Identification of activated cryptic 5′ splice sites using structure profiles and odds measure

Kun-Nan Tsai and Daryi Wang*

Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan

Received October 28, 2011; Revised December 20, 2011; Accepted January 17, 2012

ABSTRACT

The activation of cryptic 5′ splice sites (5′ SSs) is often related to human hereditary diseases. The DNA-based mutation screening strategies are commonly used to recognize the cryptic 5′ SSs, because features of the local DNA sequence can influence the choice of cryptic 5′ SSs. To improve the identification of the cryptic 5′ SSs, we developed a structure-based method, named SPO (structure profiles and odds measure), which combines two parameters, the structural feature derived from hydroxyl radical cleavage pattern and odds measure, to assess the likelihood of a cryptic 5′ SS activation in competing with its paired authentic 5′ SS. Compared to the current tools for identifying activated cryptic 5′ SSs, the SPO algorithm achieves higher prediction accuracy than the other methods, including MaxEnt, MDD, Markov model, weight matrix model, Shapiro and Senapathy matrix, Ri and ΔG. In addition, the predicted ΔSPO scores from the SPO algorithm exhibited a greater degree of correlation with the strength of cryptic 5′ SS activation than that measured from the other seven methods. In conclusion, the SPO algorithm provides an optimal identification of cryptic 5′ SSs, can be applied in designing mutagenesis experiments for various splicing events and may be helpful to investigate the relationship between structural variants and human hereditary diseases.

INTRODUCTION

Mutations at splice sites occur frequently and result in the activation of the so-called cryptic splice sites (1–3). Two typical cases in human genes, BRCA1 and BRCA2, contain several intronic genetic variants (4,5), and approximately 5% of these are associated with splice site mutation (4). These mutations have a potential effect on the activation of cryptic 5′ splice sites (5′ SSs) (4,5) that lead to cryptic splicing events. These cryptic splicing events were considered aberrant and often cause human hereditary diseases (2,6). Therefore, predicting the activation of cryptic 5′ SSs is an essential approach in investigating human hereditary diseases.

Various approaches are used in cryptic 5′ SS identification. Recently, an EST-based method named cryptic splice finder (CSF) (7) used the spliced alignment of ESTs to identify the cryptic splice site. Although the CSF program is useful for investigating splicing mutation in genetic disease, it relies considerably on the availability of sufficient EST data and accurate genomic annotations. Another approach (8,9) used information content (Ri) to detect activated cryptic 5′ SSs in human genes. Ri is the dot product of a particular sequence vector and weight matrix derived from the nucleotide frequencies at each splice site and is used to interpret mutated authentic splice sites and associated splicing regulatory sites (9). Although Ri provides useful information for analyzing the nucleotide substitutions that potentially impair splicing, the identification of activated cryptic 5′ SSs was reported to be less accurate. Sahashi et al. (10) recently used the improved Ri to estimate the splicing consequences of mutations at human 5′ SSs and discovered that Ri had low sensitivity in predicting splicing mutations. In addition to the sequence-based analyses mentioned, a thermodynamic inference scheme, based on binding free energy (ΔG) toward the stability of the RNA duplex between 5′ SS and U1 snRNA, was proposed for 5′ SS selection (11). The method considered the effects of molecular structure and revealed that the ΔG method may discriminate strong and intermediate activation of cryptic 5′ SSs in competition assays. However, the identification for the intrinsic strength of cryptic 5′ SSs using ΔG is considerably inaccurate (6). Recently, Buratti et al. (12) collected 254 cryptic 5′ SSs that were activated by mutations in human disease genes and analyzed the mutation patterns and nucleotide structures in detail. They also evaluated the performance of several computational methods, including the Shapiro and Senapathy matrix.
tested as a splicing feature for 5-
evaluate the activation of a cryptic 5-
(SPO) algorithm, was developed to quantitatively
influence the choice of cryptic 5-
structure properties of the local DNA sequence can
Importantly, a recent report indicated that the changes of
isms could induce larger structural changes in the
Their results indicated that single-nucleotide polymorph-
(28) as genome-scale structural information to analyze
DNA–protein complexes (22–24). Recently, the hydroxyl
radical cleavage patterns of DNA were discovered to be
used for monitoring structural changes of DNA molecules
hydroxyl radical cleavage patterns (18,19) were widely
regulatory molecules. To detect such interactions, the
hydroxyl radical cleavage pattern was used for assessing
structure of DNA molecules and their related biolo-
gical order to realize the influence of single base pair substitutions in
The preference of DNA-based mutation screening
strategies (12,30) was used to investigate cryptic 5' SSs in genetic diseases, and the feature was applied in the pre-
diction tool (30). In fact, some signals that may influence
the choice of 5' SSs in the local DNA sequence have been
tested as a splicing feature for 5' SS prediction (31). To our
knowledge, the association of DNA structure and the
choice of cryptic 5' SSs are rarely discussed, and a
structure-based method for the screening of activated
cryptic 5' SSs for human disease genes is not available.
In this study, an advanced version with structure-based method,
named structure profiles and odds measure
(SPO) algorithm, was developed to quantitatively
evaluate the activation of a cryptic 5' SS in competing with its authentic 5' SS. The SPO algorithm combined structural profiles with odds measure to assess the activation likelihood for a cryptic 5' SS. The results indicates that the SPO algorithm was more efficient than the other seven approaches, including S&S (13), WMM (14), MM (15), MaxeEnt (16), MDD (17), R(10) and ΔG methods (32), in identifying an activated cryptic 5' SS in competition with its paired authentic 5' SS. In addition, the ∆SPO score from the SPO algorithm was a more effective score than the others in identifying the inherent strength of 5' SSs in human disease genes.

