Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

The matrices and constraints of GT/AG splice sites of more than 1000 species/lineages

Hai Nguyen[a,b], Urmi Das[a], Benjamin Wang[a,c], Jiuyong Xie[a,*]

[a] *Department of Physiology & Pathophysiology, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB R3E 0J9, Canada*
[b] *University of Winnipeg, Winnipeg, MB R3B 2E9, Canada*
[c] *University of Illinois Urbana-Champaign, IL, USA*

ARTICLE INFO

Keywords:
Splice sites
Intron
Diversity
Alternative splicing

ABSTRACT

To provide a resource for the splice sites (SS) of different species, we calculated the matrices of nucleotide compositions of about 38 million splice sites from > 1000 species/lineages. The matrices are enriched of a GGTAAGT (5′SS) or (Y)$_6$N(C/t)AG(g/a)t (3′SS) overall; however, they are quite diverse among hundreds of species. The diverse matrices remain prominent even under sequence selection pressures, suggesting the existence of diverse constraints as well as U snRNAs and other spliceosomal factors and/or their interactions with the splice sites. Using an algorithm to measure and compare the splice site constraints across all species, we demonstrate their distinct differences quantitatively. As an example of the resource's application to answering specific questions, we confirm that high constraints of particular positions are significantly associated with transcriptome-wide, increased occurrences of alternative splicing when uncommon nucleotides are present. More interestingly, the abundance of alternative splicing in 16 species correlates with the average constraint index of splice sites in a bell curve. This resource will allow users to assess specific sequences/splice sites against the consensus of every Ensembl-annotated species, and to explore the evolutionary changes or relationship to alternative splicing and transcriptome diversity. Web-search or update features are also included.

1. Introduction

Splice sites (SS) demarcate exons and introns allowing the proper joining of exons during the expression of most eukaryotic genes. Their selective usage during alternative splicing produces more than one transcript from a single gene thereby contributing to transcriptomic and proteomic diversity (Black, 2003; Nilsen and Graveley, 2010). Their importance has been clearly demonstrated by splice site mutations that cause diseases (Tazi et al., 2009; Scotti and Swanson, 2016; Daguenet et al., 2015; Feng and Xie, 2013). Genomic analyses of different individual species/groups have given a glimpse of the consensus and diversity of both constitutive and alternative splice sites (Dou et al., 2006; Thanaraj and Stamm, 2003; Rogozin and Milanesi, 1997; Sibley et al., 2016; Szczesniak et al., 2013; Abril et al., 2005; Garg and Green, 2007; Burset et al., 2001). However, a centralized source for an overview of the annotated, millions of splice sites among all the currently sequenced eukaryotes remains to be created.

In biological or biomedical research, one often needs to assess the strength of particular sequences as splice sites or compare them between species, in fields such as genetics, cell biology, biochemistry or physiology. A resource with quantitative, comparable measurements of the splice site consensus and constraints of different species would be a very helpful reference. We thus compiled this resource for the consensus and diversity of the splice sites of the GT/AG introns of the Ensembl-annotated eukaryotic species as a reference for simple search or further exploration.

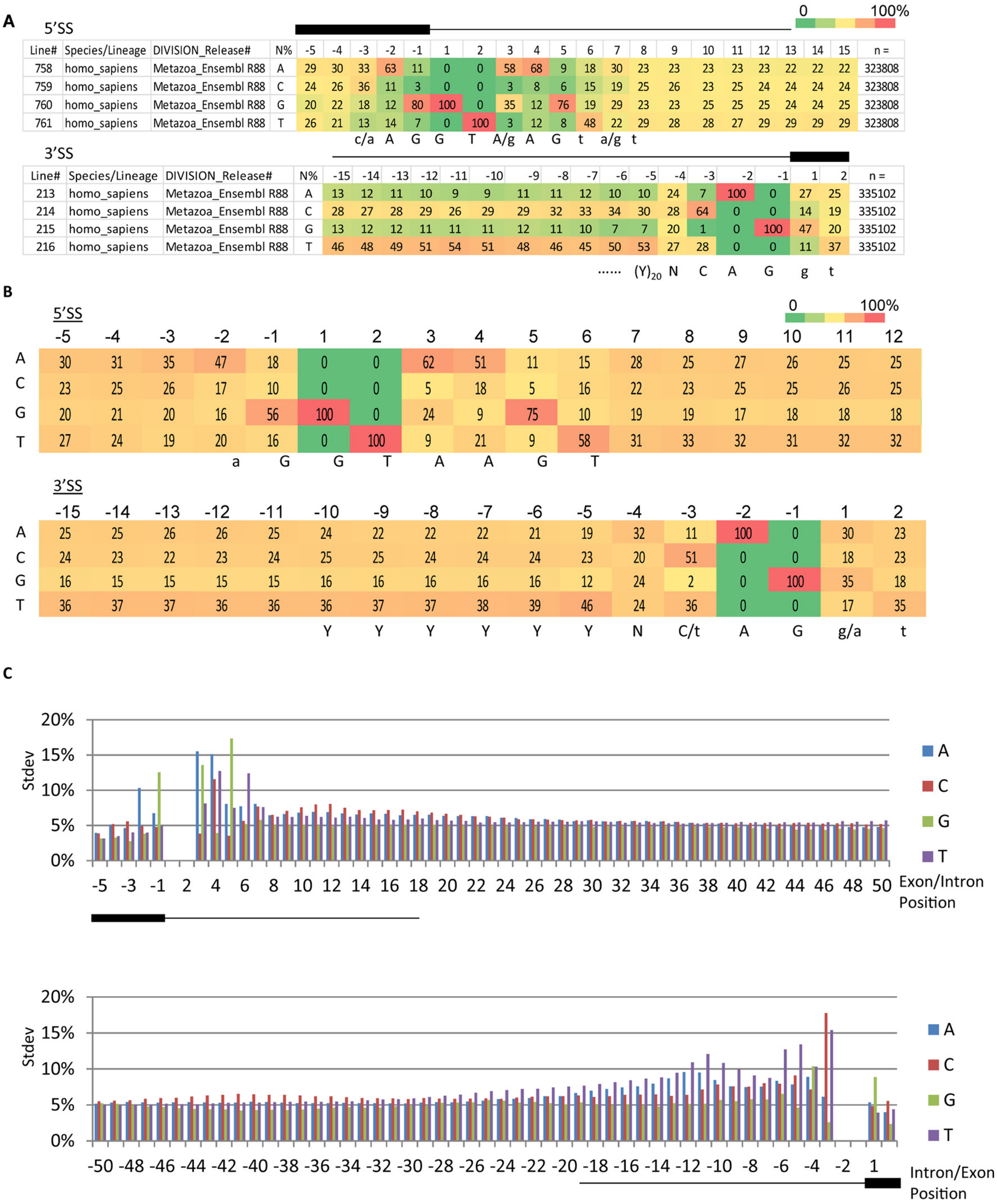
The GT/AG splice sites present in the majority of eukaryotic introns, characterized in humans with a consensus AGGTRAGT at the 5′ splice site and **A**(Y)$_n$NYAGG (underlined: intron start/end GT/AG, **A**: branch point, Yn: polypyrimidine tract Py, N: A,C,G or T, n:6–35) at the 3′ splice site, with variations (except the GT/AG) in other species (Sibley et al., 2016; Moore, 2000; Burge et al., 1998; Spingola et al., 1999; Mount et al., 1992; Lorkovic et al., 2000). These sequences are recognized through direct base-pairing by the snRNAs of snRNP splicing factors (U1, U2, U5 and U6, with the participation of U4) or through contact by accessory proteins such as U2AFs during the dynamic assembly of spliceosomes (Will and Luhrmann, 2011; Shi, 2017). In this report, we used the Ensembl-annotated databases to compile a complete list of the matrices and constraints of the splice sites. Since the branch point is not as easy to assess accurately as the other motifs, it is not

Abbreviations: RNA, ribonucleic acid; SS, splice site

[*] Corresponding author.

