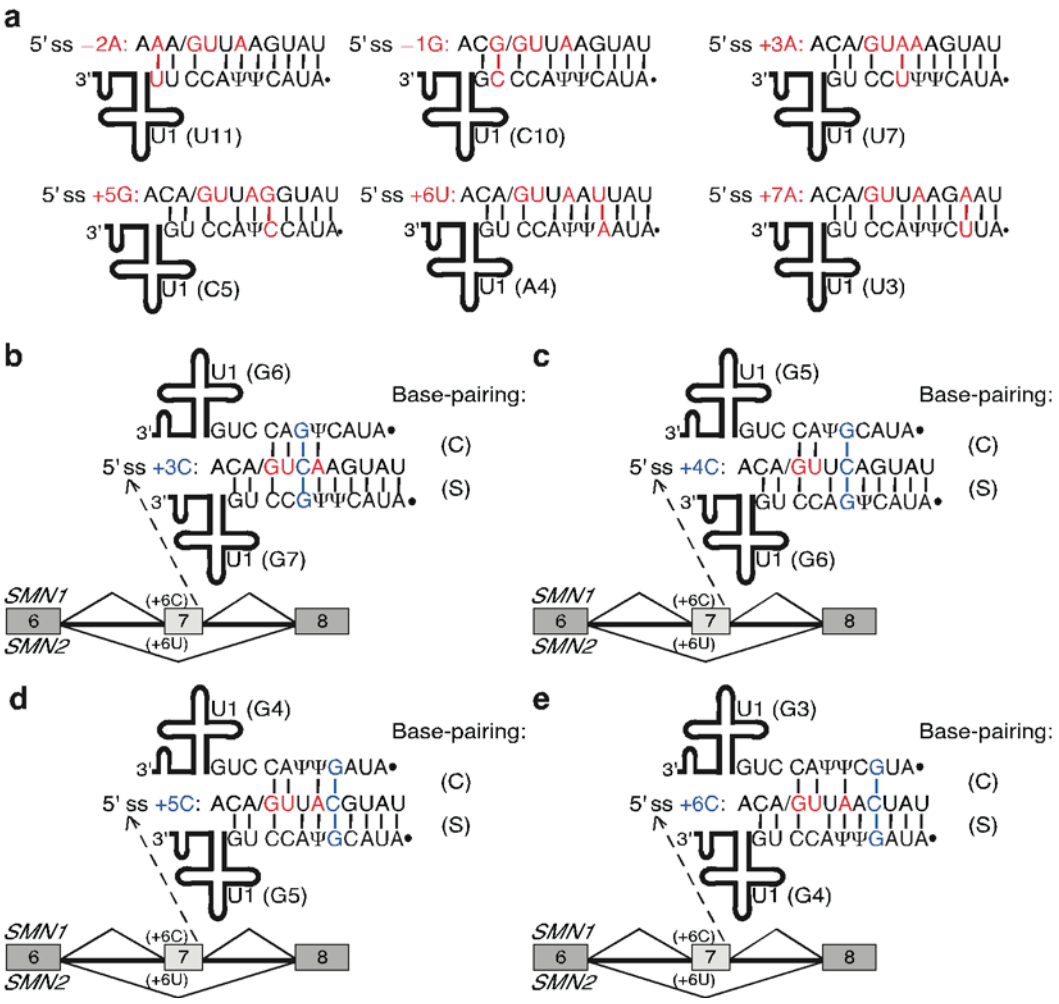
AUTHORS: Xavier Roca and Adrian R. Krainer

SUPPLEMENTARY FIGURES



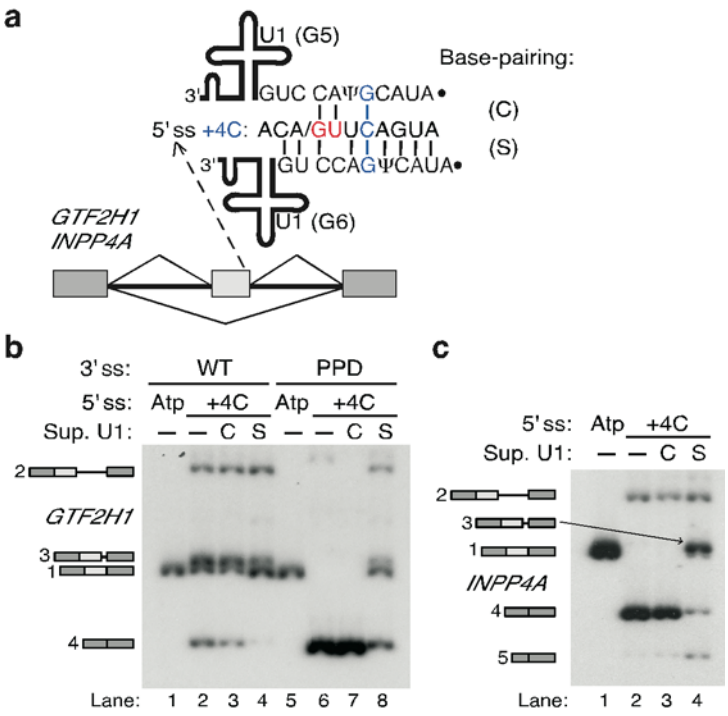
Supplementary Figure 1 Detection of endogenous and transfected *GTF2H1* and *INPP4A* mRNAs. cDNAs from HeLa cells transfected with the various minigenes (lanes 1-4) or mock-transfected (lane 5) were amplified using primers located in the exons flanking the exon containing the atypical 5' ss. The RT-PCR products from the mock-transfected samples correspond to the endogenous *GTF2H1* and *INPP4A* transcripts (E, lane 5), confirming the efficient use of the atypical 5' ss. A small amount of intron 8 retention was seen for endogenous *GTF2H1* (left panel, lane 5). Although the primers can amplify the transcripts from both transfected and endogenous genes, in transiently transfected cells the RT-PCR products are derived mostly from the minigenes, which are expressed at higher levels. In the mock-transfected samples (E, lane 5), ~ 50-fold more total cDNA template was added to the reaction. The identity of the RT-PCR products, schematically depicted on

the left of each panel, is described in the **Fig. 1** legend. For the *GTF2H1* minigene, the mutations at the exon 8 5' ss activated cryptic 5' ss at positions +20 (GAG/GUGCAG) and +25 (GCA/GUAACU) in intron 8 (band #3). For the *INPP4A* minigene, the analogous mutations activated a cryptic 5' ss 31 nucleotides upstream of the 5' ss in exon 6 (CGU/GUAUGA), and to a small extent another cryptic 5' ss at position +176 (AGG/GUAAGA) in intron 6. The arrow indicates a cryptic 5' ss at position +5 (AAA/GUAAUG) of *INPP4A* intron 6 that was activated only in the context of the +3A mutation. In addition, band #5 in *INPP4A* corresponds to an mRNA species generated by skipping of exon 6 and use of a cryptic 5' ss (GCG/GUGAGU) 63 nucleotides upstream of the 5' ss of exon 5.

