Published online 5 April 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 14* **5837–5844**
doi:10.1093/nar/gkr203

Cryptic splice sites and split genes

Yuri Kapustin[1,*], Elcie Chan[2], Rupa Sarkar[2], Frederick Wong[2], Igor Vorechovsky[3], Robert M. Winston[2], Tatiana Tatusova[1] and Nick J. Dibb[2,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20814, USA, [2]Institute of Reproductive and Developmental Biology, Imperial College London, Du Cane Road, London W12 ONN and [3]Division of Human Genetics, University of Southampton Medical School, Southampton SO16 6YD, UK

Received October 6, 2010; Revised March 17, 2011; Accepted March 22, 2011

ABSTRACT

We describe a new program called cryptic splice finder (CSF) that can reliably identify cryptic splice sites (css), so providing a useful tool to help investigate splicing mutations in genetic disease. We report that many css are not entirely dormant and are often already active at low levels in normal genes prior to their enhancement in genetic disease. We also report a fascinating correlation between the positions of css and introns, whereby css within the exons of one species frequently match the exact position of introns in equivalent genes from another species. These results strongly indicate that many introns were inserted into css during evolution and they also imply that the splicing information that lies outside some introns can be independently recognized by the splicing machinery and was in place prior to intron insertion. This indicates that non-intronic splicing information had a key role in shaping the split structure of eukaryote genes.

INTRODUCTION

Eukaryotic genomes contain large numbers of splice sites, known as cryptic splice sites (css), which are generally held to be disadvantageous sites that are dormant or used only at low levels unless activated by mutation of nearby authentic or advantageous splice sites (1,2). Once activated, css may be used extremely efficiently, resulting in a wide range of genetic disease (3–5). It is generally accepted that css are suppressed by nearby stronger splice sites and that splice site selection can be viewed as a competition between the various potential splice sites in a pre-mRNA for the splicing machinery (1,2).

For genes with many introns it is suspected that up to 50% of mutations that cause disease do so by affecting splicing, either through the activation of css, exon skipping or disruption of alternative splicing (4–7). Css are found in exons as well as introns and their recognition by the splicing machinery is similar to splice site recognition in general and is dependent upon information both at the splice site and outside this region at enhancer and silencer sequences (8–10).

It is important to be able to predict the positions of css that might be activated in genetic disease and a number of DNA-sequence scanning programs have been developed for this purpose. Such programs are often highly informative but are handicapped by the complex nature of the nucleotide information that is required to define a splice site (4,8,11,12).

Our previous work indicates a connection between css and introns. We identified a small number of css in the exon regions of actin genes by experiment and discovered that eight out of nine of these exonic css sites exactly match the positions of introns in actin genes from other species, which led us to conclude that these particular actin introns were inserted into css during evolution (13,14). This finding may help to explain why and how eukaryotes acquired introns; however, it is important to establish if our results for the actin gene family are generally applicable.

We have been unable to identify a DNA-scanning program that reveals the same strong correlation between predicted actin css and intron positions that we observed through experiments (13). However, this is probably because DNA-scanning programs were not designed specifically for this purpose and because of the difficulties such programs face in distinguishing between css and false-positive non-functional splice sites (12). Here, we describe a program called cryptic splice finder (CSF) that can reliably identify css by EST-to-genomic alignment. It does this by identifying transcripts that have been generated through the low level use of css by normal genes. Unlike the scanning programs, CSF cannot predict the position of splice sites that are created *de novo* by gene mutation. However, this program provides a useful complementary

*To whom correspondence should be addressed. Tel: +212 813 8774; Fax: +212 826 8280; Email: yurikapustin@gmail.com
Correspondence may also be addressed to Nick J. Dibb. Tel: +020 7594 2103; Fax: +020 759 42154; Email: n.dibb@imperial.ac.uk

resource for studies of genetic disease and it also enabled us to establish that there is a strong and general correlation between the positioning of css and of introns. The evolutionary implications of this finding are discussed.

MATERIALS AND METHODS

CSF

CSF (http://www.ncbi.nlm.nih.gov/IEB/Research/csf) predicts css based on spliced alignment of ESTs. Each EST is aligned against the genome independently from other ESTs. In order to be considered by CSF, a gap in the alignment must be flanked by a minimum number (25 by default) of matching residues. CSF searches for EST alignments that form the patterns illustrated in Figure 1A. It can be seen that the majority of ESTs must share a common gap or deletion and in addition must include minor transcripts that share only one of the common deletion endpoints. CSF defines the common deletion endpoints as authentic splice sites and the deletion end point of the minor transcript(s) as cryptic or alternative splice sites (arrowed). For constitutively spliced genes, these minor deletion endpoints occur very infrequently and are therefore candidate css. For alternatively spliced genes, CSF identifies both css and more frequently occurring alternative splice sites (see 'Results' section). As illustrated in Figure 1A, css can be 5′ or 3′ and upstream or downstream of the alternative authentic splice site. In the majority of cases the authentic splice sites defined by CSF are the same as the splice sites defined by reference sequences (i.e. NM_000518.4) and so represent commonly used splice sites (see below for exceptions).