E-mail address: xiej@umanitoba.ca (J. Xie).

https://doi.org/10.1016/j.gene.2018.03.031
Received 17 October 2017; Received in revised form 8 March 2018; Accepted 12 March 2018
Available online 26 March 2018



A

5′SS

Line#	Species/Lineage	DIVISION_Release#	N%	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	n =
758	homo_sapiens	Metazoa_Ensembl R88	A	29	30	33	63	11	0	0	58	68	9	18	30	23	23	23	23	23	22	22	22	323808
759	homo_sapiens	Metazoa_Ensembl R88	C	24	26	36	11	3	0	0	3	8	6	15	19	25	26	23	25	24	24	24	24	323808
760	homo_sapiens	Metazoa_Ensembl R88	G	20	22	18	12	80	100	0	35	12	76	19	29	23	23	25	25	25	24	24	25	323808
761	homo_sapiens	Metazoa_Ensembl R88	T	26	21	13	14	7	0	100	3	12	8	48	22	29	28	28	27	29	29	29	29	323808
						c/a	A	G	G	T	A/g	A	G	t	a/g	t								

3′SS

Line#	Species/Lineage	DIVISION_Release#	N%	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	n =
213	homo_sapiens	Metazoa_Ensembl R88	A	13	12	11	10	9	9	11	11	12	10	10	24	7	100	0	27	25	335102
214	homo_sapiens	Metazoa_Ensembl R88	C	28	27	28	29	26	29	29	32	33	34	30	28	64	0	0	14	19	335102
215	homo_sapiens	Metazoa_Ensembl R88	G	13	12	12	11	11	11	12	11	10	7	7	20	1	0	100	47	20	335102
216	homo_sapiens	Metazoa_Ensembl R88	T	46	48	49	51	54	51	48	46	45	50	53	27	28	0	0	11	37	335102
													......	$(Y)_{20}$	N	C	A	G	g	t	

B

5′SS

	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11	12
A	30	31	35	47	18	0	0	62	51	11	15	28	25	27	26	25	25
C	23	25	26	17	10	0	0	5	18	5	16	22	23	25	25	26	25
G	20	21	20	16	56	100	0	24	9	75	10	19	19	17	18	18	18
T	27	24	19	20	16	0	100	9	21	9	58	31	33	32	31	32	32
				a	G	G	T	A	A	G	T						

3′SS

	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2
A	25	25	26	26	25	24	22	22	22	21	19	32	11	100	0	30	23
C	24	23	22	23	24	25	25	24	24	24	23	20	51	0	0	18	23
G	16	15	15	15	15	16	16	16	16	16	12	24	2	0	100	35	18
T	36	37	37	36	36	36	37	37	38	39	46	24	36	0	0	17	35
						Y	Y	Y	Y	Y	Y	N	C/t	A	G	g/a	t

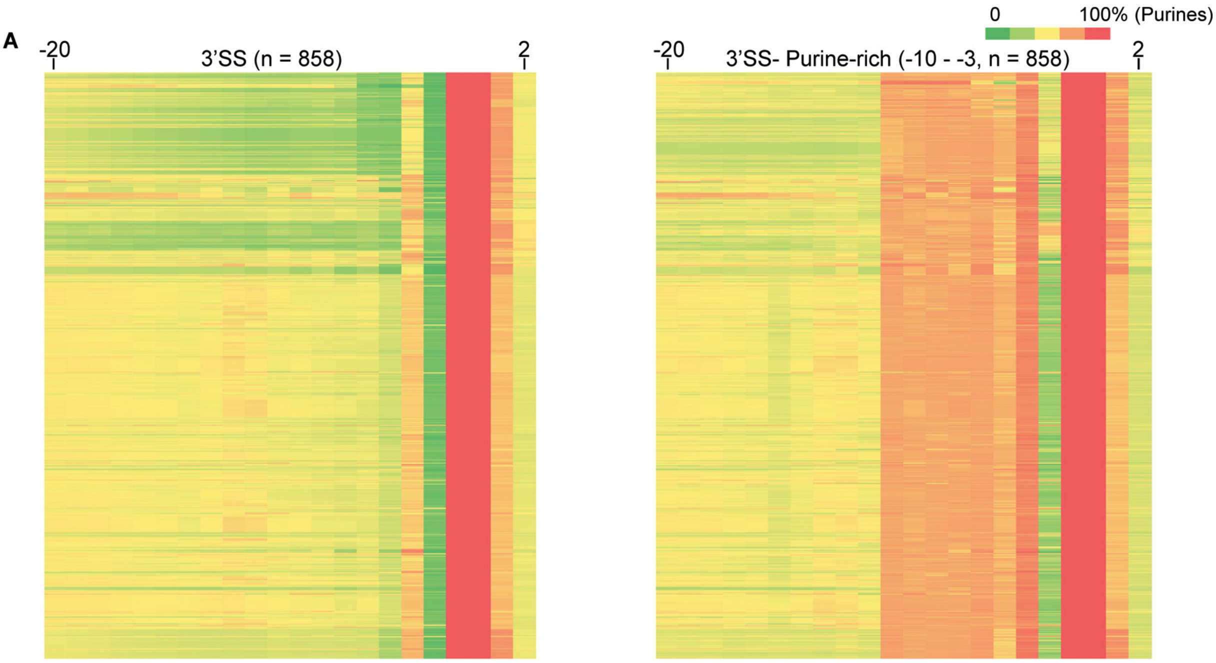
Fig. 1. Matrices and diversity of the splice sites. A. An example of the format of splice site matrices, with the human 5′ and 3′ splices sites in heat maps of the percentages of nucleotides at each position. Heavy black bar: exons, black line: introns. Below the matrices, the lowercased nucleotides are enriched well above background but < 50%, and the uppercased are above 50%. B. Average percentages of the nucleotide compositions of 30,995,943 5′ splice sites of 680 species (upper) or of 31,473,266 3′ splice sites of 682 species (lower). C. Standard deviations of the average percentages of the nucleotides in B.

included here. Also not included are the minor AT/AC introns (< 0.5%) (Burge et al., 1998; Verma et al., 2017; Wu and Krainer, 1999; Hall and Padgett, 1994; Levine and Durbin, 2001).

2. Results

2.1. The matrices of GT/AG splice sites of > 1000 eukaryotic species/lineages

We calculated the percent nucleotide compositions of the 5′ and 3′ GT/AG splice sites of 1074(5′)/1076(3′) species or their lineages/strains (hereafter 'lineage', to represent all in the same species, S_Tables I–II). An example of the resulting matrix format is shown in Fig. 1A, with the average percentages of > 300 thousands of human splice sites. The matrices are enriched of (c/a)AGGT(A/g)AGt (5′SS) or $(Y)_{20}$NCAGgt (3′SS, upstream beyond the $(Y)_{20}$ is T/A-rich), similar to those based on about 3000 human splice sites in total (Zhang, 1998). There is also a slight increase of A_7T_8, which could also participate in U1 snRNA base-pairing and splicing (Freund et al., 2005; Roca et al., 2012). However, the A_4 and G_5 of the 5′SS and the G_1 of the 3′SS are 3.5%, 4.5% and 7% less than those in the previous one on average. There is also substantial enrichment of 5′SS G_3 (35% vs 3% of C_3 or T_3),