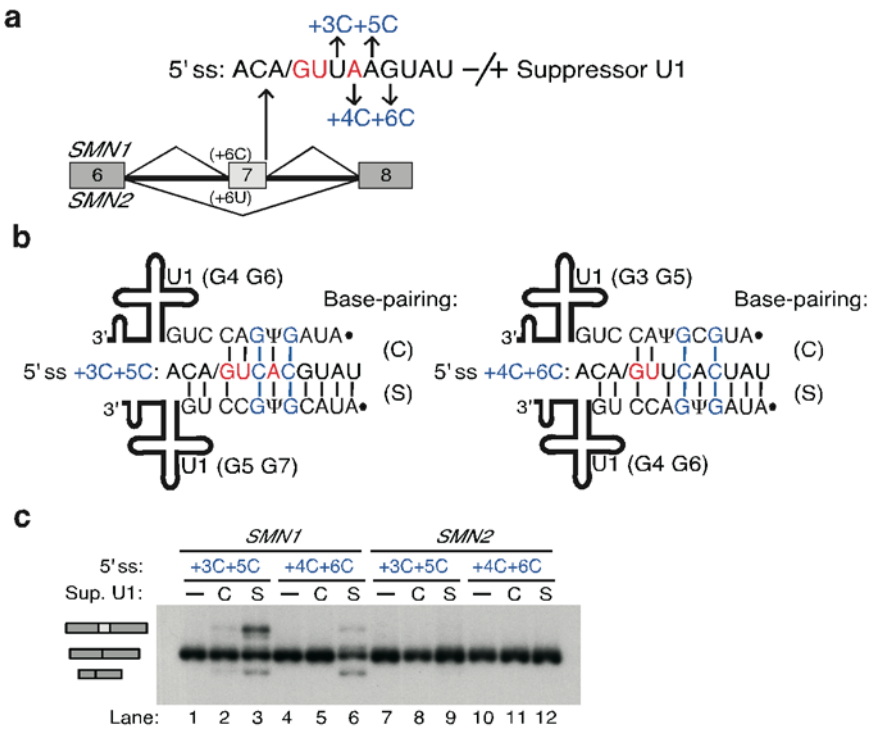


Supplementary Figure 2 Schematic of the base-pairing profiles for the mutant 5' ss used in Figs. 2 and 3. a, Base-pairing profiles of mutants –2A, –1G, +3A, +5G, +6U and +7A with their corresponding suppressor U1 snRNAs restoring base-pairing in the shifted register (**Fig. 2**). Mutations at the 5' ss or U1 are highlighted in red. For these mutant 5' ss, there is no suppressor U1 in the canonical register, because the mutant nucleotide already base-pairs to endogenous U1 in the canonical register. **b-e,** Diagrams show the atypical 5' ss carrying the +3C **(b)**, +4C **(c)**, +5C **(d)** and +6C **(e)** mutations, along with the corresponding suppressor U1 snRNAs in the canonical (C) or shifted (S) base-pairing registers (**Fig. 3**). The 5' ss mutations and the compensatory U1 mutations are highlighted in blue. Note that some of the suppressor U1 snRNAs were used for more than one 5' ss

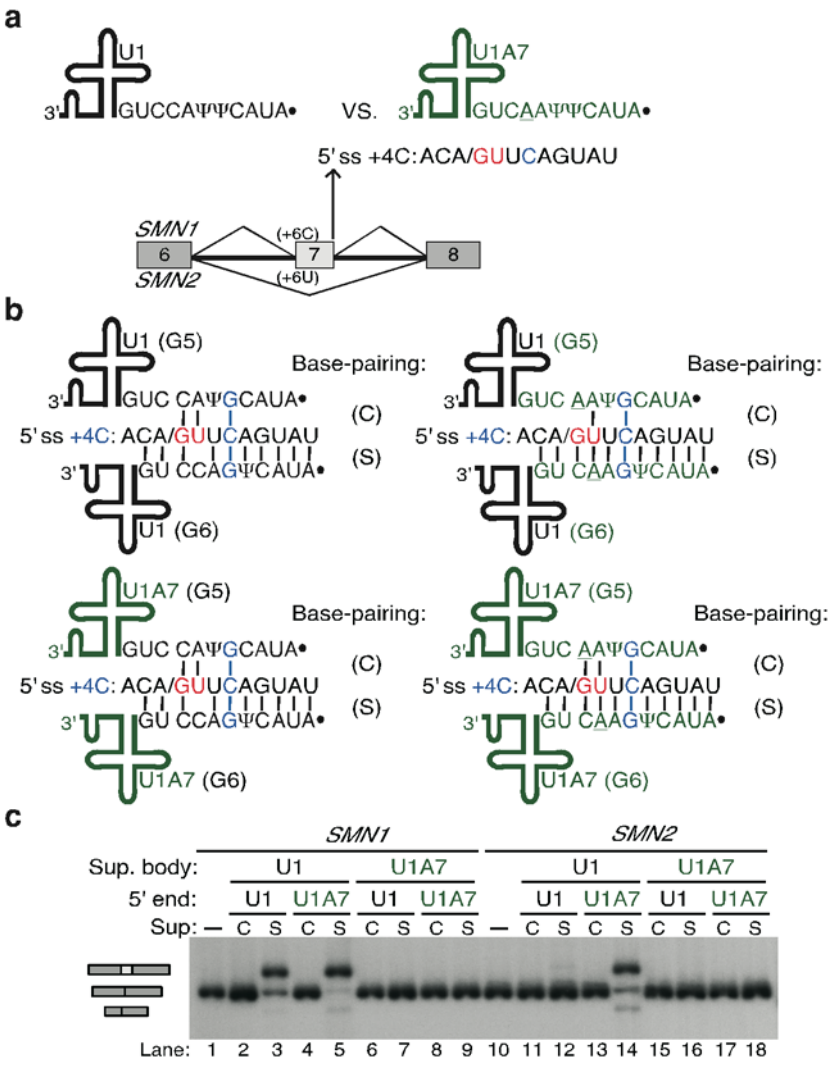
mutation: U1 (G5) compensates for 5' ss mutation +4C in the canonical register, and for mutation +5C in the shifted register. **Fig. 3** shows that U1 (G5) rescued exon 7 inclusion for mutant +5C but not for +4C, and that U1 (G6) rescued mutation +4C (shifted register) but not +3C (canonical register). These results indicate that the distinct effects of the suppressor U1s are not due to their relative expression levels, and further strengthen the demonstration of shifted base-pairing between the atypical 5' ss and U1.



Supplementary Figure 3 Suppressor U1 snRNA analysis for atypical 5' ss in their natural context. a, Base-pairing profiles of the +4C mutant 5' ss in the *INPP4A* and *GTF2H1* minigenes with the corresponding suppressor U1 snRNAs carrying compensatory mutations in the canonical (C) or shifted (S) registers. Mutant nucleotides and compensatory U1 mutations are highlighted in blue. **b,** Suppressor U1 analysis of the +4C mutant 5' ss in the *GTF2H1* minigene. The top labels indicate the version of the 3' ss upstream exon 8, the version of the atypical 5' ss, and the mock or suppressor U1 snRNA used. In lanes 5-8, the 3' ss upstream of exon 8 was weakened by a point mutation at the polypyrimidine tract ('polypyrimidine tract down mutation' or PPD), so as to compromise exon 8 inclusion. Suppressor U1 in the shifted but not in the canonical register rescued exon 8 inclusion in both the wild-type and the PPD minigenes (lanes 2-4, 6-8). **c,** Suppressor U1 analysis in the *INPP4A* minigene. Whereas the suppressor U1 in the canonical register did not rescue exon 6 inclusion (lane 3), the suppressor in the shifted register did, albeit weakly (lane 4).

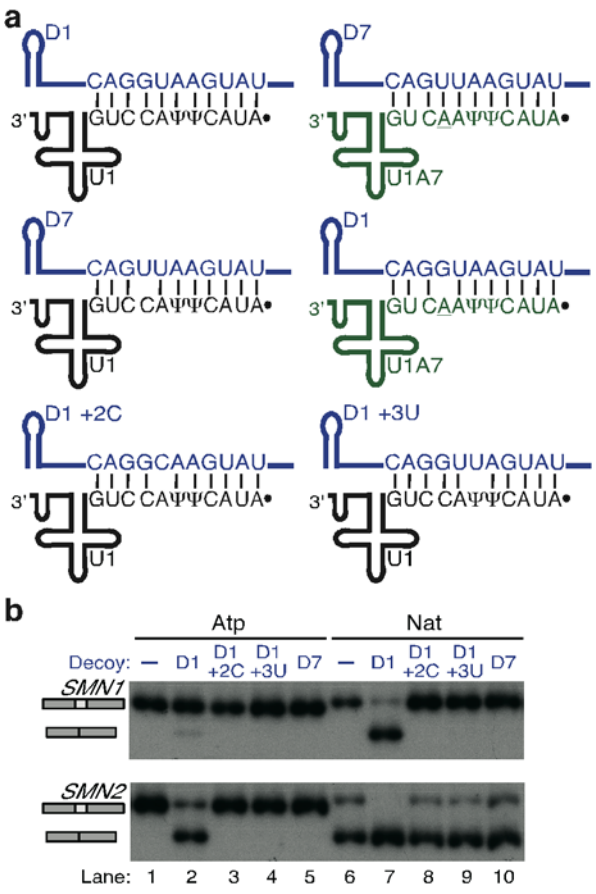


Supplementary Figure 4 Analysis of double mutations at atypical 5' ss. a, Schematic of the +3C+5C and +4C+6C 5' ss mutations in the *SMN1/2* context. Mutant nucleotides are shown in blue. **b,** Base-pairing profiles of the two double-mutant 5' ss with the corresponding suppressor U1 snRNAs carrying compensatory mutations in the canonical (C) or shifted (S) registers. Compensatory U1 mutations are indicated and highlighted in blue. **c,** Analysis of the mutant 5' ss. The top labels indicate the minigene (*SMN1* or *2*), the mutant 5' ss in exon 7, and the mock or suppressor U1 snRNA used. In this case, the suppressor U1 snRNAs rescued exon 7 inclusion only in the *SMN1* context (lanes 1-6). For +3C+5C (lanes 1-3), both suppressors rescued splicing, but the suppressor U1 (G5 G7) in the shifted register was more effective. For +4C+6C (lanes 4-6), only U1 (G4 G6) in the shifted register had an effect. In this experiment, aberrant splicing products were also seen upon use of suppressor U1s, for unknown reasons.