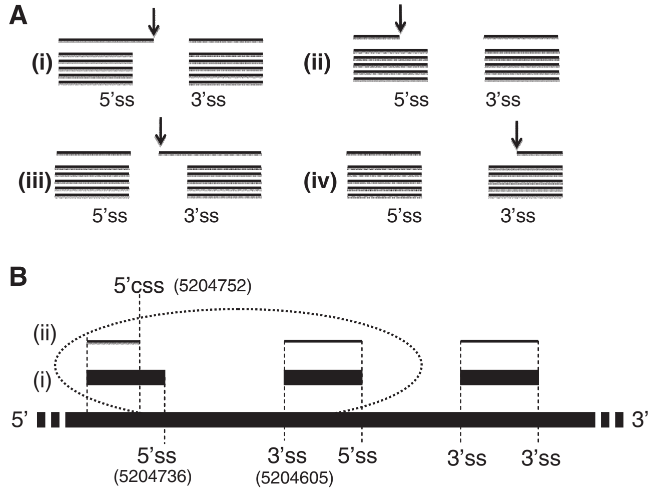


Figure 1. (**A**) CSF searches for transcript alignments that form one of four patterns (i–iv). All of these patterns contain a group of major transcripts that share a common deletion and a minor transcript that shares only one of the deletion endpoints. CSF defines the common deletion endpoints as authentic splice sites and the less common deletion endpoint of the minor transcript(s) as cryptic or alternative splice sites (arrowed). (**B**) Schematic of the HBB gene for human β-globin, which contains two introns that are constitutively spliced from pre-mRNA. As illustrated the vast majority of ESTs align as shown and define the three exons of this gene. The circle shows a pattern of ESTs that CSF is designed to recognize and that is reported in Figure 2A. The numbers in brackets show the genome coordinates of the three splice sites that are identified and listed by CSF (Figure 2A).

In more detail, coordinates of the splice sites from adjacent exon–intron–exon sequences are pooled into four-tuples which are then loaded into a relational database alongside the data linking them to their alignments. The database runs a query to detect tuple pairs satisfying the css condition: the overlapping introns match at one end and mismatch at the other, with a mismatching intron end from one tuple residing within the exon from the other tuple. For each tuple pair returned by the query, the splice sites of the intron that has more supporting ESTs are declared authentic and the remaining site is declared cryptic. Splice site coordinates are mapped against the NCBI36/hg18 human genome assembly.

As illustrated (Figure 1A), a css detected by CSF is further classified as to whether it is a 5′ or 3′ css and whether it is located in an exon or intron. The number of transcripts supporting a splice (either authentic or css), is printed and the complete list of such ESTs is linked under the count. Where applicable, the count is followed by a number in parentheses. For exonic css, parentheses always appear on the left-hand (authentic) side of the CSF report and on the right-hand side (cryptic) for intronic css. The numbers in parentheses show the transcripts that formally satisfy the css conditions (see Notes about CSF section).

Notes about CSF. CSF only lists authentic 5′- and 3′-splice sites when one of the two authentic splice sites is a more common alternative for another nearby splice site that is listed under the cryptic (alternative) column. This means that CSF only lists a subset of the total number of authentic splice sites for any one gene.

CSF provides a very good way of screening large numbers of transcripts for the minority that are likely to have been generated by use of a css or nearby alternative splice site. However, candidate transcripts do need to be checked, by for example, using the Splign alignment option (15). CSF classifies a minority of less well-supported deletions as authentic splices because these happen to meet the CSF conditions. These sites can be easily recognized from the CSF output because they have lower levels of support than other authentic splice sites and do not match reference sequence splice sites. Caution needs to be taken with the interpretation of these particular 'authentic' splice sites and also with the interpretation of css that are paired with authentic splice sites that have a low-parenthesis score. See Supplementary Data for illustrative examples.

Searching CSF. Two methods of searching for css are provided (Supplementary Figure S1). In the first method, a landmark EST accession is submitted and CSF returns a list of css within the genomic range of the entered EST. In the second method, an arbitrary genomic interval can be specified and CSF will return a hierarchic list of css for that interval. Entire chromosomes can also be entered such as NC_000001 (human chromosome 1) for which

CSF currently lists 3232 css. Links are provided in order to view the sequence alignments of individual css. CSF can predict css for *Homo sapiens*, *Bos taurus*, *Mus musculus*, *Danio rerio* and *Arabidopsis thaliana* and we intend to expand this range as further transcript data becomes available

Statistical analysis. There are 91 known different intron positions within the coding region of the actin gene family, these have been identified by sequencing actin genes from over 160 different species (16–18). The positions of all of these introns together with the 14 css that we have identified are plotted in Supplementary Figure S2A. Actin genes usually have 375 codons and therefore three times this number of possible positions for introns or css and so the probability of a single css exactly matching an intron position by chance is $91/(3 \times 375)$. The exact Fisher's test gives the P-value of 1.6×10^{-10} for 11 or more matches out of 14 occurring by chance. Similarly, we identified 135 css in the coding region of 51 different genes of the ribosomal protein gene database. The 51 genes are 190 codons in size on average and have a total of 957 introns (from 22 species). The probability of a css exactly matching an intron by chance is therefore $957/(190 \times 3 \times 51) = 0.033$. The binomial probability of 33 or more matches out of 135 occurring by chance is $P = 2.2 \times 10^{-15}$.

RT–PCR. Total RNA was extracted using Trizol and the cDNA first strand was synthesized from 0.1 µg of RNA using superscript III (Invitrogen) and random hexamers. PCR products were generated with Taq polymerase (NBI) for 25 or 35 cycles and separated on 5% native polyacrylamide gels. PCR bands were excised and cloned into pGEM-T Easy vectors (Promega) for sequencing by colony PCR followed by ABI Prism Big Dye Terminator cycle sequencing (Applied Biosystems).