B

Species/Lineages	Ensembl Divisions	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	n=
brugia_malayi	Metazoa_GR35	53	82	86	84	84	49	32	70	40	100	100	77	45	3731
caenorhabditis_brenneri	Metazoa_GR35	55	86	87	86	84	56	30	64	37	100	100	78	41	3407
caenorhabditis_briggsae	Metazoa_GR35	57	84	85	85	83	61	38	61	45	100	100	72	45	3483
caenorhabditis_elegans	Metazoa_GR35	54	89	92	91	87	49	19	59	31	100	100	78	37	2705
caenorhabditis_japonica	Metazoa_GR35	56	86	87	86	82	55	29	66	38	100	100	77	42	2731
caenorhabditis_remanei	Metazoa_GR35	52	84	84	85	83	57	35	65	43	100	100	76	41	3088
dendroctonus_ponderosae	Metazoa_GR35	45	77	80	80	77	57	44	77	34	100	100	74	40	1756
loa_loa	Metazoa_GR35	53	86	87	86	85	46	23	73	35	100	100	81	41	2847
onchocerca_volvulus	Metazoa_GR35	52	84	86	86	85	48	26	71	38	100	100	79	43	3442
pristionchus_pacificus	Metazoa_GR35	54	82	79	81	83	55	29	71	43	100	100	81	40	10561
strongyloides_ratti	Metazoa_GR35	51	77	72	76	77	64	48	73	55	100	100	72	43	836
tribolium_castaneum	Metazoa_GR35	51	78	78	81	82	57	46	76	33	100	100	71	42	3050
trichinella_spiralis	Metazoa_GR35	47	76	76	78	77	61	42	76	46	100	100	64	45	4265
aphanomyces_astaci	Protists	43	78	67	70	68	74	62	89	13	100	100	75	43	8520
aphanomyces_invadans	Protists	44	78	68	68	70	73	60	89	13	100	100	76	42	7478
babesia_bigemina	Protists	49	79	74	73	69	76	38	87	21	100	100	72	46	746
babesia_bovis	Protists	46	78	78	63	71	65	55	93	12	100	100	84	37	578
babesia_microti_strain_ri	Protists	49	82	75	41	79	85	53	72	37	100	100	67	42	514
blastocystis_hominis	Protists	51	33	84	85	74	84	66	90	19	100	100	71	49	3886
capsaspora_owczarzaki_atcc_30864	Protists	50	76	71	78	60	75	79	91	9	100	100	50	32	4437
cyclospora_cayetanensis	Protists	41	77	68	76	69	78	42	86	29	100	100	77	46	2342
eimeria_acervulina	Protists	41	77	67	78	58	81	44	87	28	100	100	81	39	1566
eimeria_brunetti	Protists	41	77	66	79	68	76	46	88	28	100	100	77	47	1621
eimeria_mitis	Protists	43	78	62	78	67	76	49	89	30	100	100	75	47	1622
eimeria_necatrix	Protists	39	75	68	75	65	78	42	88	29	100	100	77	40	1366
eimeria_tenella_gca_000499545	Protists	39	73	69	79	62	80	42	90	25	100	100	80	37	1235
entamoeba_invadens_ip1	Protists	44	40	83	78	47	82	78	85	60	100	100	68	52	1407
monosiga_brevicollis_mx1	Protists	38	76	68	69	68	70	66	88	14	100	100	70	38	5563
perkinsus_marinus_atcc_50983	Protists	43	78	66	72	76	62	74	77	15	100	100	92	15	5747
plasmodiophora_brassicae	Protists	42	74	69	70	68	70	64	90	13	100	100	77	38	3380
reticulomyxa_filosa	Protists	60	72	71	74	76	77	77	84	62	100	100	71	57	25711
salpingoeca_rosetta	Protists	34	79	55	72	67	65	68	91	20	100	100	73	44	7399
saprolegnia_diclina_vs20	Protists	46	80	71	69	72	81	53	86	7	100	100	76	43	8378
saprolegnia_parasitica_cbs_223_65	Protists	47	80	70	68	71	79	57	83	14	100	100	74	41	9325
theileria_equi_strain_wa	Protists	43	85	83	81	52	84	33	88	14	100	100	84	46	790
vitrella_brassicaformis_ccmp3155	Protists	36	83	52	78	65	84	47	88	18	100	100	77	37	11715
cutaneotrichosporon_oleaginosus	Fungi	34	44	67	69	61	66	82	84	44	100	100	59	51	1463
hanseniaspora_valbyensis_nrrl_y_1626	Fungi	53	72	70	69	77	78	70	78	75	100	100	66	46	244
millerozyma_farinosa_cbs_7064	Fungi	60	82	53	65	80	76	75	72	41	100	100	73	30	182
mixia_osmundae_iam_14324	Fungi	47	77	73	71	65	75	70	90	11	100	100	64	39	5281
mixia_osmundae_iam_14324_gca_000708205	Fungi	46	77	72	71	65	76	70	90	12	100	100	64	39	5072
mortierella_elongata_ag_77	Fungi	47	72	69	70	72	70	81	89	13	100	100	64	43	6120
ophiostoma_piceae_uamh_11346	Fungi	30	78	78	48	71	81	69	83	19	100	100	57	38	1781
rhodotorula_toruloides	Fungi	42	78	70	74	59	81	62	93	13	100	100	63	37	3322
rhodotorula_toruloides_np11	Fungi	39	76	71	75	59	82	62	96	8	100	100	63	37	3174
sporidiobolus_salmonicolor	Fungi	45	74	71	74	62	81	68	94	14	100	100	62	39	3115
chlamydomonas_reinhardtii	Plants	33	79	65	70	67	75	47	87	32	100	100	71	39	5995
physcomitrella_patens	Plants	41	75	70	71	71	74	56	93	13	100	100	79	40	15420

Fig. 2. Examples of diverse, preferred nucelotides even under selection for rare compositions of the splice sites. A. Heatmaps of purine (A & G) matrices of 3′ splice sites (Left) or under selection pressure for purine-rich compositions between −10 and −3 (Right). Note that some of the green- to yellow-colored positions (lower percentages) between −10 and −3 in the Left panel are still flanked by red ones (higher percentages) under this selection pressure. The maps are based on the average matrix data from 858 corresponding species/lineages. B. Examples of some of the species containing at least one position at the top of the purine-rich list (between −10 and −3, with percentages). For 5′SS information, please see S_Table IIIa.

which is associated with weak splice sites and found in A → G disease mutations (Madsen et al., 2006; Roca et al., 2008). Moreover, the G_1 at the 3′SS has only 47%, indicating that more than half of the human non-first exons start with a non-G nucleotide. Together, these differences suggest that weak splice sites are highly prevalent in the human genome, likely in favor of alternative splicing (Thanaraj and Stamm, 2003; Stamm et al., 2000; Clark and Thanaraj, 2002; Shepard et al., 2011). The matrices of another species, *D. melanogaster*, showed < 5% differences from the previously reported at these positions of both splice sites (Mount et al., 1992). The new matrices are thus similar to the previous ones in the majority of nucleotides overall but have revised the matrices of some positions substantially in the nominally complete genomes.

The average matrices from approximately 700 unique species are in Fig. 1B (for data including lineages, see S_Figs. 1a–b & 2). Overall, the highly enriched nucleotides/sequence of the splice sites are aGGTAAGT (5′SS) or $(Y)_6$N(C/t)AG(g/a)t (3′SS, upstream beyond the $(Y)_6$ is T/A-rich), similar to that from previous analyses of the human genome (Sibley et al., 2016; Zhang, 1998).

Despite the enrichment of these nucleotides, there are atypical registers of 5′SS base-pairing by U1 snRNA (Roca et al., 2012; Roca and Krainer, 2009), as well as known differences in the splice site consensus of some species, for instance between *S. cerevisiae* and humans (Sibley et al., 2016; Spingola et al., 1999; Zhang, 1998). These differences suggest sequence diversity of splice sites within and across species. Indeed, further examination of the matrices of individual species indicated that many of them were highly distinct from others. Altogether this matrix variation is reflected in the standard deviations of the matrices in Fig. 1C. The highly variable positions are mainly between −2–6 (5′SS, G_5 being the highest) and −20–1 (3′SS, C_{-3} being the highest), with > 10% deviation for some nucleotides/positions.

Analysis of the four Ensembl divisions fungi, metazoa, plants and protists indicates that their nucleotide compositions are quite different (S_Table_I). For example, at the 5′SS, both fungal and protist G_{-1} is about 50%, much less than that of metazoa and plants (~70%), while as fungal G_5 is much more (83%) than that of metazoa (70%), plants (54%) and protists (61%). Also at the 5′SS, A4 in fungi, plants and protists is < 48% but in metazoa it is 67%. At the 3′SS, fungal and protist C_{-3} is around 46% but metazoan and plant C_{-3} is much higher (60% and 67%, respectively). Moreover, the 3′SS polypyrimidine tract of fungi is not as much T-rich as the other three divisions (32–39% vs 40–62%). Together, these suggest that the splice sites have diverged at different positions among these divisions. We thus examined the diverse matrices among individual species in more detail.

2.2. The matrices of the splice sites of different species/genera/phyla are highly diverse

Detailed examination of the matrix consensus indicated that both the 5′ and 3′ splice sites could contain highly enriched distinct nucleotides in many species (S_Fig. 1 and S_Table IIa & IIb). To test if they are indeed preferred nucleotides, we selectively compiled the average matrices of splice sites enriched with less common nucleotides (Fig. 2A, and S_Table IIIa and IIIb): at least 3 Pys between −2 and 6 positions of 5′SS or at least 5 purines between nucleotides −10 and −3 of 3′SS, as we have done previously (Sohail et al., 2014). This selection identified many unusually enriched nucleotides.

Diverse splice site matrices of many species could be found in the Supplementary tables by using the SORT function of Excel for each position/nucleotide. For example, the yeast *Rhodotorula_graminis_wp1*, with a G/C-rich genome (67%) (Firrincieli et al., 2015), stood out in the selection by its C_4 (instead of A_4 at 5′SS) in 91% of the 32,901 Py-enriched 5′SSs (S_Table IIa & IIIa), the distinct C_4 was also enriched in its whole genome (89% of the 33,969 5′SSs). It also contains a highly enriched G_3 (81%) in the genome. The G_3C_4 is enriched as well in four other species/strains of the *Rhodotorula*, but not *Saccharomyces*, genus (S_Table IIa), as seen in previous reports (Illias et al., 1998; Visser et al., 2000). In *S. cerevisiae*, a T_4, instead of A_4, is enriched, as reported (Spingola et al., 1999), as well as in two other *Saccharomyces* species (S_Table IIa). In the protist *Acanthamoeba_castellanii_str_neff*, with C-rich introns (30%) and a G/C-rich (58%) genome (S_Table IIa), it has not only C_4 but also C_6 (instead of T_6) as the most abundant nucleotide at the respective position, similar to that reported (Wong et al., 1992). In another protist *Entamoeba_dispar_saw760*, T_3T_4, instead of A_3A_4, are enriched, consistent with the previous consensus (Wilihoeft et al., 2001). The enriched T_3T_4 is also seen in 8 other species/strains of the *Entamoeba* genus. The presence of the Ts could be explained by interaction with the U1 or U6 snRNAs or U1C protein (see Discussion). The C_4 matches with the G_5 of AUAC**G_5**UAC at the 5′ end of 5 diverged U1 snRNA sequences, and C_6 with the G_2 of C**G_2**UGGAC of another diverged U1 snRNA in *A. castellanii* in the Rfam database of RNA families (Kalvari et al., 2018).