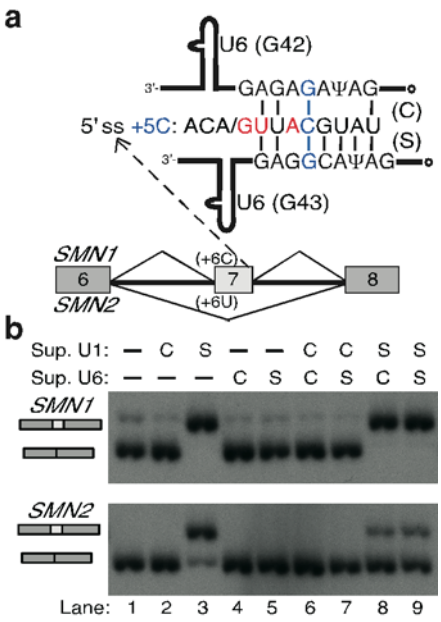
Supplementary Figure 5 Use of U1 and U1A7 snRNA suppressors to rescue splicing via mutant atypical 5' ss. a, Schematic of the U1 (black) and U1A7 (green) snRNAs and the mutant atypical (+4C) 5' ss in the *SMN1/2* context. U1 and U1A7 differ by one nucleotide at their 11-nucleotide 5' ends (underlined in U1A7), and have several additional nucleotide differences in the snRNA body. **b,** Base-pairing of the four different combinations of suppressor U1/U1A7: the U1 snRNA body with the U1 snRNA 5' end; the U1 body with the U1A7 5' end; the U1A7 body with the U1 5' end; and the U1A7 body with its 5' end. For each combination, the compensatory mutation was introduced in the canonical (G5, (C)) or shifted (G6, (S)) register, indicated in blue. **c,** Analysis of the

+4C mutant 5' ss with U1/U1A7 suppressors. The top label indicates the minigene (*SMN1* or *2*), the suppressor body (U1 or U1A7), the 5' end (U1 or U1A7), and the compensatory mutation (control (–), C or S). None of the suppressors with the U1A7 body rescued exon 7 inclusion, but both 5' ends of U1 and U1A7 in the U1 body did.



Supplementary Figure 6 Specificity of the U1/U1A7 snRNA decoys. **a,** Base-pairing profiles of the different decoys used with the 5' end of U1/U1A7. The D1 and D7 decoys have perfect (11 bp) complementarity to their cognate snRNAs (top diagrams). In addition, the D1 decoy has one mismatch (10 bp) to the U1A7 snRNA in the canonical register, and the D7 decoy has one less potential base pair (10 bp) to U1 in the shifted register (middle diagrams). To assess the specificity of the D1 and D7 decoys, two additional decoys were constructed, D1+2C and D1+3U, which can have the same number of base pairs (10 bp) with U1 as the D7 decoy (bottom diagrams). **b,** Decoys need perfect complementarity (11 bp) to their cognate snRNA to have a strong effect. The top label indicates the 5' ss in *SMN1/2* exon 7 and the decoy used. As in b, expression of the D1 decoy resulted in exon 7 skipping. In contrast, the D1+2C, D1+3U and D7 decoys, which have one nucleotide

mismatch (10 bp) when base-pairing with U1, did not affect recognition of the natural or the atypical 5' ss. These results indicate that the snRNA decoys need perfect complementarity to their cognate snRNA, and explain why the D7 decoy does not significantly affect exon 7 inclusion.



Supplementary Figure 7 U6 snRNA does not base-pair to atypical 5' ss in a shifted register. **a,** Schematic of the suppressor U6 experiment. In this case, the suppressor U6 snRNAs carry only a point mutation to compensate for the +5C mutation in the atypical 5' ss. The U6 G42 and G43 mutations rescue base-pairing to the +5C 5' ss in the canonical (C) or shifted (S) register, respectively. Note that the G43 suppressor U6 has much stronger predicted base-pairing to the mutant 5' ss than the G42 suppressor. **b,** RT-PCR analysis of co-transfections of the *SMN1/2* constructs carrying +5C 5' ss and the various suppressor U1 and U6 snRNAs. The suppressor U1s (C or S) used are shown in **Supplementary Fig. 2d** online. As in **Fig. 3**, the suppressor U6s alone did not show any effect on exon 7 inclusion. Combined with suppressor U1, none of the suppressor U6 enhanced the levels of exon 7 inclusion.

SUPPLEMENTARY TABLES

	H. sapiens	*M. musculus*	*D. melanogaster*	*C. elegans*	*A. thaliana*
Search					
NNHGTYRAGT[i]	40	35	10	11	55
NYGGTYRAGT	7	5	3	4	17
NYAGTRRAGT	4	5	0	3	5
NYAGTYYAGT	1	2	1	1	7
NYAGTYRBGT	1	3	2	14	4
NYAGTYRAHT	1	1	2	23	9
NYAGTYRAGV	4	5	2	7	18
Homology[ii]	1	3			
Total	59	59	20	63	115
Human-mouse orthologous 5' ss					
Conserved	27	27			
Not conserved	13	8			
Intron not in ortholog[iii]	1	7			
No ortholog[iv]	10	0			
Not in databases[v]	8	17			
Total	59	59			

Supplementary Table 1 Summary of the in silico searches for atypical 5' ss. [i]See Methods for details about the SpliceRack[1] searches. [ii]Number of 5' ss that were added based on homology to the gene from the other mammalian species. [iii]The other mammalian counterpart does not have an intron mapped at the same location. [iv]Genes for which we were not able to find an ortholog in the other mammalian species. [v]Genes that we were unable to find in the ENSEMBL or UCSC genome resources, and that correspond to putative genes awaiting confirmation. See Supplementary Table 3 for a complete list of atypical 5' ss in all five species.