RESULTS

Principle of css detection by CSF

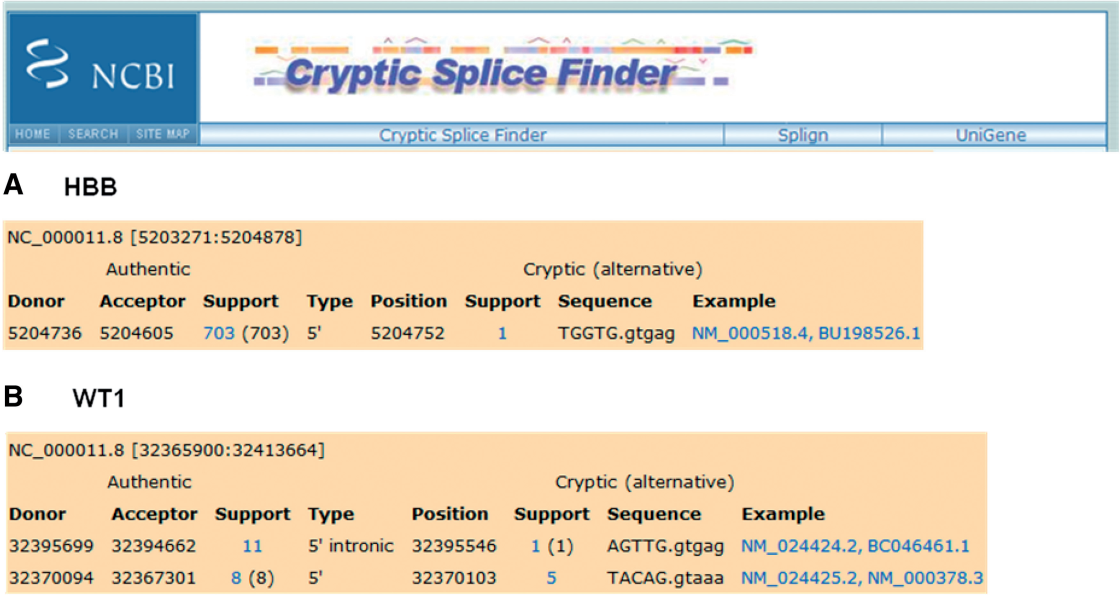
Css are often used highly efficiently in genetic disease following the mutation of nearby more competitive splice sites. We, therefore, reasoned that css might be used at a low but detectable frequency by normal genes. Figure 1A shows the patterns of EST alignments that CSF is designed to identify. It can be seen that CSF identifies groups of ESTs or transcripts that share a common deletion, when aligned to the genome, together with a minor transcript(s) that shares just one of the common deletion endpoints. CSF defines the common deletion endpoints as authentic splice sites and the unusual deletion endpoint of the minor transcript as a cryptic or alternative splice site (see 'Materials and Methods' section for further details). Figure 1B illustrates how ESTs align to the human HBB gene for β-globin, which has two introns. The circle identifies a pattern of alignments that is recognized by CSF because it includes a minor transcript that has an unusual deletion endpoint (position 5 204 752) that is a predicted css (Figures 1B and 2A). The reason why such css predictions turn out to be accurate (see below) is because of the restriction that the predicted css is paired with a commonly used splice site. This distinguishes mRNA deletions that are generated by low-level aberrant splicing from deletions that are generated by non-splicing mechanisms such as errors during transcription or during the generation of the EST.

Figure 2A shows the CSF output for HBB, which is one of the first genes in which css were identified (19). CSF identifies a single EST called BU198526 as having been generated by aberrant splicing. BU198526 aligns to HBB as illustrated in Figure 1B(ii), this alignment can be viewed in detail by a link to Splign (Figure 2). As explained above, the reason why BU198526 was identified by CSF is because it forms a pattern alignment with other ESTs as illustrated in Figure 1A(ii). The CSF output (Figure 2A) shows that BU198526 shares a deletion endpoint at position 52 044 605 that is in common with 703 other ESTs (link provided) but differs in having an unusual 5′ deletion endpoint at position 5 204 752, which is the predicted css. A comparison with the known css of HBB confirms that 5 204 752 is indeed a css (Supplementary Table S1). We show below that css predictions by CSF are very reliable even if supported by just a single EST, as in this case. However, HBB has 11 known css (Supplementary Table S1), which illustrates that CSF is limited in its predictions, most probably by the amount of available transcript data.

For alternatively spliced genes, CSF identifies both css and also nearby alternative splice sites that act to shorten or lengthen exons. This is because the two types of splice sites are rather similar. For example, the css identified in HBB (Figure 2A) might be considered an alternative splice site if used at a greater frequency than 1 in 703.

WT1 is an example of gene that encodes at least three alternatively spliced mRNAs and for which there are only 80 ESTs. CSF identifies two css (Figure 2B), however, the CSF output also shows that the predicted css at 32 370 103, for example, is supported by five ESTs including the reference sequence NM_000378.3 and that the 'authentic' splice site nine bases away at 32 370 094 is supported by only eight transcripts. The similar usage of the authentic and css shows that these are really alternative 5′ ss. However, it is useful to have this type of information because alternative splice sites are also implicated in genetic disease and in this particular case disruption of the splice site at position 32 370 094 gives rise to Frasier syndrome, possibly due to the increased use of splice site 32 370 103 (20,21).