More examples of these selected splice sites are given in Fig. 2B, showing the top list of species with at least one such position between −10 and −3 of the 3′SS. Under the purine-rich selection, nematode *Brugia malayi* allows for > 80% of purine nucleotides between the −10 and −7 positions but much less between −6 and −3, only 32% purines at −5 in particular. The other 12 species of the nematode phyla showed similar preferences to different degrees. Moreover, some species of protists, fungi and plants also showed nucleotide preferences at various positions under this selection.

Therefore, the splice site compositions are far more diverse than that from studying smaller groups of species, to such an extent that completely different nucleotides have been highly enriched in some species, genera or phyla. At least some of these nucleotides could be explained by co-evolved *trans*-acting factors particularly the complementary nucleotides in the U snRNAs. The preferred retention of the enriched nucleotides under sequence selection pressure suggests that some positions are highly constrained.

2.3. There are diverse constraints of splice sites among the species of different divisions

To obtain splice site constraints that can be compared among all of the species, we choose the simple 25% random distribution frequency of nucleotides as the common background (Fig. 3, n > 100 of 5′ or 3′ splice sites per species, and see Discussion & Materials and methods for details of the choice). The averages of the absolute values of matrix deviation from 25% within the highly variable regions (−2–6 of 5′ splice site, and −15–1 of 3′ splice site, excluding the 5′GT and 3′AG positions) are divided by the highest value 0.375 (e.g. 0.14/0.375 = 0.37 for the human 3′SS, Fig. 3A), resulting in the constraint index (CI) of the splice sites. Complete lists of the CI of all these species are in Supplementary Tables IVa (5′SS) and IVb (3′SS).

According to the CI, the histogram distribution of the constraints of 5′ and 3′ splice sites of the four divisions are shown in Fig. 3B–C. The major peak of the metazoa 5′SS CI is at higher values and with a narrower span than the other three divisions, whose peaks partially overlap and protists' spans the widest range.

The CI of the 3′ splice site showed more distinct distribution peaks among the four divisions. The peak with the lowest constraint is in fungi (at 0.16), while it is higher in plants (0.32 and 0.36) and metazoa (0.38). The protists again have multiple peaks that span or go beyond the ranges of the other divisions.

Also indicated in the histograms are the CI positions of the typical model organisms. It is interesting to note that *S. cerevisiae* has CIs (0.64 for 5′SS, and 0.31 for 3′SS) that are distinctively higher than the peaks (0.46 for 5′SS, and 0.16 for 3′SS) of all of these fungal species, suggesting that its splice site is not typical of most fungal species in this regard.

Of the metazoan species, the arthropod and nematode groups showed a wider distribution than vertebrates (S_Fig. 3), indicating that

A

d	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	n =	Ave_d(-15 - 1)	Ave_d(-15 - 1)_ACGT	CI_3SS_Human
A	0.12	0.13	0.14	0.15	0.16	0.16	0.14	0.14	0.13	0.15	0.15	0.01	0.18	0.75	0.25	0.02	0.00	335102	0.13		
C	0.03	0.02	0.03	0.04	0.01	0.04	0.04	0.07	0.08	0.09	0.05	0.03	0.39	0.25	0.25	0.11	0.06	335102	0.07		
G	0.12	0.13	0.13	0.14	0.14	0.14	0.13	0.14	0.15	0.18	0.18	0.05	0.24	0.25	0.75	0.22	0.05	335102	0.15		
T	0.21	0.23	0.24	0.26	0.29	0.26	0.23	0.21	0.20	0.25	0.28	0.02	0.03	0.25	0.25	0.14	0.12	335102	0.20		
AVE_d_ACGT	0.12	0.13	0.14	0.15	0.15	0.15	0.14	0.14	0.14	0.17	0.17	0.03	0.21	0.38	0.38	0.12	0.06	335102	0.14	0.14	0.37

(= 0.14/0.375)

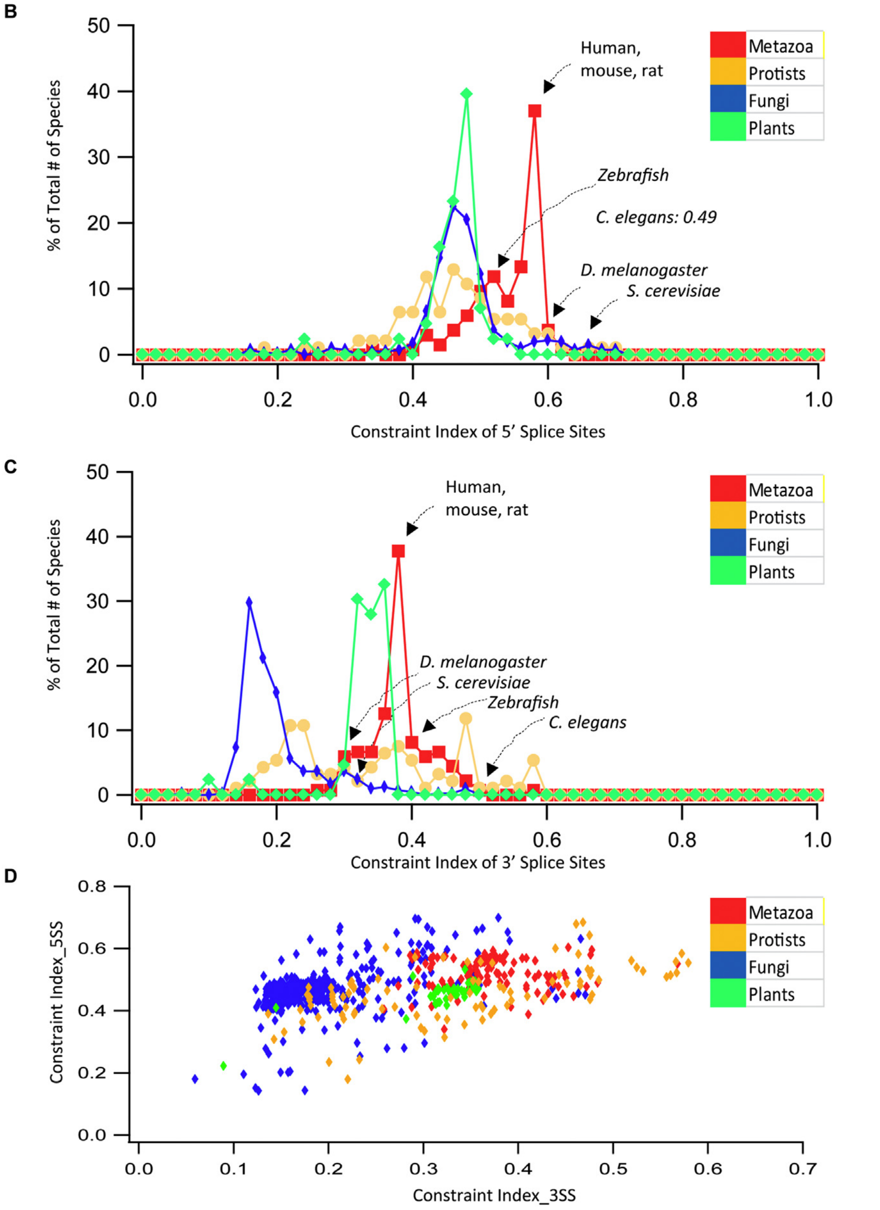


Fig. 3. The diverse constraint indexes of the splice sites of different species/divisions. A. An example for calculating the constraint index of the splice site of a species. For the human 3′ splice site, the deviation (d) of each position/nucleotide is the absolute value after taking 25% off its matrix percentage. The average of the four nucleotides/positions excluding 3′AG is then divided by 0.375, the average of a non-variable nucleotide/position or sequence, to obtain a constraint index (CI) of the 3′ splice site of the species. The CI of the 5′ splice site is calculated (between −2 to 6) similarly excluding the 5′AG. B–C. Histograms of the 5′SS (B) and 3′SS (C) constraint indexes of the four Ensembl divisions, with the positions of some of the common model organisms indicated. D. Dot plot of the 5′SS versus 3′SS constraint indexes of different species among the four Ensembl divisions.

the two phyla have evolved more diverse constraints than vertebrates.