	HUMAN	*Homo sapiens*					
	Search	NNHGTYRAGT	40				
	Accession	Gene Name	Int. #	Int. Length	Sequence	Orthologs	Ortholog sequence
H1	NM_006314	*CNKSR1*	15	170	AAGAAATAgttaagtcttggggtgtgatgggctggggg	M32	
H2	NM_006252	*PRKAA2*	1	28900	AGTGAAGAgttgagtacgcgctccggaccgctgtgcgg	NO, XM_131633	AGTGAAGAgtgagtgcagcggcgcgggcgggcgggcag
H3	NM_024911	*GPR177*	2	34708	CCAAATCAgttaagtgtactctcctctcatccctttct	NO, NM_026582	CCAAATCAgtaagtaccccaactgctatggtcaactct
H4	NM_015485	*RWDD3*	2	1827	AGTTAAATgttgagtatgaatctggctattttctgctt	NO, NM_028456	ACATCAAGgtgctggaaaacttcattttgaatatgaat
H5	NM_006135	*CAPZA1*	4	4351	TCTGTACTgttaagtatctgactgcttcgttatcagag	M15	
H6	NM_012134	*LMOD1*	2	840	CTCAAGAAgttaagtagttgtgggcatgagccaacagg	NO, NM_053106	TGAAGAAGgtaaaaggtggcaggcagggctaggggct
H7	NM_025145	*C10orf79*	17	2434	AACACACAgttgagttcctatatgactagcttgcttgc	Intron not in mouse ortholog,*D19Ertd652e*	
H8	NM_198795	*TDRD1*	25	3361	GTTAAAAAgttaagtaagttaaatcgtatgttttcgcc	NO, NM_031387	GAAGCTAAgtaagtttcccgtttcccctttctgtgatgc
H9	NM_005316	*GTF2H1*	8	117	AGAAAAACAgttaagtataaatgcagaggtgcagtaact	M19	
H10	NM_033150	*COL2A1*	22	371	GAGAAAGAgttaagtgaatgtggaggctccatcccatg	NO, NM_031163	GAGAGAGAgtaagtaggcaaggagctccagctcacagg
H11	NM_020818	*KIAA1409*	26	9379	CAAAATTCgttgagtatcttctttcatccttccgttaa	M27	
H12	NM_173805	*FLJ38723*	6	479	CAGGCCCAgttgagtctcatggctctccctggggaagg	NO ORTHOLOG	
H13	NM_198830	*ACLY*	5	442	AAGAAAGAgttgagtgagagggacggtagggaggtggc	M5	
H14	NM_002265	*KPNB1*	19	1607	CGTACACCgttgagtatacaacccagttcccttcgtcc	M4	
H15	NM_014906	*PPM1E*	1	198971	TACAAATAgttaagtagctgggcttggctggtggtggg	M26	
H16	NM_032853	*MUM1*	3	1300	GAGGAAAAgttaagtgttgtgttgttttaatactgttt	YES, *Mum1*	AAGGAAAAgttaagtgtgcttgtttttgttttgtttt
H17	NM_022737	*LPPR2*	8	342	CTCGAAAAgttaagtgtggcgcaggtaagggggccgg	M21	
H18	NM_017814	*TMEM161A*	8	104	ATGCTGCAgttaagtgggtcgcttggtgggtgggggga	YES, NM_145597	CTGCTACAgttaggcgggtctcccagggtaggattgga
H19	NM_003419	*ZNF345*	1	1213	AAGAGGAAgttaagtatcaagagtgggggggattgttac	NO ORTHOLOG	
H20	NM_012344	*NTSR2*	2	1834	GGTTCTCAgttaagtcctctgtaacatcaccccagaag	M6	
H21	NM_004027	*INPP4A*	6	284	AGGGAACAgttaagtaatgtgttgtagttcgtgggatt	M1	
H22	NM_004543	*NEB*	11	671	TTAGCACAgttaagtgggaagggatggctggttggtca	NO, XM_130232	TCAGCTCAgtaagccgggatccggggccagatgtcctg
H23	NM_198851	*C22orf34*	4	187	CTAAAATAgttgagttcatagaggcagagggtggggaa	NO ORTHOLOG	
H24	NM_016173	*HEMK1*	1	378	CCAAAACAgttaagtttagtgttgtcaagaggacagga	M24	
H25	NM_198196	*CD96*	11	12792	CAAAAAATgttaagtataatcgtgggtcccttgagtcc	M9	
H26	NM_145026	*SPATS1*	1	9492	GACCTACAgttgagttctaagaaatccaggctgtggag	NO, NM_027649	TTCCAAGGgtagatgaagggcttcccccagggaagtag
H27	NM_033481	*FBXO9*	3	2848	GAAGAAAAgttaagtattatagatattgtaacaaatta	M23	
H28	NM_020320	*RARSL*	2	5285	CAAAAAGAgttaagtataatacacttgtaacaatctat	M16	
H29	NM_004956	*ETV1*	10	1540	GAACAGAAgttaagttgtccaccactgaaattccgtgt	M7	
H30	NM_001735	*C5*	14	979	TGGAAAGAgttaagtaatgcatgtctcgctgcagtgtt	M13	
H31	NM_024657	*MORC4*	9	402	AAGAAACAgttgagtatacatgagatggggttctctac	M25	
H32	XM_378381	*DKFZp686F0839*	9	142	ATGGAGCAgttaagtattctaaagggaacctaaagcaa	NO ORTHOLOG	
H33	XM_290670	*LOC23117*	15	21700	AAATTAAAgttaagtggttttccttttttttttttgt	NO ORTHOLOG	
H34	XM_377716	*LOC402057*	1	87	ATCCTGTTgttaagtgacacgtgactgtatgctaatcg	NO ORTHOLOG	
H35	XM_497955	*NO NAME*	2	10117	AGAACTACgttaagttcatcccagccacagaaattgac	NOT FOUND IN ENSEMBL, UCSC	
H36	XM_379482	*LOC401320*	2	4835	CAAAAACAgttaagtgaaacaaacaaacaatcacacaa	NOT FOUND IN ENSEMBL, UCSC	
H37	XM_059830	*LOC136242*	3	1225	TATTTACCgttgagtatttgggcatccccatccccaaa	NO, *1700016G05Rik*	TATTTACCgtgagtattgagcatcccctccccaaatag
H38	XM_374277	*NO NAME*	1	466	CCCCCAAAgttaagtattattattgttaatatcctcat	NOT FOUND IN ENSEMBL, UCSC	
H39	XM_088578	*LOC375748*	1	3764	CAGAATGAgttgagttggtaattagtttaaaccaaaac	Maps different in ENSEMBL, UCSC	
H40	XM_088525	*FAM125B*	1	13498	GGGGAACAgttaagtgaggcccgggcgccgcggaggct	M12	
	Search	NYGGTYRAGT	7				
H41	NM_144569	*SPOCD1*	10	2355	ATCTGGTGgttaagttggagcctgacctggtgctgggc	NO ORTHOLOG	
H42	NM_001225	*CASP4*	8	1775	TGAAAATGgttaagtattgagagttgttggtggtatgt	NO, NM_007609	TGAGAACAgtaagtatatttccttttagagacctgaat
H43	NM_017828	*COMMD4*	2	222	CCAAGATGgttgagtgcacagggtctagtctgggtgga	M22	
H44	NM_014705	*DOCK4*	4	4972	TCTATGTGgttaagtagctttttaatgacacgattttt	NO, XM_358307	TCTATGTCgtaagtaatgtttaaagagcatggttttta
H45	NM_022444	*SLC13A1*	11	2334	AGAAATTGgttgagtatcatttataattctgcataaat	M17	
H46	XM_496074	*LOC440287*	5	221	CCAAGATGgttgagtgcacagggtctacactgggtgga	NO ORTHOLOG	
H47	XM_496078	*LOC440292*	2	222	CCAAGATGgttgagtgcacagggtgtagtctgagtgga	NO ORTHOLOG	
	Search	NYAGTRRAGT	4				
H48	NM_001851	*COL9A1*	19	2534	GACCTCCAgtaaagtatttttttaaaaaaatatttaacta	M41	
H49	XM_372840	*LOC391209*	1	160	TCCCTGCAgtagagttccttggctggtcaatctacagc	NOT FOUND IN ENSEMBL, UCSC	
H50	XM_114158	*NO NAME*	9	94	AGGAGCCAgtggagttcctgccatggtgtccctgagga	NOT FOUND IN ENSEMBL, UCSC	
H51	XM_379433	*NO NAME*	4	1689	ATACCTTAgtaaagtttactgcttctacggatggccta	NOT FOUND IN	