To test CSF we analyzed a database called DBASS (database of aberrant splice sites), which lists 340 human genes that have one or more css that are activated in genetic disease (3). There are 814 different css listed in DBASS and CSF predicts 609 css from the same set of 340 genes. Fifty-eight percent of these predictions are supported by only single ESTs (Supplementary Table S1). Before comparing the css identified by CSF with those of DBASS, we first asked whether the css predictions that are supported by just single ESTs were likely to have been generated by aberrant splicing. We reasoned that because CSF identifies deletion endpoints irrespective of their sequence, then if these rare deletion endpoints



A HBB

NC_000011.8 [5203271:5204878]

Authentic			Cryptic (alternative)				
Donor	**Acceptor**	**Support**	**Type**	**Position**	**Support**	**Sequence**	**Example**
5204736	5204605	703 (703)	5'	5204752	1	TGGTG.gtgag	NM_000518.4, BU198526.1

B WT1

NC_000011.8 [32365900:32413664]

Authentic			Cryptic (alternative)				
Donor	**Acceptor**	**Support**	**Type**	**Position**	**Support**	**Sequence**	**Example**
32395699	32394662	11	5' intronic	32395546	1 (1)	AGTTG.gtgag	NM_024424.2, BC046461.1
32370094	32367301	8 (8)	5'	32370103	5	TACAG.gtaaa	NM_024425.2, NM_000378.3

Figure 2. (**A**) CSF output for the human gene HBB for β-globin. (**B**) CSF output for WT1 (see text). The coordinates that are used refer to the NCBI36/hg18 human genome assembly. It should be noted that HBB and WT1 genes align in a 3′–5′ direction with respect to their genome coordinates.

were generated by splicing they should look like splice sites. The predictions were divided into 5′ and 3′ css and Table 1 shows that the predicted css have a very good match to the expected 5′- or 3′-splice site consensus sequence, which is a strong proof that the vast majority of even the rarest deletions identified by CSF were generated by aberrant splicing.

Of the two sets of DBASS and CSF css, only 46 are in common (Supplementary Tables S1 and S2). However, there are only 61 cases where the DBASS and CSF css are located within the same exon or intron (Supplementary Table S1, column 5), giving a match rate of 46/61 or 75%. Thirteen of the 15 DBASS css that did not match were either created *de novo* by mutation and would, therefore, not be expected to be identified by CSF or were of opposite type, for example, a 5′ css in DBASS and a 3′ css in CSF and would, therefore, be unlikely to match (Supplementary Table S1). Only two out of 48 css predictions that could be meaningfully compared with DBASS did not exactly match (Supplementary Table S1). This result together with the clear consensus sequence results shown in Table 1 indicates that CSF predictions are highly reliable and can only generate a very low level of false positives even when supported by single ESTs. The small number of 46 css in common between the CSF database and DBASS (Supplementary Table S1) presumably indicates that both DBASS and CSF have identified relatively few of all possible css within this set of 340 genes.

We confirmed that css predicted by CSF could also be detected by experiment. Five css predictions by CSF that DBASS shows to be activated in patients were detected in human cell lines that do not have the causative genetic mutations (Figure 3, lanes 1–5). The last lane shows an example of a css prediction by CSF that has not been reported in patients. The PCR products marked with asterisks were confirmed by sequencing (Supplementary Figure S3). These results confirmed the CSF predictions and also show that the css we analyzed were already active in a normal genetic background.

Css were previously reported to be active at very low levels in normal globin genes (22) and our results systematically extend this finding. It is clear from Figure 3 that some splice sites that have been classified as css are used by normal genes at a relatively high frequency (lanes 2–5), which is in accordance with the CSF analysis (Supplementary Table S2). These css can therefore also be regarded as alternative splice sites that are further activated in disease.

Css and introns

We previously identified nine css within the coding region of the actin gene family by experiment and reported an eight out of nine exact match to the position of introns in actin genes from other species (13,14). CSF identified eight css within the coding region of the actin gene family of which three were identical in position to the previously identified actin css. Of the five new css identified by CSF, three exactly match the position of introns in other species, therefore extending our previous study and adding further support to the validity of CSF predictions (Supplementary Figure S2).

Because CSF predictions are sufficiently accurate, we could use this program to establish whether there is a general relationship between the position of css and introns. We therefore analyzed an extensive intron database compiled of some 80 genes from a wide range of species that encode ribosomal proteins (23). Figure 4A shows a small part of this database that records the positions of introns against a 14 amino acid stretch of the gene RPS5 from 23 species (file kindly provided by N. Kenmochi). Introns that fall between codons are indicated by red shading and introns that fall within codons are indicated in blue (Phase 1) and green (Phase 2). For example, there is an intron located within a codon for valine (V) in the fungus *Cryptococcus neoformans* but in no other listed species. This observation is typical in that although introns are wide-spread among eukaryotes, the vast majority of individual intron positions are found in only a minority of species (Supplementary Table S3) (18).

We screened for css in ribosomal protein genes from human, mouse, *Danio rerio* and *Arabidopsis thaliana* and identified 74, 41, 16 and 4 css, respectively, within coding regions, these diminishing numbers simply reflect the amount of EST data that is available. Alignment of the predicted css indicates that the large majority are correct because they conform to splice site consensus sequences (Supplementary Table S4). All of these css were used at low frequencies, similar to the HBB example of Figure 2A. Figure 4B shows how we record the position of css; in this example, there is a css in the human RPS5 gene that exactly matches the position of introns in *Drososphila*

Table 1. The alignment of css predictions by CSF reveals consensus sequences typical of splice sites

Base (%)	−5	−4	−3	−2	−1	1	2	3	4	5
5′ css predictions										
T	28.8	17.6	21.2	14.1	11.8	0.6	90.6	5.9	11.8	10.0
C	22.4	24.1	27.6	17.1	4.7	0.6	7.6	11.8	20.0	12.9
G	24.1	34.7	25.3	17.6	63.5	97.1	0.6	40.6	30.0	63.5
A	24.1	19.4	25.3	50.6	19.4	1.2	0.6	41.2	37.6	12.9
3′ css predictions										
T	31.8	22.9	21.2	4.5	2.2	17.9	26.3	31.8	26.3	30.2
C	32.4	30.2	59.8	0.6	0.6	23.5	24.6	29.6	32.4	26.3
G	20.1	24.0	5.0	1.1	96.6	33.0	18.4	21.2	22.3	21.8
A	15.6	22.9	14.0	93.9	0.6	25.7	30.7	17.3	19.0	21.8

This Table is compiled from 169 and 179 examples of 5′ and 3′ css predictions, respectively, that are supported by only single ESTs (Supplementary Table S1). The relative proportions of the bases T, C, G and A are shown as a percentage at five positions both upstream and downstream of the predicted cryptic cleavage site. The most frequently occurring bases are shaded.