To see if the constraint indexes of the two splice sites correlate with each other, we plotted the corresponding indexes of each species in Fig. 3D. Overall, there is a weak positive correlation between the constraints of the two splice sites (Pearson coefficient = 0.45), and a similar extent within fungi (0.45). The correlation appears much stronger within plants (0.78) and protists (0.56), but no correlation within metazoans (−0.02). This supports that the constraints of the two splice sites co-evolved overall, and more so within plants and protists but not within metazoans, where it is mainly the 3′SS that has been undergoing changes among different species.

2.4. The nucleotide and splice site constraints are associated with the abundance of alternative splicing in different species

Constrained nucleotides at the splice sites are likely preferred by the spliceosome components for constitutive splicing. Thus, a natural inference is that the non-constrained nucleotides at the same positions are more predictive of alternative splicing, similarly as mutations within or deviations from the consensus sequences of splice sites in aberrant or alternative splicing of particular gene(s) (Tazi et al., 2009; Feng and Xie, 2013; Thanaraj and Stamm, 2003). To verify this point using the constraint data at the transcriptome level, we examined the transcriptome-wide relationship of different 3′SS nucleotides with the corresponding percentages of alternative exons in *C. elegans* and lamprey.

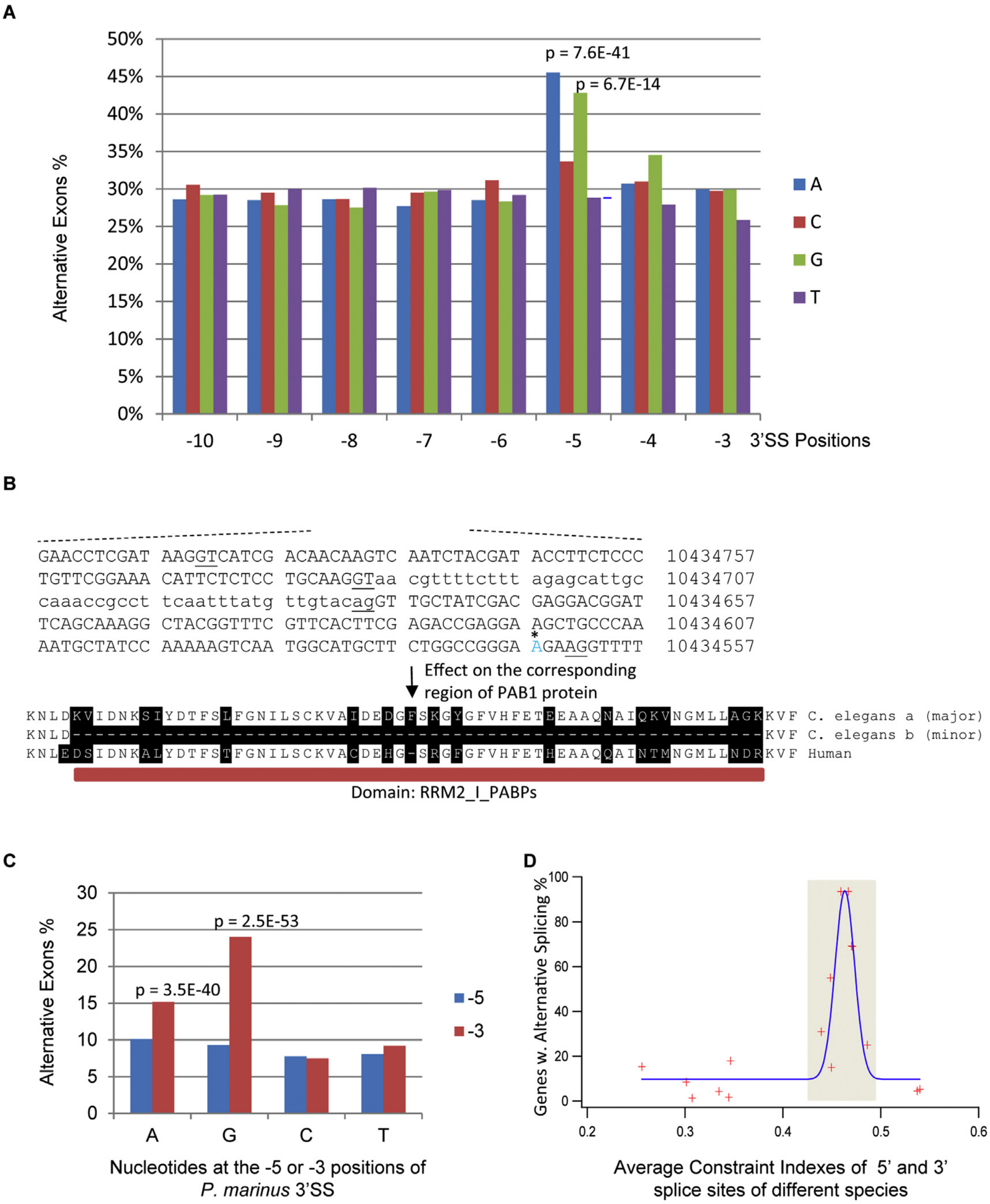


Fig. 4. Relationship of the constraint index with the abundance of alternative splicing in different species. A. An example of a constrained position associated with increased level of alternative splicing when a rare nucleotide(s) is present. Shown are the percent alternative exons associated with each nucleotide/position of the 3′SS of *C. elegans*. B. An example of the consequence of the nematode A_{-5}-associated alternative splicing: a minor *Pab1* (Polyadenylate-binding protein, *Pab1*) variant resulting in a protein isoform lacking a whole RRM domain. RRM2_I_PABPs: RNA recognition motif 2 found in type I polyadenylate-binding proteins. C. Another example of a highly constrained nucleotide/position whose change is associated with increased presence of alternative exons, position −3 of the 3′SS with extremely constrained Cs in the lamprey *P. marinus*. The percentages of alternative exons were counted based on the 'Constitutive' exon data in Biomart for A and C, except that results from DEXseq analysis of additional data from male or hermaphrodites were included in A. D. Correlation of the percentages of genes containing alternative exons with the constraint indexes of the same species, fitted in a Gaussian curve (n = 16 species). The species are: *A. flavus*, *A. oryzae*, *C. elegans*, *C. intestinalis*, *C. neoformans*, *D. rerio*, *D. melanogaster*, *F. graminearum*, *G. gallus*, *H. sapiens*, *M. musculus*, *P. falciparum*, *R. norvegicus*, *S. commune*, *T. melanosporum and T. thermophila*. Inclusion of data from three plant species also gave a bell curve but more scattered, suggesting that the abundance of alternative splicing in plants is not as well associated with the constraint indexes as in these species.

The result showed that when the most constrained T_{-5} in *C. elegans* was replaced with the other nucleotides, particularly A and G, the percentages of immediate downstream alternative exons increased significantly (Fig. 4A). An example of such an effect on splicing is shown in the *pab1* gene in Fig. 4B. The alternative use of a weak minor 3′SS causes truncation of the RRM2 domain of the Pab1 protein.

At the 3′SS of the lamprey *P. marinus*, the −3 position is highly enriched for Py nucleotides (C_{-3}: 80%, at the top 4% of 1076 species/lineages, plus 14% as T, S_Table IIb), even under the purine-selection pressure for purine-rich sequences within the −10 to −3 region (68% as C plus 9% as T, S_Table IIIb). When the C_{-3} is replaced with a purine, particularly G, the percentage of its immediate downstream alternative exons increased significantly, while as the other position −5, which is also Py-rich (77%), is not as sensitive to this change in the same species.

Together, these support that when a highly constrained nucleotide is replaced with uncommon ones, the possibility for alternative splicing increases significantly at a transcriptome scale.

While changing the constrained nucleotides predicts increased chances for alternative splicing, it is also reasonable that the more constrained (from randomness) a position is, the less chance the same position could have a non-constrained nucleotide. Therefore, a balance between the constraint level and alternative splicing should exist. To test this relationship, we calculated the average constraint indexes of the 5′ and 3′ splice sites of each species. Of a number of non-plant species with known percentages of genes containing alternative splicing from transcriptome-wide analysis (Kim et al., 2007; Aanes et al., 2011; Daines et al., 2011; Ramani et al., 2011; Xiong et al., 2012; Sorber et al., 2011; Loftus et al., 2005; McGuire et al., 2008; Zhao et al., 2013; Gehrmann et al., 2016; Tisserant et al., 2011; Wang et al., 2010), their abundance of alternative splicing appears to correlate overall with the constraints in a bell-shaped curve (Fig. 4D). An equation fitted using Gaussian function for this relationship is:

$$P_{AS} = 9.6752 + 84.216*e^{\wedge}(-(((CI - 0.4633)/0.014082)^{\wedge}2))$$

where P_{AS} is the percentage of genes containing alternative splicing, and CI is the average constraint index of the splice sites of a species.