						ENSEMBL, UCSC	
	Search	NYAGTYYAGT	1				
H52	NM_006256	*PKN2*	7	1280	CTTGTCCAgttcagtaaccagatttttaaaaatcatgt	M46	
	Search	NYAGTYRBGT	1				
H53	NM_139025	*ADAMTS13*	25	466	GCAGTCCAgttatgtcctgtcctccttcctgtcaggca	NO, NM_001001322	AGAGCATGgttaggtttcttcctgtttcctgttgggga
	Search	NYAGTYRAHT	1				
H54	XM_370939	LOC388221	5	3534	AATTTGTAgttaaatttgacctcagtgccctgactaca	NOT FOUND IN ENSEMBL, UCSC	
	Search	NYAGTYRAGV	4				
H55	NM_004476	*FOLH1*	8	882	TCTACACAgttaagagactattttaattttaactcttt	M52	
H56	NM_153696	*PSMAL*	3	887	TCTACACAgttaagagactattttaattttaactcttt	NO ORTHOLOG	
H57	NM_024721	*ZFHX4*	4	9276	GAGCAGCAgttgagatcgacctcaggtaatggttccta	NO, NM_030708	GGCCTCAGgtaaatatgtattcctctcagaagctagcc
H58	XM_375358	*Klkbl4*	5	1216	CATCTGCAgttaaggcattcctcctcccagagaaggct	YES, XM_134585	CGTCTGCAgttaatgcattccttctcctagaggaggct
	Manually added human/mouse orthologs						
H59	NM_015199	*ANKRD28 (KIAA0379)*	21	5662	CTACAATGgttaagtatacaaacacaaatgcatatcat	M39	
	MOUSE	*Mus musculus*					
	Search	NNHGTYRAGT	35				
	Accession	Gene Name	Int. #	Int. Length	Sequence	Orthologs	
M1	NM_030266	*Inpp4a*	4	260	AGGGAACAgttaagtatcgtgcagtcgggcaacagtct	H21	
M2	NM_007495	*Astn1*	17	8557	CATCATTAgtcaagtcagaaatgtctcctcctcctctt	Intron not in human ortholog, NM_004319	
M3	NM_007622	*Cbx1*	5	1420	TTAGCCCTgttgagtaccagcccctgtcacctctgact	Intron not in human ortholog, NM_006807	
M4	NM_008379	*Kpnb1*	19	697	CGTACACCgttgagtatatagcccagctcccaaaatcc	H14	
M5	NM_134037	*Acly*	5	447	AAGAAAGAgttgagttctgagagggaacggcgtgggca	H13	
M6	NM_008747	*Ntsr2*	2	1495	GGTTCTCAgttaagtcctctaccacaataccctgaggg	H20	
M7	NM_007960	*Etv1*	9	1873	GAACAGAAgttaagttcattttataagattgcttataa	H29	
M8	NM_030225	*Dlst*	4	113	TGTGTGCAgttaagtacctgatttcttgtgaatggact	NO, NM_001933	TGTATGCAgtaagtacctgctttcttgggaatggaatt
M9	NM_032465	*Cd96*	10	1677	CAAGAACTgttgagtattgtcatgggtcctccattcat	H25	
M10	NM_009985	*Ctsw*	6	86	CAACAACAgttgagtaccaggccttgtgggaggtcagt	NO, NM_001335	CAACAACAgtgagtgcactgccccgccactctggacat
M11	NM_028412	*Ciz1*	1	507	CAGGATTAgttaagtggcgttggctatatctgcgaggg	NO, NM_012127	CCGCAGAGgtgtgtgtgtgctggcgggaggccgtgccc
M12	NM_175184	*2610528K11Rik*	1	12860	GGGGAACAgttaagtgaggcccgcgcgggaagacgggc	H40	
M13	NM_010406	*Hc*	14	1361	TGCAAAGAgttaagtgaatttccgtggtgaccatgggg	H30	
M14	NM_011978	*Slc27a2*	5	2964	TACAAAGAgttaagtaccccgaaagatgagcaaatgtg	NO, NM_003645	TACAGAAAgtaagtacattgaaaaatgagagcatacgt
M15	NM_009797	*Capza1*	5	4924	TCTGTACTgttaagtatctgcctgtccctcggagaaca	H5	
M16	NM_181406	*Rarsl*	2	4784	AAAAAAGAgttaagtatgcttgtagtgatctcaaattg	H28	
M17	NM_019481	*Slc13a1*	11	2557	AGAAATTAgttgagtatcattttgaacactgaatttat	H45	
M18	NM_019715	*Kcmf1*	3	3286	CCTAGAAAgttgagtaagcagtcatcccagagcacagg	Intron not in human ortholog, NM_020122	
M19	NM_008186	*Gtf2h1*	8	103	AGGAAACAgttaagtagacatgtagaagctgtcactga	H9	
M20	NM_133214	*BC017612*	2	20057	TGCATACAgttgagtgccattaagtatctctgctgtga	NO, *C11orf75*	GCGGGCAGgtgggctgggagcggcgcggcggcggcggc
M21	NM_144935	*BC018242*	7	329	CTCGAAAAgttaagtgtggctcaggtaaggggcccgg	H17	
M22	NM_025417	*Commd4*	1	227	CCAAAATTgttaagtgcactggtttacttagggttggg	H43	
M23	NM_023605	*Fbxo9*	3	1684	GAAGAAAAgttaagtgctgaagtgctctaacaaattaa	H27	
M24	NM_133984	*Hemk1*	2	338	CCAAAACAgttaagtgtaatgttagggaggacagggag	H24	
M25	NM_029413	*Morc4*	12	498	AAGCAACAgttgagtatacacggggcggtgttctccat	H31	
M26	XM_484060	*Ppm1e*	1	109399	TACAAATAgttaagtagcggggctcggccgcgggtggg	H15	
M27	XM_484173	*9030205A07Rik*	21	9472	CAAAATCCgttgagtatctttcctctttgtcttgtgac	H11	
M28	XM_488661	*NO NAME*	5	3011	CCAGCACTgttaagtggaattcatacaattctttctgg	NOT FOUND IN ENSEMBL, UCSC	
M29	XM_358492	*NO NAME*	1	15613	GGCATTCAgttaagtaaattctactgagtccattcagt	NOT FOUND IN ENSEMBL, UCSC	
M30	XM_485305	*NO NAME*	2	205	CCAAAACAgttaagtataaggtcaaataagtgggtttt	NOT FOUND IN ENSEMBL, UCSC	
M31	XM_485339	*1700008P02Rik*	4	1296	AAATCAATgttgagtttatacataaataatattagaat	Intron not in human ortholog	

M32	XM_110525	*Cnksr1*	15	161	AAAAAATAgttaagtcttggaatgagatgggattgggg	H1	
M33	XM_357710	*LOC384557*	11	941	TGTAAACAgttaagtgactctcaatggactctggttgc	Different in ENSEMBL	
M34	XM_134048	*NO NAME*	4	14935	TAAAAACAgttaagtactgcatgtttgtccacatactt	NOT FOUND IN ENSEMBL, UCSC	
M35	XM_487089	*NO NAME*	8	241	AGTTTGGAgttgagtgaggagggcaagagcaggataga	NOT FOUND IN ENSEMBL, UCSC	
	Search	NYGGTYRAGT	5				
M36	NM_145930	*AW549877*	1	92	CCAACATGgttgagtgtctagtaaagtttctaatgtgt	Intron not in human ortholog	
M37	NM_153795	*BC032204*	9	85	ATCTTCCGgttaagtatagggcctgcctggagggagtg	NO, *URP2_HUMAN*	ATTCTTCGgtgagttgggggccagagtaggcagccctg
M38	NM_010576	*Itga4*	16	2149	GAAAAATGgttgagtatttttcttgtatttaaaacataa	NO, NM_000885	AAAAAACAgtaggaatattttccttttattcaaattatt
M39	XM_127673	*Ankrd28*	21	1747	CTACAATGgttaagtatcagataaaaacctgctttgag	YES, *ANKRD28 (KIAA0379)*	CTACAATGgttaagtatacaaacacaaatgcatatcat
M40	XM_131242	*4930579F01Rik*	7	1282	GTCACACGgttaagtacagtttgtggttacttcagatg	NO, *C4orf17*	ACCAAAAGgttaagtacagtttttggtaactattgagt
	Search	NYAGTRRAGT	5				
M41	NM_007740	*Col9a1*	19	2944	GACCTCCAgtaaagtatttttttttattgtttaccaag	H49	
M42	XM_483979	*NO NAME*	2	128	TATTTCTAgtggagtggtgtccatttgggcccaaaatc	NOT FOUND IN ENSEMBL, UCSC	
M43	XM_283061	*Hectd1*	7	886	TCAGGTCAgtaaagtgctgttgtatacatttagttttc	Intron not in human ortholog	
M44	XM_484630	*NO NAME*	1	1470	TAGCACCAgtaaagttcctgttcccatgcatgcccact	NOT FOUND IN ENSEMBL, UCSC	
M45	XM_142197	*Rgag1*	2	620	GTCTCTCAgtaaagtggtaaaagcatcagaaatgctcc	Intron not in human ortholog, NM_020769	
	Search	NYAGTYYAGT	2				
M46	NM_178654	*Pkn2*	6	873	CTTGTCCAgttcagtaaccatacttttaaagtcttgcc	H53	
M47	XM_486735	*No NAME*	2	16518	TTAGGTTAgtttagtgatgtgaaatgccctgtccctac	NOT FOUND IN ENSEMBL, UCSC	
	NYAGTYRBGT		3				
M48	NM_010655	*Kpna2*	12	42411	AAAGACCAgtcgggtctacagagtgagttccaggacag	Intron not in human ortholog, NM_002266	
M49	XM_139844	*EG240055*	1	111	TTCAGCCAgttgggtagcatgatcttccaagccacacc	Does not map to the gene in ENSEMBL, UCSC	
M50	XM_489266	*NO NAME*	1	759	TTTCCCTAgttatgtcaccatctcttataaaaaatcca	NOT FOUND IN ENSEMBL, UCSC	
	Search	NYAGTYRAHT	1				
M51	NM_177011	*9330154K18Rik*	6	639	TGCAAGCAgttgaattagggaaccctaagttccttctg	Intron not in ENSEMBL, gene not found in UCSC	
	Search	NYAGTYRAGV	5				
M52	NM_016770	*Folh1*	8	588	TCAACACAgttgagatatttttctaatgattctgttca	H56	
M53	XM_488594	*LOC432562*	1	2429	CAAGGCCAgtcaagcacctgaccttgaaagttcttgtg	NOT FOUND IN ENSEMBL, UCSC	
M54	XM_358474	*NO NAME*	3	2725	TGTTGCCAgtcaagaaaaacacttcactgtttccacaa	NOT FOUND IN ENSEMBL, UCSC	
M55	XM_488085	*NO NAME*	5	103	ATGGCATAgtcaaggctgagaatgggaagcttatcatc	NOT FOUND IN ENSEMBL, UCSC	
M56	XM_358170	*NO NAME*	1	119	ACTTCACAgttgagccctgaaacaacactgtggagcac	NOT FOUND IN ENSEMBL, UCSC	
	Manually added human/mouse orthologs		3				
M57	NM_023431	*Mum1*	3	976	AAGGAAAAgttaagtgtgcttgtttttgttttgtttt	H16	
M58	NM_145597	*Tmem161a*	8	87	CTGCTACAgttaggcgggtctcccagggtaggattgga	H18	
M59	XM_134585	*Klkbl4*	5	857	CGTCTGCAgttaatgcattccttctcctagaggaggct	H58	