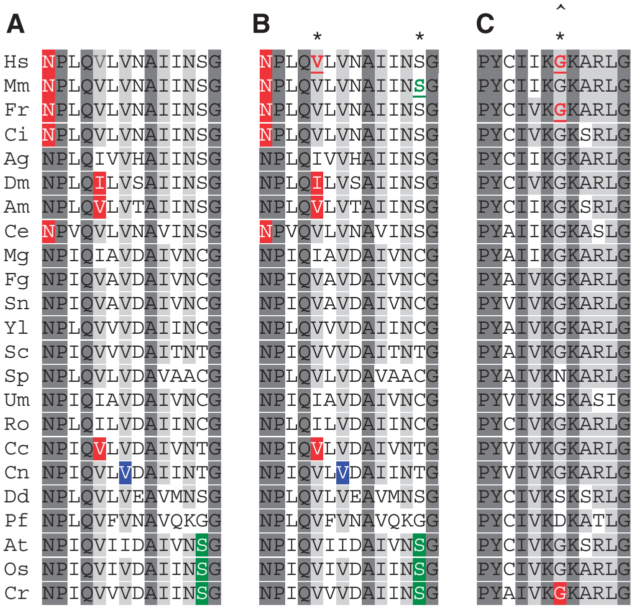

A
B
C
*
*
ˆ
*
Hs NPLQVLVNAIINSG NPLQVLVNAIINSG PYCIIKGKARLG
Mm NPLQVLVNAIINSG NPLQVLVNAIINSG PYCIIKGKARLG
Fr NPLQVLVNAIINSG NPLQVLVNAIINSG PYCIVKGKARLG
Ci NPLQVLVNAIINSG NPLQVLVNAIINSG PYCIVKGKSRLG
Ag NPLQIVVHAIINSG NPLQIVVHAIINSG PYCIIKGKARLG
Dm NPLQILVSAIINSG NPLQILVSAIINSG PYCIVKGKARLG
Am NPLQVLVTAIINSG NPLQVLVTAIINSG PYCIIKGKSRLG
Ce NPVQVLVNAVINSG NPVQVLVNAVINSG PYAIIKGKASLG
Mg NPIQIAVDAIVNCG NPIQIAVDAIVNCG PYAIIKGKARLG
Fg NPIQVAVDAIVNCG NPIQVAVDAIVNCG PYAIVKGKARLG
Sn NPIQVAVDAIVNCG NPIQVAVDAIVNCG PYVIVKGKARLG
Yl NPLQVVVDAIINCG NPLQVVVDAIINCG PYAIVKGKARLG
Sc NPIQVVVDAITNTG NPIQVVVDAITNTG PYAIVKGKARLG
Sp NPLQVLVDAVAACG NPLQVLVDAVAACG PYAIVKNKARLG
Um NPIQIAVDAIVNCG NPIQIAVDAIVNCG PYVIVKSKASIG
Ro NPLQILVDAIINCG NPLQILVDAIINCG PYCIVKGKARLG
Cc NPIQVLVDAIVNTG NPIQVLVDAIVNTG PYVIVKGKARLG
Cn NPIQVLVDAIINTG NPIQVLVDAIINTG PYVIVKGKARLG
Dd NPLQVLVEAVMNSG NPLQVLVEAVMNSG PYCIVKSKSRLG
Pf NPLQVFVNAVQKGG NPLQVFVNAVQKGG PYCIVKDKATLG
At NPIQVIIDAIVNSG NPIQVIIDAIVNSG PYCIVKGKSRLG
Os NPIQVIVDAIINSG NPIQVIVDAIINSG PYCIVKGKARLG
Cr NPIQVVVDAIINSG NPIQVVVDAIINSG PYAIVKGKARLG

Figure 4. Comparison of intron and css positions for a small part of the ribosomal protein gene database (23). (**A**) An alignment of 14 amino acids of the RPS5 gene that marks the position of introns for 23 species. (**B**) The same alignment including two css positions identified by CSF. (**C**) An alignment of part of RPL7A that illustrates the conservation of two css that also match an intron in *Chlamydomonas reinhardtii* (Cr). * - marks the position of css that match introns; ˆ - marks conserved css.