The peak area covers the constraints (0.43–0.5) of 77% of metazoa, in contrast to only 7% of fungi, 18% of protists and 2% of plants, among the 681 Ensembl species examined here. Therefore, the CI could distinctively separate most of the metazoan species from the others. This is consistent with the increased level of alternative splicing in metazoa (Nilsen and Graveley, 2010; Maniatis and Tasic, 2002).

Since the genomes of different species have different nucleotide compositions with some nucleotides particularly enriched instead of being always 25% each, we have thus also tested using the current genome compositions of each species as the background (S_Tables V & VI). This test reduced the effect of exceptionally enriched nucleotides in some species including the A/T-rich (80%) protist *P. falciparum*. Its correlation with the abundance of alternative splicing is in a partial bell curve with the left side heavily clustered with low constraint species but the right side with only three samples (Supplementary Fig. 3C). Although the right side also shows a similar trend to decrease as the constraints go further up, this half will need more samples to confirm if the trend still holds true in the future.

From the bell curve, it appears that human beings have evolved a near-optimal constraint of the splice sites for a much higher level of alternative splicing. The other species with either higher or lower constraints show less alternative splicing than humans. If true, this suggests that there is an optimal constraint (from randomness or the genome background) of the splice sites to allow for specific splice signals while still flexible enough for the high abundance of alternative splicing in a species. Moreover, this also suggests that the abundance of alternative splicing is mainly dependent on the average index of the splice site constraints of the species.

2.5. *A web interface for searching the matrices and constraints of specific species/sequences*

Besides the Supplementary Tables in the Excel file, we also designed a web interface to output the matrices and constraints of a particular species upon input of the species name. Users could also input their sequence (or mutant) of choice to check the nucleotide frequency of this sequence in a particular species. A test site for this is currently at: https://intronrepa.herokuapp.com/intronrepa/, and will be put at a permanent site later with update information at the Xie lab website at http://home.cc.umanitoba.ca/~xiej/index.html.

2.6. *Updates*

Ensembl Genomes 36 and Ensembl 90 have been released in the summer of 2017. We have also analyzed this release and found that it added < 8% of splice sites over the previous version. We examined their matrices the same way and obtained similar results with < 1% change on average of the nucleotide percentages in the highly variable regions. Future releases will be analyzed and updated at the web interface and the Xie lab website at http://home.cc.umanitoba.ca/~xiej/index.html.

3. Discussion

3.1. *A reference database for the splice site matrices/constraints of > 1000 species/lineages*

The matrix and constraint database of the annotated eukaryotic species has covered the 5′GT, 3′ polypyrimidine tract and 3′AG. It will be useful at least in the studies on genes or gene functions: as a reference for making effective mutations close to or within the splice sites in splicing assays, for assessing the strength of a particular splice site or mutant sequence in a species, for developing species-specific algorithm for more accurate prediction of unknown exons, for comparing splice sites and their evolutionary changes among different species, or for exploring the evolution of splice signals.

For instance, to make mutations within the 3′ splice site to assess the effect of potential regulatory elements there on splicing (Xie and Black, 2001; Xie et al., 2005), one would want the mutation itself not to abolish splicing completely. Particularly the −5 to −15 positions are often Py-rich and sensitive to changes to purine nucleotides in mammalian species, which could abolish splicing and is often avoided. However, in some other species like the protist *Thecamonas trahens* (S_Table IIb), the genome is G-rich (32%, the top 6th) and G is tolerated (> 30%) even in the Py region. Here mutation to G is expected to still have splicing products in reporter assays. Therefore, the splice site matrices will be helpful for designing appropriate mutations for such experiments in specific species.

The highly diverse splice site matrices among the different species/genera/phyla suggest that a prediction algorithm developed based on the splice sites of a species or a group of species will miss a substantial number of sites that do not score high enough in other species/phyla. It would thus be necessary to develop more specific algorithms for improved accuracy in prediction and annotation of novel splice sites or exons in some species/phyla in the future. For this purpose, the matrix database will be a useful reference.

The diverse matrices also provide a resource for exploring potentially novel spliceosome components that have diverged during evolution, such as the novel 5′ end of U1 snRNA for the C_4 and C_6 of 5′SS in *A. castellanii* or potentially in the *R. graminis* as well, or for novel interactions if any.

3.2. *The splice site constraints and evolvement of alternative splicing*

Alternative splicing is controlled by both *cis*-acting elements and *trans*-acting factors (Black, 2003; Lee and Rio, 2015; Chen and Manley, 2009). The *cis*-acting elements appear to play an important role in its species differences (Sohail and Xie, 2015; Barbosa-Morais et al., 2012). Here our analysis of 16 species (Fig. 4D) suggests that splice site constraints determine the abundance of alternative splicing in a species.

The molecular basis of this relationship might be related to the preferred interactions of the highly constrained nucleotides with a spliceosome component(s) in the species/genera/phyla. For the constrained T_{-5} of nematode 3′SS, its corresponding U_5 is strongly preferred by U2AF65 of *C. elegans* in competitive binding assays (Hollins et al., 2005), therefore, its change to another nucleotide would likely reduce this preferred interaction and probability for spliceosome assembly. As for the distinct T_4 of *S. cerevisiae* 5′ splice site, its corresponding U_4 pairs with the pseudo-uridine (ψ^5) of U1 snRNA (Libri et al., 2002), interacts with the U1 snRNP protein U1C (Du and Rosbash, 2002), and could also pair with the A_{49} of U6 snRNA (Lesser and Guthrie, 1993; Kandels-Lewis and Seraphin, 1993; Yan et al., 2016). The T_3T_4 in the *Entamoeba_dispar* could be explained by interaction with the $\psi^5\psi^6$ of U1 snRNA (Libri et al., 2002; Tan et al., 2016), and/or the $A^{44}A^{45}$ of U6 RNA (Lesser and Guthrie, 1993; Sawa and Abelson, 1992), if the snRNA motifs are conserved in this species. The highly constrained nucleotides are thus likely for preferred interactions with a spliceosome component(s) that have diverged properties in the corresponding species (e.g. U2AF65 of *C. elegans*). In contrast, a rare nucleotide might have reduced interaction with splicing factors, therefore decreasing the probability for spliceosome assembly resulting in alternative splicing.

The current data for the relationship between the constraints and alternative splicing appear to correlate relatively well in the bell curve (Fig. 4D & S_Fig. 3C), though more species data are still needed for improved accuracy in the future. The curve suggests that the constraint index of the splice sites of a species could predict the threshold of the abundance of alternative splicing and thus transcriptome complexity of different species or a group of them, though the inclusion level of specific alternative exons could still be changed by regulatory elements/factors. For instance, the average species constraint indexes of fungi peak at 0.24 (Fig. 3C), far below the 0.44–0.49 range for the peak of the abundance of alternative splicing (Fig. 4D). Consistently, the highest abundance of alternative splicing among fungal species is only 17% of genes in *S. commune*, with a CI index of 0.346 *(Gehrmann et al., 2016)*. In contrast, most of the vertebrate species are within the peak range of the bell curve and are expected to have much higher levels of alternative splicing, which is consistent with the species comparison data on the evolutionary changes of alternative splicing (Barbosa-Morais et al., 2012).

3.3. Considerations for future improvements of the algorithms for constraint calculations and cross-species comparison

Of the published different methods for assessing the strengths/constraints of sequence motifs, maximum entropy modeling (MEM) and information content are based on Shannon entropy in units of bits (log2) (Yeo and Burge, 2004; Fields, 1990; Schneider et al., 1986). The MEM also considers dependencies on non-adjacent positions with iterative training/scaling using large numbers of samples and selected decoy sequences (Yeo and Burge, 2004). Unfortunately, many positions have 0% specific nucleotides (e.g. 5′SS C_5 in many strains of yeast *S. cerevisiae*). These in yeast or other species cannot be measured using the logarithmic calculations. Moreover, practically storing and handling of the complete genomes and sequences for training/scaling of > 1000 species/lineages using MEM is a challenge. We have thus calculated the splice site constraints based on the 25% random distribution of nucleotides or the species nucleotide composition as a background (Figs. 3 & 4D, or S_Fig. 3C).