	FRUITFLY	*Drosophila melanogaster*			
	Search	NNHGTYRAGT	10		
	Accession	Gene Name	Int. #	Int. Length	Sequence
D1	NM_057884	*ppk*	2	185	GATGAGGAgttgagtatttcagcatatacttttatag
D2	NM_137551	*CG15117*	4	61	GATACAGAgttgagttattccgacacctcgaagggatt
D3	NM_139436	*CG15822*	1	1762	AAAACCAAgtcgagttacttcacatggaacaacacgca
D4	NM_168190	*NO NAME*	1	283	GCCAAAGCgtcaagtcgtggccgtagctcaaacattgt

D5	NM_140439	*CG5842-RA (nan)*	4	60	CGAATACAgttaagtattgtccttatatttaatataca
D6	NM_169384	*CG31386*	3	6050	AAAAACCAgttgagtaaattatgtactgcaacaacaaa
D7	NM_142173	*CG4210-RA*	1	63	TGATTCAAgttgagttttataagattcctattttaata
D8	NM_142432	*CG7985-RA (CG7985)*	1	11970	CTTTAACAgttaagttcgaaaatggttgaaaaaccaat
D9	NM_167318	*CG1786-RA (Cyp318a1)*	1	63	TCCGAACAgttgagtgtagagtcgaaggccattagagg
D10	NM_078684	*CG3291-RA (pcm)*	5	84	CAGGAACAgttgagtgtgaagccgccgtctggccctgc
	Search	NYGGTYRAGT	3		
D11	NM_166710	*CG32013*	1	89	CGGCGGTGgtcgagtcggcaccgacgtcaacggcggct
D12	NM_169454	*CG4795-RA, (Cpn)*	2	67	AAACTACGgttgagtatcctgtagccgaaaaagtcctg
D13	NM_169605	*CG3631-RB*	1	72	AACAAACGgttaagtcagatagaaagctatttgtgcca
	Search	NYAGTRRAGT	0		
	Search	NYAGTYYAGT	1		
D14	NM_167187	*CG17754-RC*	6	1279	GTGCTACAgtttagtccaaacccaatccactttcggtt
	Search	NYAGTYRBGT	2		
D15	NM_166504	*CG11474-RB*	1	71	GGTGGGTAgttgtgtgcacttctaccttgaccttgagc
D16	NM_130679	*CG3526-RB*	1	1038	TGCGTTTAgttgtgtgtttatgcttaatatgcgtcatt
	Search	NYAGTYRAHT	2		
D17	NM_168054	*CG32259-RB*	4	308	CGGCGCCAgttaattgcaaatactgctcgataccctac
D18	NM_170040	*CG31139-RA*	1	56	CTCAAATAgttaactatatagaatataatataatataa
	Search	NYAGTYRAGV	2		
D19	NM_141447	*CG14609-RA*	6	58	GGAGAATAgttgagcgacgtctacaattaaaaagtagc
D20	NM_133036	*CG15373-RA*	5	109	TCGCACCAgttgagcattgcattattaagaatagaaga

	WORM	*Caenorhabditis elegans*			
	Search	NNHGTYRAGT	11		
	Accession	Gene Name	Int. #	Int. Length	Sequence
C1	NM_071062	*Y46H3D.4*	4	43	CTTTAAAAgttgagttttgatcttagaaagttggcaat
C2	NM_071184	*srbc-22*	3	63	CAAAATCAgttgagtttatttatccaattgaacatttt
C3	NM_074995	*F47H4.2*	1	1995	TGAAAAATgttgagttattttcaaggtcggaaagtaaa
C4	NM_075290	*Y80D3A.8*	4	119	AGCGATGAgtcaagtgattttccgatttgagttgttaa
C5	NM_068590	*C01G5.9*	3	357	TCAAAAATgttaagttttatctttaaaatttgtttaga
C6	NM_069297	*T05A1.7*	2	150	GTGGATATgtcaagtagagaagaagataatgaatcact
C7	NM_070325	*F02H6.3*	3	66	AAAAAGATgttaagtggaagagtttcaaattcctagta
C8	NM_061415	*F46F5.12*	2	44	TGGCATCAgttgagtcaattttcgctgcaacaattaat
C9	NM_061925	*NO NAME*	4	61	TAGAAAAAgttgagttcgaaaatttgaatttcgcgcca
C10	NM_064042	*W02B12.4 (esterase)*	3	578	TTCAAGAAgttgagtttttgaattactattaccggtac
C11	NM_064547	*K09E4.4*	3	53	TCTAAACAgttgagtctgcacggcaatcgaggttgata
	Search	NYGGTYRAGT	4		
C12	NM_070973	*Y50D4B.7*	5	159	CTGGGGTGgttaagttttatttgtaattttaaacgaac
C13	NM_071110	*Y46H3A.5*	2	61	TGCGCATGgtcgagtttctctttttcacggccacgta
C14	NM_072502	*F18E3.10*	1	131	GGTTCCCGgttgagttgatgtcatacatcggtgtgaat
C15	NM_066589	*F59B2.3*	2	50	ATTGAACGgttgagttgttcattttaacataattttct
	Search	NYAGTRRAGT	3		
C16	NM_059201	*T05E8.1*	5	185	TCGCGTCAgtgaagtttgagtaagagatctatcttcca

C17	NM_074054	*srx-42*	7	1497	ACAAAGCAgtgaagtgaaggaaaaaaagaatgaataaa
C18	NM_061650	*F45C12.9*	2	75	TCTTTGCAgtggagttatgaactggaacacctgaaatt
	Search	NYAGTYYAGT	1		
C19	NM_171206	*NO NAME*	1	45	GAATCTTAgttcagtgcatattacgatttttcagatga
	Search	NYAGTYRBGT	14		
C20	NM_058800	*D1037.5 (phospholipase)*	4	540	AACTAACAgttgggtttgatagtttttaaactactgag
C21	NM_059316	*B0207.8*	3	48	AAAATATAgttaggttgaacaatttccggacatgttca
C22	NM_060727	*T09E11.5*	1	45	GTGGATCAgttaggtttttcagtgataacctccagcat
C23	NM_060728	*T09E11.4*	1	44	GTTGATCAgttaggtttttcagtgataacctccagcat
C24	NM_071183	*srbc-21*	2	50	CCAAAACAgttgcgtttatccgcatttttctaaattat
C25	NM_072749	*Y51B11A.1*	1	104	ATCAGACAgttatgtttcttttaaagaaattgccgacc
C26	NM_074793	*nhr-271*	5	78	CGGAAATAgttaggtaattgttaaattgtaatttgaaa
C27	NM_075237	*Y69H2.2*	10	310	TGAAACTAgttgggtttgaaaaaatcgataatttattc
C28	NM_067539	*asm-3*	7	53	TTGGAATAgttgggttttttgaagagtttgggaaaaaa
C29	NM_068159	*R08C7.6*	4	1215	GATTGATAgttaggtttttcggtgcttctcatagttt
C30	NM_070351	*LLC1.2*	4	3370	AATTAATAgttgggttttttttttggaattttgaggg
C31	NM_062809	*R12C12.7*	1	362	TTTTTGTAgttgggtgatatttcgagataaaatcgcaa
C32	NM_075907	*tbx-31*	2	223	TTTTCACAgttatgtaataaggggaacatcaataggtt
C33	NM_065413	*unc-79*	2	82	GCCAGTCAgttgggtatttttccaagggaattggataa
	Search	NYAGTYRAHT	23		
C34	NM_058984	*Y119C1B.9*	1	397	AATAATCAgttaatttttagtttctatcaattgaaat
C35	NM_059145	*C32E12.4*	21	53	TGGCTATAgttaaattggaaattgtagattattgcaca
C36	NM_059245	*F46F11.7*	1	61	TCGGTACAgttaatttctgaataaggaatgataacaac
C37	NM_071952	*K11D12.8*	3	97	TTGGATCAgttgaataaaatctaaaataaatatttgac
C38	NM_074745	*NO NAME*	5	123	TTGCAGCAgttgaatatgacggtaagtgaagattgagc
C39	NM_075053	*srz-45*	3	86	TTCTGATAgttgattatgataacaatgaagataccgat
C40	NM_075147	*C14A6.5*	7	944	TTGTCCCAgttgatttttcgaaatcttctaaccgtttt
C41	NM_067921	*cyp-25A5*	1	293	GACCACCAgttaattggctttagggaaatttgaatatt
C42	NM_068589	*C01G5.7*	3	46	ATTAGTTAgttaactttaactttaaacttaaaaatgct
C43	NM_070559	*Y51H4A.2*	1	82	AGCACTCAgtcaatttagtgtagttttgtgtgtgtgtt
C44	NM_182271	*K02E7.4*	2	246	TATGAACAgttaattgtaataataatattgtatgttat
C45	NM_061573	*Y51H7C.12*	2	1095	ATTGTTCAgttaatttttaaactgtgatactttaataa
C46	NM_061922	*sri-47*	2	49	CCAAAATAgttaatttgaattgaaaatttcagtgcaaa
C47	NM_062134	*F18A12.7*	6	48	TCTGGTCAgttaaatttttaacaaaaactgacttgaat
C48	NM_063632	*T19H5.1*	3	794	ATTCAATAgtcaattgaatgcaatcacctttgacggat
C49	NM_064069	*NO NAME*	1	748	CAGTAACAgttaattttaatatttttttggttccgatt
C50	NM_076485	*H28G03.3*	2	44	TAAAAACAgttaaatttttcagtacttaaacatctaaa
C51	NM_076877	*F46C8.8*	4	57	AGGTATCAgttaaatgttcgctagtatatttcctcgtt
C52	NM_077618	*W03G11.4*	4	797	TTCCCCCAgtcaattaatcactttctttttttttgttt
C53	NM_077707	*R09A8.2*	6	105	AAAGGCTAgttgatttaatgaagtactcacttaatgat
C54	NM_078201	*Y7A5A.9*	5	614	TTGTTCCAgttaacttttccgaaaaaacttaatatgtt
C55	NM_064757	*ver-2*	1	258	AAACGTTAgttaaatactgttgaatttgtgatgaattt
C56	NM_067102	*Y66A7A.5*	5	933	AAAATCCAgttgaatttttctaaatttttttattttg
	Search	NYAGTYRAGV	7		
C57	NM_071343	*srg-61*	3	56	GTTCACCAgttaagcaaaaggtaaaggccgctacttgt
C58	NM_072056	*C24G6.3*	13	74	CCTGAACAgttgagatttttttaatttttaaattaaa
C59	NM_073280	*Y32F6A.1*	6	43	GTCTATCAgttgaggatatatttatagaagttaatagt
C60	NM_069993	*sru-17*	5	52	CAGGAATAgtcaagagagcaggagacaattattgtgat
C61	NM_063629	*R09D1.9 (chitinase)*	3	44	AGAATACAgttgaggtgcgcttatattttgtaaataat
C62	NM_063631	*R09D1.11 (chitinase)*	3	44	AGAATACAgttaaggtacacttatagtttgtaattaat
C63	NM_063632	*T19H5.1*	1	44	AGGAGACAgttgagatattttgagtattgacaatgac
	PLANT	*Arabidopsis thaliana*			