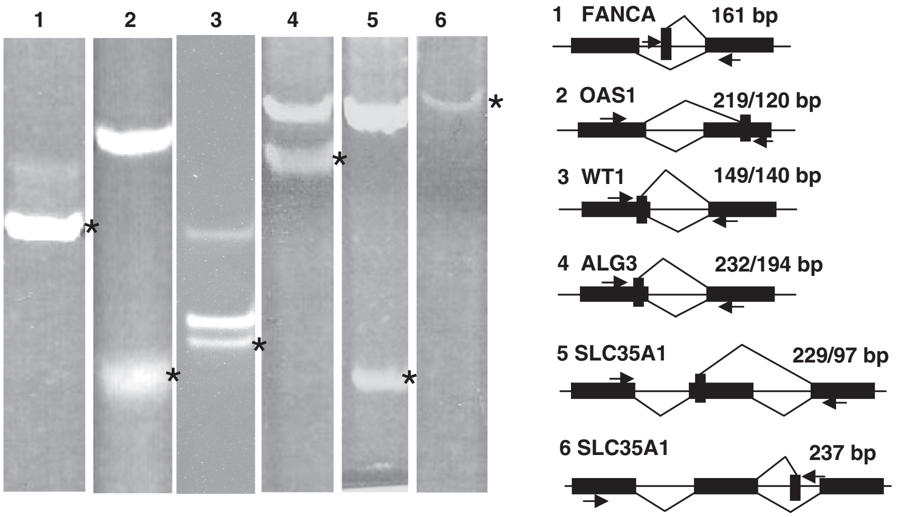


Figure 3. Experimental confirmation of CSF predictions. Predicted css are shown by the vertical boxes for the indicated genes. Active css would be expected to generate PCR fragments of the sizes shown in the gene diagrams. PCR products marked with asterisks were sequenced in order to confirm the use of the predicted css (Supplementary Figure S3). Messenger RNA was prepared from the human cell lines K562 (lane 1); HEPG2 (lanes 2, 3) and primary mesenchymal stem cells (lanes 4–6) and used for RT–PCR with the indicated primers (see Supplementary Data).

(Dm) the honeybee (Am) and a fungus (Cc) and a nearby css in the mouse (Mm) that exactly matches the position of introns in three plant species (At, Os, Cr). We have identified 135 css to date and of these at least five are conserved between species (Figure 4C and Supplementary Figure S4). Thirty-three out of 135 css of the ribosomal gene family exactly match the position of introns in other species (Supplementary Figure S4). This proportion is smaller than the 11/14 match observed for the actin gene family (Supplementary Figure S2), but is still highly significant ($P = 2.2 \times 10^{-15}$, see 'Materials and Methods' section).

DISCUSSION

The CSF program is designed to identify transcripts that are generated through the low level use of css by normal genes. In addition, CSF also identifies a subset of alternative splice sites that are similar to css, but are used at a greater frequency. Both types of splice sites are reported to be activated in genetic disease as a result of mutations that disrupt the function of nearby competitive splice sites (4,5). Our analyses show that css predictions by CSF are very reliable and so we would expect, for example, that many more of the predicted css and alternative splice sites of Supplementary Table S1 will be discovered to be associated with genetic disease. The identification of css by CSF is limited by the amount of available transcript data but this will improve as further transcript sequences become available, particularly with the advent of mRNA deep sequencing (24).

CSF was designed primarily to identify css in highly conserved gene families such as actin in order to advance our understanding of intron origin. We found that about 25% of the css within the coding sequence of the large family of genes that encode ribosomal proteins exactly match the position of introns that are present in equivalent genes from other species (Figure 4 and Supplementary Figure S4). This compares to a 78% match between actin css and introns, however, there is far more phylogenetic data available for the actin gene family and so relatively more introns have been discovered. Consequently, the css that are recorded in Supplementary Figure S4 predict the positions of as yet undiscovered introns in the ribosomal gene family.

The well-known and valuable early and late models of intron origin both assume that the splicing machinery evolved for the purpose of removing introns that were either present in the most ancestral genomes or were inserted after the separation of the prokaryotes (25–27). At the time the models were proposed, non-intronic splicing information was not generally thought to be of major relevance and so is not an important feature of either model (28).

However, it is now established that exon junction sequences are older and better conserved than most if not all introns and were sites for intron insertions during evolution (18,29–33). Indeed, a number of intron properties such as their phasing with respect to the coding sequence (29–30,34) and their location with respect to protein structure have now been largely attributed to the flanking exon junction sequence (35,36).

Consequently, our finding that css often match the position of exon junction sequences in gene homologues, strongly indicates that css were targeted by intron insertions during evolution. Our data also indicates that the information that lies outside some introns is not only conserved with homologs that lack such introns but is also capable of being independently recognized by the splicing machinery and can define the position of the 'missing' introns (13). This is a striking observation because although the splicing information that flanks introns contributes to their recognition, there is no mechanistic reason for this information to be autonomous rather than auxiliary in nature for the purpose of intron removal (37). It is unlikely that a css is recognized from the immediate splice-site information alone (8–10), suggesting that other information such as splicing enhancer sequences are also conserved between some gene homologues independently of intron presence.

All of the available evidence indicates that the splicing information that flanks introns was largely in place prior to their insertion (18,29–33). The key question we have started to address here is whether such information might have been functional. Our data so far does not prove but it certainly supports the suggestion that information of this nature could have been used by the splicing machinery for splicing purposes prior to the arrival of introns, which is a rather different concept to the early and late models of intron origin (28).

The splicing information that flanks introns is perhaps similar to the splicing information that enables the mRNA of intronless genes to interact with components of the splicing machinery for the purpose of mRNA biogenesis, nonsense mediated decay or alternative splicing (38–43). For genes that have introns there are still many reports of intron-independent splicing between exonic splice sites (44–53). If intron-independent splicing is an ancestral mechanism then it may be far more prevalent across species than is currently realized.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Naoya Kenmochi for his help and would also like to thank Anil Chandrashekran, Andy Newman, Peter Rogan, Alison Russell, Malcolm Parker Terrie Sadusky and Arlin Stoltzfus.