The constraint calculation using either 25% or genome background also has limitations. The 25% is the frequency of random distribution of nucleotides. Its use as a common background reflects the deviation of the splice sites from randomness but does not reflect the difference of real genome compositions among different species. As for using the genome compositions as backgrounds, their accuracy depends on the completeness of the individual genomes to reflect the best approximation to the true value of the whole genome. Moreover, it is also not clear to what extent the true constraints at different positions of the splice sites are due to the whole genome background of each species, making it difficult for accurate assignment of the weights/penalties for constraint calculation.

There are also various other known constraints including the atypical recognition of 5′SS and pseudouridylation of U1 and other snRNAs (Roca et al., 2012; Roca and Krainer, 2009; Reddy et al., 1981; Hudson et al., 2013; Yu et al., 2011). Some of the pseudo-uridylations are inducible under nutrient-deprivation or heat-shock stress (Wu et al., 2011). Adding further to the complexity in accurate determination of the constraints, a small group of the GT/AG introns can also be spliced by the minor spliceosome with different consensus sequences of the splice sites (Dietrich et al., 1997; Turunen et al., 2013). To what extent these constraints should be applied is a challenge before their species distribution and effects are clear to the about 700 species. Accurate assessment of the contributions to splicing by these constraints and by the genome background should help future improvement of a ‘universal’ algorithm for the constraints of the highly diverse splice sites.

4. Materials and methods

4.1. Genome data

The GenBank-format files of the genomes of all the species examined here were downloaded from the release 88 or Genome release 35 of the Ensembl database, of which the transcripts are based on experimental evidence (Aken et al., 2016).

4.2. Calculation of matrices and constraints

For matrices, the 5′ GT(−5–+50) or 3′ AG(−50–+2) splice sites were counted for their nucleotide compositions (percentages) at each position according to the intron coordinates in the annotated GenBank files of each species/lineage/strain, using Python scripts.

For constraints, we used a simple calculation:

$$d = [\Sigma_{x=A,C,G,T}(|\,P(x)-25\%|)]/4$$

to reflect the absolute average deviation of a position from random nucleotide distribution (d: deviation of a nucleotide/position from random distribution, P: observed frequency of nucleotide x at a position of the splice sites in the annotated genome, and 25% is the frequency of random distribution as used by Fields (Mount et al., 1992; Fields, 1990)). Accordingly, a position with an invariable nucleotide should have an average deviation of 0.375 from random distribution of all four nucleotides. For constraint indexes of the splice sites of a species, the average deviation of all the counted variable positions is divided by 0.375, resulting in a constraint index between 0 and 1.0. Using 25% thus allows comparison of the splice site constraints of all species/lineages in the same background that does not change with the version or completeness of the Ensembl genomes.

Constraints using the genome nucleotide compositions of different species as backgrounds were calculated similarly except that GT/AG positions were counted and the average deviations were not normalized to 0.375.

4.3. Analysis of alternative splicing

The alternative exons were counted as the ‘non-constitutive’ exons based on the information on ‘Constitutive’ in BioMart (Aken et al., 2016), or in combination with our analysis of *C. elegans* male and female (hermaphrodite) RNA-seq data deposited in the NCBI SRA database (Kramer et al., 2016), using DEXseq (Anders et al., 2012).

4.4. Statistical analysis

Statistical analysis was done with hypergeometric test for the different percentages of alternative exons.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2018.03.031.

Acknowledgements

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2016-06004) and a Manitoba Research Chair fund to J.X. We thank the anonymous reviewers for helpful suggestions.

References

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G., Lim, A.Y., Hajan, H.S., Collas, P., Bourque, G., Gong, Z., Korzh, V., Alestrom, P., Mathavan, S., 2011. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res. 21, 1328–1338.

Abril, J.F., Castelo, R., Guigo, R., 2005. Comparison of splice sites in mammals and chicken. Genome Res. 15, 111–119.

B.L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. Garcia Giron, T. Hourlier, K. Howe, A. Kahari, F. Kokocinski, F.J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y.A. Tang, J.H. Vogel, S. White, A. Zadissa, P. Flicek, and S.M. Searle, The Ensembl gene annotation system. Database (Oxford) 2016 (2016).

Anders, S., Reyes, A., Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. Genome Res. 22, 2008–2017.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C.M., Wilson, M.D., Kim, P.M., Odom, D.T., Frey, B.J., Blencowe, B.J., 2012. The evolutionary landscape of alternative splicing in vertebrate species. Science 338, 1587–1593.

Black, D.L., 2003. Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. 72, 291–336.

Burge, C.B., Padgett, R.A., Sharp, P.A., 1998. Evolutionary fates and origins of U12-type introns. Mol. Cell 2, 773–785.

Burset, M., Seledtsov, I.A., Solovyev, V.V., 2001. SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res. 29, 255–259.

Chen, M., Manley, J.L., 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat. Rev. Mol. Cell Biol. 10, 741–754.

Clark, F., Thanaraj, T.A., 2002. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum. Mol. Genet. 11, 451–464.

Daguenet, E., Dujardin, G., Valcarcel, J., 2015. The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. EMBO Rep. 16, 1640–1655.

Daines, B., Wang, H., Wang, L., Li, Y., Han, Y., Emmert, D., Gelbart, W., Wang, X., Li, W., Gibbs, R., Chen, R., 2011. The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. Genome Res. 21, 315–324.

Dietrich, R.C., Incorvaia, R., Padgett, R.A., 1997. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol. Cell 1, 151–160.

Dou, Y., Fox-Walsh, K.L., Baldi, P.F., Hertel, K.J., 2006. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. RNA 12, 2047–2056.

Du, H., Rosbash, M., 2002. The U1 snRNP protein U1C recognizes the 5′ splice site in the absence of base pairing. Nature 419, 86–90.

Feng, D., Xie, J., 2013. Aberrant splicing in neurological diseases. Wiley Interdiscip. Rev. RNA 4, 631–649.

Fields, C., 1990. Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. Nucleic Acids Res. 18, 1509–1512.

Firrincieli, A., Otillar, R., Salamov, A., Schmutz, J., Khan, Z., Redman, R.S., Fleck, N.D., Lindquist, E., Grigoriev, I.V., Doty, S.L., 2015. Genome sequence of the plant growth promoting endophytic yeast *Rhodotorula graminis* WP1. Front. Microbiol. 6, 978.

Freund, M., Hicks, M.J., Konermann, C., Otte, M., Hertel, K.J., Schaal, H., 2005. Extended base pair complementarity between U1 snRNA and the 5′ splice site does not inhibit splicing in higher eukaryotes, but rather increases 5′ splice site recognition. Nucleic Acids Res. 33, 5112–5119.

Garg, K., Green, P., 2007. Differing patterns of selection in alternative and constitutive splice sites. Genome Res. 17, 1015–1022.

Gehrmann, T., Pelkmans, J.F., Lugones, L.G., Wosten, H.A., Abeel, T., Reinders, M.J., 2016. Schizophyllum commune has an extensive and functional alternative splicing repertoire. Sci. Rep. 6, 33640.

Hall, S.L., Padgett, R.A., 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J. Mol. Biol. 239, 357–365.

Hollins, C., Zorio, D.A., MacMorris, M., Blumenthal, T., 2005. U2AF binding selects for the high conservation of the *C. elegans* 3′ splice site. RNA 11, 248–253.

Hudson, G.A., Bloomingdale, R.J., Znosko, B.M., 2013. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. RNA 19, 1474–1482.

Illias, R.M., Sinclair, R., Robertson, D., Neu, A., Chapman, S.K., Reid, G.A., 1998. L-Mandelate dehydrogenase from *Rhodotorula graminis*: cloning, sequencing and kinetic characterization of the recombinant enzyme and its independently expressed flavin domain. Biochem. J. 333 (Pt 1), 107–115.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., Petrov, A.I., 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46, D335–D342.

Kandels-Lewis, S., Seraphin, B., 1993. Involvement of U6 snRNA in 5′ splice site selection. Science 262, 2035–2039.

Kim, E., Magen, A., Ast, G., 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35, 125–131.

Kramer, M., Rao, P., Ercan, S., 2016. Untangling the contributions of sex-specific gene regulation and X-chromosome dosage to sex-biased gene expression in *Caenorhabditis elegans*. Genetics 204, 355–369.

Lee, Y., Rio, D.C., 2015. Mechanisms and regulation of alternative pre-mRNA splicing. Annu. Rev. Biochem. 84, 291–323.

Lesser, C.F., Guthrie, C., 1993. Mutations in U6 snRNA that alter splice site specificity: implications for the active site. Science 262, 1982–1988.