	Search	NNHGTYRAGT	55		
	Accession	Gene Name	Int. #	Int. Length	Sequence
A1	NM_099990	*AT1G01080*	3	87	ATGGAACAgttaagttcgttatccataaaaagaatctt
A2	NM_100351	*acetyl-CoA C-acyltransferase (AT1G04710)*	2	86	ATTGTCGCgttaagtttctcttcttcttttgatattgt
A3	NM_100846	*AT1G09740*	1	92	TGGTCCCAgttgagtctcgttactactaattttgtgat
A4	NM_101003	*AT1G11300*	8	91	AGAATTCAgttaagtattattccaaacctttctcata
A5	NM_101933	*AT1G20810*	1	262	CACTGGCCgttaagtccctcgctctcttcttagctatc
A6	NM_101948	*AT1G20950*	16	99	CTATTACAgttgagttttgtcgtattttctatctaaaa
A7	NM_102114	*AT1G22670*	3	82	CAGTCACAgttgagtatgttgtgatagtttaaactatc
A8	NM_102414	*AT1G26520*	5	303	ATTCAACTgttaagtctctgaaacttctcttctcgtct
A9	NM_102778	*ATMRP13*	18	80	TGCTTCCAgttaagtggaatgaaaggtttacacaactc
A10	NM_102779	*ATMRP12*	18	75	TGTTTCCAgttaagtaaaatgacaggcatacacaaccc
A11	NM_103331	*AT1G36510*	3	492	ACCGTTAAgttgagtttcgaaatataggatttggggta
A12	NM_148569	*AT1G48635*	2	91	AAAACCCAgttaagttctgtgctcataaaaactgatcc
A13	NM_180628	*SCD1*	1	506	TTCCTACTgttaagttttcacaacatcttttgtaatt
A14	NM_202303	*AT1G54920*	6	99	TGTTATCAgttaagttctgtcggagttctcggtgccct
A15	NM_105876	*AT1G72175*	3	115	ATATCGCAgttaagttcgatgacatcttgatctctctg
A16	NM_105969	*AT1G73110*	1	112	AAAAGATTgttaagtctcttttctttttgaaaaa
A17	NM_106111	*AT1G74510*	1	133	AATACTCAgttgagttttgctttttttttttttgat
A18	NM_106305	*AT1G76550*	16	112	CTATCTCAgttgagttttgtctcttatttataaggcta
A19	NM_106594	*GA2*	13	95	TGAAAATTgttaagtatcaatcatcccacacaattata
A20	NM_126485	*AT2G04540*	3	89	AGAGAACAgttaagtcttcctcatctgtagatcaaatg
A21	NM_127266	*AT2G17120*	2	106	AATTGGACgttaagtgattgataatctcgttgtgtttg
A22	NM_179653	*HPT1*	3	93	TTGGCACAgttaagtttctctttaaaaatgtaactctt
A23	NM_127595	*AT2G20360*	1	130	TAATCACCgttaagtttctgccattgatttttttctta
A24	NM_129047	*AT2G34940*	8	79	CTTGTTCAgttaagtgactgctgctttttcttccactt
A25	NM_129519	*AT2G39630*	7	116	TTGCTACAgttaagtcttgcaaacttctgttctagaca
A26	NM_129584	*AT2G40260*	1	125	CCCAGATAgttgagtttcaattatatatatgaatcaat
A27	NM_111905	*AT3G10690*	1	95	AGAATTCTgttgagtgactttcgttcggctcaaagcta
A28	NM_112698	*AT3G18160*	1	92	AAAACCCAgttgagttttgtaacctctcgtctttatt
A29	NM_112821	*AT3G19340*	11	88	TACTTCACgttaagtgctttctttttgcctttccttg
A30	NM_113020	*ATMRP6*	3	121	CTCACCAAgttgagttcctctcagaagttgatcaaatt
A31	NM_113689	*MER3*	21	125	AGAATACAgttaagtcctctaaataactcctctggttt
A32	NM_148769	*AT3G29786*	2	112	AGACGACTgttaagttttgaattagaagtgaaggcatt
A33	NM_114085	*AT3G42180*	2	2411	TTTCATCAgttaagtttatacttttgattacattttct
A34	NM_114662	*ATGLDH*	2	283	AATCACAAgttaagtatcgctaactttcgctatattag
A35	NM_202688	*SPP2*	3	92	AACGTGGCgttaagtgtttacttgttatctttttgtgg
A36	NM_115099	*AT3G52390*	3	88	GAATCATAgttgagtcactccttattgattctttttgt
A37	NM_116013	*AT3G61480*	6	89	AACTTCCAgttgagtatcatcatatccgttatcccctt
A38	NM_116154	*AT3G62890*	1	118	GGATGGAAgtcaagtgtatcagacacgaaatggaggca
A39	NM_116199	*AT3G63340*	2	84	CAAATGCAgttgagtgccttttcattttatttctcttt
A40	NM_116709	*CPK22*	10	163	AAACTGAAgttaagttatcttcagcatcttggcttcac
A41	NM_118234	*AT4G21150*	16	123	TAGTTCTAgttaagttccttcaacttcatttaaaattg
A42	NM_119061	*AT4G29170*	7	142	GGCAAAGAgttaagttctgcataatacgttacttttgt
A43	NM_202922	*AT4G31240*	1	444	TCAGGTATgttaagttttgtttattacttttctcaaca
A44	NM_119844	*RCE1*	3	95	TATTACCAgttaagtagtgcttttttcccactgaaac
A45	NM_121001	*SNG2*	4	99	TTAGGAGTgttaagtgttgttttgctttgttacagtt
A46	NM_121132	*AT5G10940*	9	501	TATTTGCAgttgagttgtttgctgcttaaagcttaaat
A47	NM_121713	*AT5G17070*	7	101	CGAAAACAgttgagtttctgatttgattgttactttct
A48	NM_147869	*AT5G20165*	5	93	GGCATATAgttaagtctatatcctcctactttactgc
A49	NM_122394	*AT5G24850*	2	87	CTCGTACAgttaagttctgtttattgcttgtgttgaag
A50	NM_122718	*AT5G28350*	6	90	AACTTCCAgttgagtatcatcatatccgttatcccctt
A51	NM_180567	*NO NAME*	2	76	AGTTTGCTgttaagttttctttcagccctttatatat
A52	NM_123979	*APE2*	8	100	CCATCATCgttaagtcactttcattatttctcactagt
A53	NM_125498	*AT5G61050*	2	95	AGAATCACgttaagtttcttcatctccgtttccgagtt
A54	NM_125499	*AT5G61060*	10	105	CGTATCACgttaagtgtaccactcttcatcgcagtttc
A55	NM_125658	*AT5G62630*	4	84	TATGCCAAgttaagtgacgaaatcgataaccaaaagcc
	Search	NYGGTYRAGT	17		
A56	NM_101086	*AT1G12140*	4	85	CCTGGATGgttaagtctctcattcaatctttacttttgg