FUNDING

Reserch grants from Biotechnology and Biosciences Research Council and Atazoa. Funding for open access charge: Genesis Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Padgett,R.A., Grabowski,P.J., Konarska,M.M., Seiler,S. and Sharp,P.A. (1986) Splicing of messenger RNA precursors. *Annu. Rev. Biochem.*, **55**, 1119–1150.
2. Green,M.R. (1986) Pre-mRNA splicing. *Annu. Rev. Genet.*, **20**, 671–708.
3. Buratti,E., Chivers,M., Hwang,G. and Vorechovsky,I. (2011) DBASS3 and DBASS5: databases of aberrant 3′- and 5′-splice sites. *Nucleic Acids Res.*, **39**, D86–D91.
4. Buratti,E., Chivers,M., Kralovicova,J., Romano,M., Baralle,M., Krainer,A.R. and Vorechovsky,I. (2007) Aberrant 5′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.*, **35**, 4250–4263.
5. Wang,G.S. and Cooper,T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749–761.
6. Hastings,M.L., Resta,N., Traum,D., Stella,A., Guanti,G. and Krainer,A.R. (2005) An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nat. Struct. Mol. Biol.*, **12**, 54–59.
7. Lopez-Bigas,N., Audit,B., Ouzounis,C., Parra,G. and Guigo,R. (2005) Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.*, **579**, 1900–1903.
8. Wimmer,K., Roca,X., Beiglbock,H., Callens,T., Etzler,J., Rao,A.R., Krainer,A.R., Fonatsch,C. and Messiaen,L. (2007) Extensive in silico analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5′ splice-site disruption. *Hum. Mutat.*, **28**, 599–612.
9. Kralovicova,J. and Vorechovsky,I. (2007) Global control of aberrant splice-site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. *Nucleic Acids Res.*, **35**, 6399–6413.
10. Russo,A., Siciliano,G., Catillo,M., Giangrande,C., Amoresano,A., Pucci,P., Pietropaolo,C. and Russo,G. (2010) hnRNP H1 and intronic G runs in the splicing control of the human rpL3 gene. *Biochim. Biophys. Acta.*, **1799**, 419–428.
11. Divina,P., Kvitkovicova,A., Buratti,E. and Vorechovsky,I. (2009) Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur. J. Hum. Genet.*, **17**, 759–765.
12. Betz,B., Theiss,S., Aktas,M., Konermann,C., Goecke,T.O., Moslein,G., Schaal,H. and Royer-Pokora,B. (2009) Comparative in silico analyses and experimental validation of novel splice site and missense mutations in the genes MLH1 and MSH2. *J. Cancer Res. Clin. Oncol.*, **136**, 123–134.
13. Sadusky,T., Newman,A.J. and Dibb,N.J. (2004) Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr. Biol.*, **14**, 505–509.
14. Stoltzfus,A. (2004) Molecular evolution: introns fall into place. *Curr. Biol.*, **14**, R351–352.
15. Kapustin,Y., Souvorov,A., Tatusova,T. and Lipman,D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct.*, **3**, 20.
16. Bhattacharya,D. and Weber,K. (1997) The actin gene of the glaucocystophyte Cyanophora paradoxa: analysis of the coding region and introns, and an actin phylogeny of eukaryotes. *Curr. Genet.*, **31**, 439–446.
17. Flakowski,J., Bolivar,I., Fahrni,J. and Pawlowski,J. (2006) Tempo and mode of spliceosomal intron evolution in actin of foraminifera. *J. Mol. Evol.*, **63**, 30–41.
18. Qiu,W.G., Schisler,N. and Stoltzfus,A. (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol. Biol. Evol.*, **21**, 1252–1263.
19. Treisman,R., Orkin,S.H. and Maniatis,T. (1983) Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. *Nature*, **302**, 591–596.
20. Barbaux,S., Niaudet,P., Gubler,M.C., Grunfeld,J.P., Jaubert,F., Kuttenn,F., Fekete,C.N., Souleyreau-Therville,N., Thibaud,E., Fellous,M. *et al.* (1997) Donor splice-site mutations in WT1 are responsible for Frasier syndrome. *Nat. Genet.*, **17**, 467–470.
21. Niaudet,P. and Gubler,M.C. (2006) WT1 and glomerular diseases. *Pediatr. Nephrol.*, **21**, 1653–1660.
22. Haj Khelil,A., Deguillien,M., Moriniere,M., Ben Chibani,J. and Baklouti,F. (2008) Cryptic splicing sites are differentially utilized in vivo. *FEBS J.*, **275**, 1150–1162.
23. Yoshihama,M., Nguyen,H.D. and Kenmochi,N. (2007) Intron dynamics in ribosomal protein genes. *PLoS ONE*, **2**, e141.
24. Pickrell,J.K., Pai,A.A., Gilad,Y. and Pritchard,J.K. (2010) Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, **6**, e1001236.
25. Cavalier-Smith,T. (1985) Selfish DNA and the origin of introns. *Nature*, **315**, 283–284.
26. Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.
27. Roy,S.W. and Gilbert,W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.*, **7**, 211–221.
28. Dibb,N.J. (1993) Why do genes have introns? *FEBS Lett.*, **325**, 135–139.
29. Nguyen,H.D., Yoshihama,M. and Kenmochi,N. (2006) Phase distribution of spliceosomal introns: implications for intron origin. *BMC Evol. Biol.*, **6**, 69.
30. Ruvinsky,A., Eskesen,S.T., Eskesen,F.N. and Hurst,L.D. (2005) Can codon usage bias explain intron phase distributions and exon symmetry? *J. Mol. Evol.*, **60**, 99–104.
31. Bhattacharya,D., Simon,D., Huang,J., Cannone,J.J. and Gutell,R.R. (2003) The exon context and distribution of Euascomycetes rRNA spliceosomal introns. *BMC Evol. Biol.*, **3**, 7.
32. Dibb,N.J. and Newman,A.J. (1989) Evidence that introns arose at proto-splice sites. *EMBO J.*, **8**, 2015–2021.
33. Sverdlov,A.V., Rogozin,I.B., Babenko,V.N. and Koonin,E.V. (2004) Reconstruction of ancestral protosplice sites. *Curr. Biol.*, **14**, 1505–1508.
34. Long,M., de Souza,S.J., Rosenberg,C. and Gilbert,W. (1998) Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. *Proc. Natl Acad. Sci. USA*, **95**, 219–223.
35. De Kee,D.W., Gopalan,V. and Stoltzfus,A. (2007) A sequence-based model accounts largely for the relationship of intron positions to protein structural features. *Mol. Biol. Evol.*, **24**, 2158–2168.
36. Whamond,G.S. and Thornton,J.M. (2006) An analysis of intron positions in relation to nucleotides, amino acids, and protein secondary structure. *J. Mol. Biol.*, **359**, 238–247.
37. Burge,C.B., Tuschl,T. and Sharp,P.A. (eds), (1999) *Splicing of Precursors to mRNAs by the Spliceosomes*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
38. Brogna,S. and Wen,J. (2009) Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.*, **16**, 107–113.
39. Guang,S., Felthauser,A.M. and Mertz,J.E. (2005) Binding of hnRNP L to the pre-mRNA processing enhancer of the herpes simplex virus thymidine kinase gene enhances both polyadenylation and nucleocytoplasmic export of intronless mRNAs. *Mol. Cell. Biol.*, **25**, 6303–6313.
40. Guang,S. and Mertz,J.E. (2005) Pre-mRNA processing enhancer (PPE) elements from intronless genes play additional roles in mRNA biogenesis than do ones from intron-containing genes. *Nucleic Acids Res.*, **33**, 2215–2226.
41. Juneau,K., Nislow,C. and Davis,R.W. (2009) Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. *Genetics*, **183**, 185–194.
42. Pozzoli,U., Riva,L., Menozzi,G., Cagliani,R., Comi,G.P., Bresolin,N., Giorda,R. and Sironi,M. (2004) Over-representation of exonic splicing enhancers in human intronless genes suggests multiple functions in mRNA processing. *Biochem. Biophys. Res. Commun.*, **322**, 470–476.
43. Brody,Y., Neufeld,N., Bieberstein,N., Causse,S.Z., Bohnlein,E.M., Neugebauer,K.M., Darzacq,X. and Shav-Tal,Y. (2011) The *In Vivo* Kinetics of RNA Polymerase II Elongation during Co-Transcriptional Splicing. *PLoS Biol.*, **9**, e1000573.
44. Ng,B., Yang,F., Huston,D.P., Yan,Y., Yang,Y., Xiong,Z., Peterson,L.E., Wang,H. and Yang,X.F. (2004) Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes. *J. Allergy Clin. Immunol.*, **114**, 1463–1470.