Levine, A., Durbin, R., 2001. A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res. 29, 4006–4013.

Libri, D., Duconge, F., Levy, L., Vinauger, M., 2002. A role for the Psi-U mismatch in the recognition of the 5′ splice site of yeast introns by the U1 small nuclear ribonucleoprotein particle. J. Biol. Chem. 277, 18173–18181.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M., Hyman, R.W., 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science 307, 1321–1324.

Lorkovic, Z.J., Wieczorek Kirk, D.A., Lambermon, M.H., Filipowicz, W., 2000. Pre-mRNA splicing in higher plants. Trends Plant Sci. 5, 160–167.

Madsen, P.P., Kibaek, M., Roca, X., Sachidanandam, R., Krainer, A.R., Christensen, E., Steiner, R.D., Gibson, K.M., Corydon, T.J., Knudsen, I., Wanders, R.J., Ruiter, J.P., Gregersen, N., Andresen, B.S., 2006. Short/branched-chain acyl-CoA dehydrogenase deficiency due to an IVS3 + 3A > G mutation that causes exon skipping. Hum. Genet. 118, 680–690.

Maniatis, T., Tasic, B., 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418, 236–243.

McGuire, A.M., Pearson, M.D., Neafsey, D.E., Galagan, J.E., 2008. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol. 9, R50.

Moore, M.J., 2000. Intron recognition comes of AGe. Nat. Struct. Biol. 7, 14–16.

Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O., Fields, C., 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucleic Acids Res. 20, 4255–4262.

Nilsen, T.W., Graveley, B.R., 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463.

Ramani, A.K., Calarco, J.A., Pan, Q., Mavandadi, S., Wang, Y., Nelson, A.C., Lee, L.J., Morris, Q., Blencowe, B.J., Zhen, M., Fraser, A.G., 2011. Genome-wide analysis of alternative splicing in *Caenorhabditis elegans*. Genome Res. 21, 342–348.

Reddy, R., Henning, D., Busch, H., 1981. Pseudouridine residues in the 5′-terminus of uridine-rich nuclear RNA I (U1 RNA). Biochem. Biophys. Res. Commun. 98, 1076–1083.

Roca, X., Krainer, A.R., 2009. Recognition of atypical 5′ splice sites by shifted base-pairing to U1 snRNA. Nat. Struct. Mol. Biol. 16, 176–182.

Roca, X., Olson, A.J., Rao, A.R., Enerly, E., Kristensen, V.N., Borresen-Dale, A.L., Andresen, B.S., Krainer, A.R., Sachidanandam, R., 2008. Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics. Genome Res. 18, 77–87.

Roca, X., Akerman, M., Gaus, H., Berdeja, A., Bennett, C.F., Krainer, A.R., 2012. Widespread recognition of 5′ splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. Genes Dev. 26, 1098–1109.

Rogozin, I.B., Milanesi, L., 1997. Analysis of donor splice sites in different eukaryotic organisms. J. Mol. Evol. 45, 50–59.

Sawa, H., Abelson, J., 1992. Evidence for a base-pairing interaction between U6 small nuclear RNA and 5′ splice site during the splicing reaction in yeast. Proc. Natl. Acad. Sci. U. S. A. 89, 11269–11273.

Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A., 1986. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188, 415–431.

Scotti, M.M., Swanson, M.S., 2016. RNA mis-splicing in disease. Nat. Rev. Genet. 17, 19–32.

Shepard, P.J., Choi, E.A., Busch, A., Hertel, K.J., 2011. Efficient internal exon recognition depends on near equal contributions from the 3′ and 5′ splice sites. Nucleic Acids Res. 39, 8928–8937.

Shi, Y., 2017. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat. Rev. Mol. Cell Biol. 18, 655–670.

Sibley, C.R., Blazquez, L., Ule, J., 2016. Lessons from non-canonical splicing. Nat. Rev. Genet. 17, 407–421.

Sohail, M., Xie, J., 2015. Evolutionary emergence of a novel splice variant with an opposite effect on the cell cycle. Mol. Cell. Biol. 35, 2203–2214.

Sohail, M., Cao, W., Mahmood, N., Myschyshyn, M., Hong, S.P., Xie, J., 2014. Evolutionarily emerged G tracts between the polypyrimidine tract and 3′ AG are splicing silencers enriched in genes involved in cancer. BMC Genomics 15, 1143.

Sorber, K., Dimon, M.T., DeRisi, J.L., 2011. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res. 39, 3820–3835.

Spingola, M., Grate, L., Haussler, D., Ares Jr., M., 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. RNA 5, 221–234.

Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., Zhang, M.Q., 2000. An alternative-exon database and its statistical analysis. DNA Cell Biol. 19, 739–756.

Szczesniak, M.W., Kabza, M., Pokrzywa, R., Gudys, A., Makalowska, I., 2013. ERISdb: a database of plant splice sites and splicing signals. Plant Cell Physiol. 54, e10.

Tan, J., Ho, J.X., Zhong, Z., Luo, S., Chen, G., Roca, X., 2016. Noncanonical registers and base pairs in human 5′ splice-site selection. Nucleic Acids Res. 44, 3908–3921.

Tazi, J., Bakkour, N., Stamm, S., 2009. Alternative splicing and disease. Biochim. Biophys. Acta 1792, 14–26.

Thanaraj, T.A., Stamm, S., 2003. Prediction and statistical analysis of alternatively spliced exons. Prog. Mol. Subcell. Biol. 31, 1–31.

Tisserant, E., Da Silva, C., Kohler, A., Morin, E., Wincker, P., Martin, F., 2011. Deep RNA sequencing improved the structural annotation of the *Tuber melanosporum* transcriptome. New Phytol. 189, 883–891.

Turunen, J.J., Niemela, E.H., Verma, B., Frilander, M.J., 2013. The significant other: splicing by the minor spliceosome. Wiley Interdiscip. Rev. RNA 4, 61–76.

Verma, B., Akinyi, M.V., Norppa, A.J., Frilander, M.J., 2017. Minor spliceosome and disease. Semin. Cell Dev. Biol. http://dx.doi.org/10.1016/j.semcdb.2017.09.036. pii: S1084-9521(17)30123-4, [Epub ahead of print].

Visser, H., Vreugdenhil, S., de Bont, J.A., Verdoes, J.C., 2000. Cloning and characterization of an epoxide hydrolase-encoding gene from *Rhodotorula glutinis*. Appl. Microbiol. Biotechnol. 53, 415–419.

Wang, B., Guo, G., Wang, C., Lin, Y., Wang, X., Zhao, M., Guo, Y., He, M., Zhang, Y., Pan, L., 2010. Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. Nucleic Acids Res. 38, 5075–5087.

Wilihoeft, U., Campos-Gongora, E., Touzni, S., Bruchhaus, I., Tannich, E., 2001. Introns of *Entamoeba histolytica* and *Entamoeba dispar*. Protist 152, 149–156.

Will, C.L., Luhrmann, R., 2011. Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. 3.

Wong, J.M., Liu, F., Bateman, E., 1992. Isolation of genomic DNA encoding transcription factor TFIID from *Acanthamoeba castellanii*: characterization of the promoter. Nucleic Acids Res. 20, 4817–4824.

Wu, Q., Krainer, A.R., 1999. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. Mol. Cell. Biol. 19, 3225–3236.

Wu, G., Xiao, M., Yang, C., Yu, Y.T., 2011. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNP. EMBO J. 30, 79–89.

Xie, J.Y., Black, D.L., 2001. A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. Nature 410, 936–939.

Xie, J.Y., Jan, C., Stoilov, P., Park, J., Black, D.L., 2005. A consensus CaMK IV-responsive RNA sequence mediates regulation of alternative exons in neurons. RNA 11, 1825–1834.

Xiong, J., Lu, X., Zhou, Z., Chang, Y., Yuan, D., Tian, M., Zhou, Z., Wang, L., Fu, C., Orias, E., Miao, W., 2012. Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using Deep RNA sequencing. PLoS One 7, e30630.

Yan, C., Wan, R., Bai, R., Huang, G., Shi, Y., 2016. Structure of a yeast activated spliceosome at 3.5 A resolution. Science 353, 904–911.

Yeo, G., Burge, C.B., 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. 11, 377–394.

Yu, A.T., Ge, J., Yu, Y.T., 2011. Pseudouridines in spliceosomal snRNAs. Protein Cell 2, 712–725.

Zhang, M.Q., 1998. Statistical features of human exons and their flanking regions. Hum. Mol. Genet. 7, 919–932.

Zhao, C., Waalwijk, C., de Wit, P.J., Tang, D., van der Lee, T., 2013. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. BMC Genomics 14, 21.