A57	NM_148464	*AT1G14205*	4	119	TTTAGCCGgttaagtaacatttagttggagtttgtttt
A58	NM_104157	*AT1G52790*	2	83	CATTGATGgttaagttaaatcataatacaaaattacaa
A59	NM_104543	*AT1G56670*	4	82	CCAATACGgttaagtagagttactatttacccgtacca
A60	NM_104587	*AT1G58050*	5	86	TGCTCTTGgttaagtttaaccctttttggtaatcact
A61	NM_104588	*AT1G58060*	5	84	TGATCTTGgttaagtttaatctttagtaatgaatcagt
A62	NM_179988	*AT2G39670*	1	71	CAACCTCGgtcaagtaaagatctgatttttgcatataa
A63	NM_180001	*AT2G40640*	7	73	AATGGATGgttaagtctcgttgcctcatttcatccata
A64	NM_180313	*AT3G28270*	1	543	AATTTCTGgttaagtttttctctcttcaactctctttt
A65	NM_115438	*SBPASE*	2	92	AAAGCTTGgttgagtaacacaaagactcaatctttatt
A66	NM_115577	*AT3G57170*	4	87	ATGAAATGgttgagtagtttccaaaatgcttctctttg
A67	NM_117755	*AT4G16560*	12	255	TGCTCCTGgttaagtctctatccatttaatcttttgat
A68	NM_119341	*AT4G31900*	3	210	GCTTGCCGgttaagtgattcaagtcttgtcctttgttt
A69	NM_120931	*AT5G08460*	4	121	TAATTACGgttaagtatctaacttatcccttgttttac
A70	NM_121236	*AT5G11980*	7	92	AGTGCATGgttgagttcttctgcatcttgtgtattcct
A71	NM_124455	*AT5G50770*	3	85	CGGCTATGgtcaagttactccgaccaaagatctctcta
A72	NM_125539	*MIM*	22	85	AACTCATGgttgagtctgcagatctgtttatgttaaag
	Search	NYAGTRRAGT	5		
A73	NM_100395	*AT1G05170*	3	88	GCATTACAgtaaagtttcaatcttttgctgtgtcattt
A74	NM_101251	*AT1G13830*	2	86	ACCTTCCAgtaaagtcaataactcctaaaagtctaaaa
A75	NM_101918	*AT1G20680*	5	1458	AACATCCAgtagagttacaaagaaacattgcgcagcat
A76	NM_114647	*ATATH7*	16	92	ACTCACCAgtagagtcttctaagcatcttatctccttt
A77	NM_124406	*AT5G50270*	1	148	ATCATCCAgtgaagttcctgagttctattacttcagtc
	Search	NYAGTYYAGT	7		
A78	NM_100877	*AT1G10030*	1	1168	GACCACCAgttcagttctccctttttaatctctatttt
A79	NM_103923	*AT1G50400*	1	285	AGCTACCAgttcagtatcaccctctctctctactgctt
A80	NM_129435	*AT2G38780*	1	209	TAAACCCAgttcagtaaaaaatcatcttttgcacaaaa
A81	NM_129497	*AT2G39400*	2	98	CATGAACAgttcagtcagatattaaaatatttactaac
A82	NM_112890	*TOM40*	1	498	AGCTCTCAgttcagtatatcctaaatccctttttccgat
A83	NM_118130	*AT4G20110*	7	92	CTTGCTCAgtttagtgattctcttgtcataagccaaac
A84	NM_119538	*AT4G33800*	3	115	AATAACCAgttcagtaatttttattattatttctgta
	Search	NYAGTYRBGT	4		
A85	NM_180015	*AT2G41160*	1	87	CGGTTTCAgttacgtatcgaaattgaataacccatttc
A86	NM_180271	*AT3G17265*	1	86	CAAGAACAgttatgttttaggatatggaaacaacaaca
A87	NM_118752	*AT4G26190*	3	71	TCTTCTTAgttaggtgtcttatacatatacgaagttta
A88	NM_123512	*GLA1*	5	95	GTCGCCCAgttaggtggtttttatgtgtttgcttatct
	Search	NYAGTYRAHT	9		
A89	NM_100544	*NIH*	11	105	CTTATCCAgttaattcctctctctttatgtctctctct
A90	NM_104690	*AT1G59980*	7	91	TTAATTCAgttaattattttctctaccctcttaagcta
A91	NM_105399	*AT1G67310*	1	141	AGTCCATAgtcaaattattacaccttcttttgtctcta
A92	NM_106238	*AT1G75880*	2	87	TAGCGGTAgttaattttgagtacctatgcataatttaa
A93	NM_179823	*HVT1*	11	84	CTTATCCAgttaatttctctttttctctctatatctcg
A94	NM_179859	*ATCSLB03*	5	94	TACAATCAgttaaatttctcttactctttatatttttt
A95	NM_114201	*AT3G43330*	7	61	AAGAATCAgttaactacatgtttcctacctagctttgc
A96	NM_119567	*AT4G34060*	2	41	TGTAATCAgttaattcacttatttttatttttaactttt
A97	NM_122884	*AT5G34930*	1	117	CAGAGACAgttgaatccatcatcatcatcatcatcgtc
	Search	NYAGTYRAGV	18		
A98	NM_100923	*AT1G10490*	20	94	AAATACCAgttgagaatccctctttccctcttctcctg
A99	NM_127387	*AT2G18290*	1	101	TATTGGCAgttgagaaaacgaaaaaaaaactaaagctc
A100	NM_127544	*ATHXK2*	3	105	ACGATACAgttgagcttgttctccttttcatctctttcc
A101	NM_128895	*AT2G33350*	1	239	TTAACATAgttaagccagaacaaaattacttaagtata
A102	NM_179985	*AT2G39410*	3	194	CATGAATAgttgagatcaacccacttcacaaattctat

A103	NM_129500	*AT2G39420*	3	97	CATGAACAgttgagatacttcagccaacttcatgattt
A104	NM_113609	*AT3G26950*	6	92	CAGTCTCAgttaagatcgtttttagtttcctcaacac
A105	NM_117800	*AT4G16970*	11	89	CCTGAACAgttaagatctgacttgaatgagtacaccat
A106	NM_119057	*ATHXK1*	3	86	AAGAAGCAgttaagctcgattttcgttcaactatttca
A107	NM_119554	*AT4G33940*	3	210	TCTTTGCAgttaagaacatcttaccctgaaatttacat
A108	NM_120026	*AT4G38650*	3	87	CTTCTCCAgttaagcatttcatgctttcatctctagaa
A109	NM_120494	*AT5G04120*	3	108	CTGTTGCAgttaagatccaatccctttaacttccctgt
A110	NM_120572	*AT5G04900*	12	114	TATTCTCAgttaagacgattaatagctctgcgtagaag
A111	NM_122051	*AT5G20440*	3	93	CATGAACAgttaagacacaatcattatcatccatagtt
A112	NM_122301	*AT5G23960*	5	78	AGAAATCAgttaagaccaaaagaatcaactttaagcat
A113	NM_148058	*AT5G38386*	1	105	AGAAACTAgttaagctccattatctggatttgatattc
A114	NM_125499	*AT5G61060*	8	217	CTACTCCAgttgagcacccttctttttgtctaccattt
A115	NM_125736	*AT5G63410*	5	192	ATAACACAgttaagctcgtttctttaatgttctgtttc

Supplementary Table 2 Complete list of predicted atypical 5' ss in five species. Each atypical 5' ss is labeled with the initial of the species where it was found, and a number. For each 5' ss, we provide the accession number, the gene name, the intron number and length, and a sequence that includes the atypical 5' ss (8 nucleotides in the exon, 30 in the intron). The query sequence used to obtain each hit is also shown. For the human and mouse collections, the orthologous 5' ss in the other species is indicated, whether it is conserved or not. Cases for which the orthologous intron or gene could not be found are also indicated.