45. Rovescalli,A.C., Cinquanta,M., Ferrante,J., Kozak,C.A. and Nirenberg,M. (2000) The mouse Nkx-1.2 homeobox gene: alternative RNA splicing at canonical and noncanonical splice sites. *Proc. Natl Acad. Sci. USA*, **97**, 1982–1987.
46. Baumbusch,L.O., Myhre,S., Langerod,A., Bergamaschi,A., Geisler,S.B., Lonning,P.E., Deppert,W., Dornreiter,I. and Borresen-Dale,A.L. (2006) Expression of full-length p53 and its isoform Deltap53 in breast carcinomas in relation to mutation status and clinical parameters. *Mol. Cancer*, **5**, 47.
47. Cagliani,R., Bardoni,A., Sironi,M., Fortunato,F., Prelle,A., Felisari,G., Bonaglia,M.C., D'Angelo,M.G., Moggio,M., Bresolin,N. *et al.* (2003) Two dystrophin proteins and transcripts in a mild dystrophinopathic patient. *Neuromuscul. Disord.*, **13**, 13–16.
48. Chikaev,N.A., Bykova,E.A., Najakshin,A.M., Mechetina,L.V., Volkova,O.Y., Peklo,M.M., Shevelev,A.Y., Vlasik,T.N., Roesch,A., Vogt,T. *et al.* (2005) Cloning and characterization of the human FCRL2 gene. *Genomics*, **85**, 264–272.
49. Cocquet,J., Chong,A., Zhang,G. and Veitia,R.A. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics*, **88**, 127–131.
50. Cox,P.R., Siddique,T. and Zoghbi,H.Y. (2001) Genomic organization of Tropomodulins 2 and 4 and unusual intergenic and intraexonic splicing of YL-1 and Tropomodulin 4. *BMC Genomics*, **2**, 7.
51. Galante,P.A., Sakabe,N.J., Kirschbaum-Slager,N. and de Souza,S.J. (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*, **10**, 757–765.
52. Lukas,J., Gao,D.Q., Keshmeshian,M., Wen,W.H., Tsao-Wei,D., Rosenberg,S. and Press,M.F. (2001) Alternative and aberrant messenger RNA splicing of the mdm2 oncogene in invasive breast cancer. *Cancer Res.*, **61**, 3212–3219.
53. Aebi,M., Hornig,H., Padgett,R.A., Reiser,J. and Weissmann,C. (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell*, **47**, 555–565.