MATERIALS AND METHODS

Data sets

Two sets of human mutation splicing sequence data were used for the development and evaluation of the SPO al-
gorithm. The first data set, HMD1, was collected from published studies (6,8,12) containing 490 authentic and
cryptic 5' SS data pairs (Supplementary Table S1), which were experimentally validated. Of the 490 data pairs, 275
were inactivated pairs and 215 were activated pairs. These
490 pairs of splice site sequences were used to train the
SPO algorithm in determining a scoring threshold for the
successful prediction of cryptic 5' SS activation. The second data set, HMD2, contained 52 data pairs
(Supplementary Table S2) from two competition assays, competition scheme I (CS-I) and competition scheme II
(CS-II), which contained 26 authentic and cryptic 5' SS data pairs (11). The CS-I compared mutations of cryptic 5' SSs with wild types of authentic 5' SSs, whereas CS-II
compared mutations of cryptic 5' SSs with weakened types of authentic 5' SSs. From CS-I and CS-II, each
group of 26 cryptic 5' SSs was subdivided into 6 strong, 13 intermediate and 7 weak cryptic 5' SSs according to
their splicing strength. The HDM2 sequences were solely used to correlate the scoring method with the actual activ-
ation strength for cryptic 5' SS independent from those
490 paired splicing sequences from HDM1. In total, 189249 5' SSs (10) from the entire human genome were
extracted as source data for the SPO algorithm.

SPO algorithm

For the likelihood of activating a cryptic 5' SS, the SPO algorithm was developed based on the combination of
structural profiles with odds measure. The structural profiles consider the local DNA structural change
between the before and after mutation that occurs in a
5' SS and the odds measure computes the actual relative probability for a splicing event to occur. Figure 1 shows the
SPO algorithm. The details of defining and combining these two numerals (‘SP’ for structural profiles and ‘O’ for odds)
into the proposed ‘SPO’ algorithm are as follows:

(1) First, a 5' SS pattern was defined as \{X_1, X_2, \ldots, X_m\}, where X_m represents the m-th nucleotide and consists of nucleotide bases \{A, G, C, T\}. X_1, X_2 and X_3 obtain from exonic region, and X_6, X_7, X_8 and X_9 obtain from intronic region. X_4 and X_5 are the center consensus of a 5' SS. Following the convention for the splice site coordinate, the center
consensus $X_4$ and $X_5$ assume the position of GT. Second, the hydroxyl radical cleavage pattern from ORChID (28) was used as DNA structural profiles and provided high-resolution quantitative information of the local shape of DNA molecules. Before mutation occurrence, the DNA structural profile for a 5' SS pattern was defined as $(Y_{1b}, Y_{2b}, \ldots, Y_{mb})$, where $Y_{mb}$ represents the structural profile of the $m$-th nucleotide. After mutation occurrence, the DNA structural profile for a 5' SS pattern was defined as $(Y_{1a}, Y_{2a}, \ldots, Y_{ma})$, where $Y_{ma}$ represents the structural profile of the $m$-th nucleotide. The DNA structural change for a 5' SS between the before and after mutation occurrence was defined as $S(Y_{ma}, Y_{mb})$ and was computed using Euclidean distance. In detail, $S(Y_{ma}, Y_{mb})$ was given by:

$$S(Y_{ma}, Y_{mb}) = \sqrt{\sum_{i=1}^{m} (Y_{ia} - Y_{ib})^2}$$

(1)

The $S_p$ value and $S_q$ value were defined as the structural change for a cryptic 5' SS and an authentic 5' SS individually. Finally, the $Sc$ value was defined as $(S_p + S_q + 1)$ and used to assess the activation likelihood for a cryptic 5' SS. Here, to avoid $Sc = 0$ causing a non-meaning Or value (see the next paragraph), 1 as a constant was used to keep $Sc = 1$ when $S_p + S_q = 0$.

(2) The improved odds measure was used to identify activated cryptic 5' SSs. All known 189 249 5' SSs (10) in human genome were extracted as source data $N$. The odds ($Os$) were computed for each of these 4679 non-redundant sequences from the 189 249 5' SSs. The $Os$ was defined as a square root of $(M/N)/(1 - M/N)$, where $M$ is the number of occurrences of a particular splicing sequence in the source data $N$. If a splicing sequence did not appear in the source data $N$, the $Os$ were defined as a square root of $(0.25/N)/(1 - 0.25/N)$ to avoid the infinity caused by odds ratio calculations. Note 0.25 as a parameter quoted from Sahashi’s study (10). To increase the computation speed for $Os$, all 5' SS sequences in the source data $N$ were permuted for each splicing sequence. This was followed by pre-computing and indexing of all $Os$ in the database to efficiently retrieve $Os$ for any given splicing sequence. After mutation occurrence, an improved odds ratio (Or) was defined as the $Os$ value of a cryptic 5' SS divided by the $Os$ value of its paired authentic 5' SS. Finally, the Or value was used to assess the activation likelihood for a cryptic 5' SS.

(3) The SPO value was defined as the $Sc$ value multiplied by the Or value. Finally, the SPO value was used as $\Delta$SPO score for identifying activated cryptic 5' SSs.

**Performance analysis**

The performance of the proposed SPO algorithm in the identification of activated cryptic 5' SSs was evaluated with the other seven reported approaches, that is, S&S (13), WMM (14), MM (15), MaxeEnt (16), MDD (17), $R_1$ (10) and $\Delta G$ (32). Comparative evaluation was conducted by using a 5-fold cross-validation of 490 paired splicing sequences that were included in the HMD1 data set. First, all 490 pairs of splicing sequences were divided equally into five partitions. Each partition was a testing set, and the remaining four partitions were used for training. In total, five testing sets were used, and each training set was four times the size of its corresponding testing set. The indices that were used to evaluate the performance included the following: sensitivity, specificity, accuracy, precision and $F$-measure, which may be defined as TP/(TP+FN), TN/(FP+TN), (TP+TN)/(TP+FN+TN+FP), TP/(TP+FN) and $2 \times$ (sensitivity x specificity)/(sensitivity + specificity), respectively. The TP, TN, FP and FN represented the count of true positive, true negative, false positive and false negative cases, respectively. The receiver operating characteristic (ROC) curves from the sensitivity and 1 − specificity of the eight methods were constructed based on varying delta scores for determining the activation of a cryptic 5' SS. The area under the ROC curve (AUC) was used as a measurement for their performance. In addition to these methods, Pearson’s coefficient was also used to evaluate the correlation between the predicted scores and the activation strength of cryptic 5' SSs from the HMD2 data set.

**Determining $\Delta$SPO score threshold for an activated cryptic 5' SS**

A 5-fold cross-validation of 490 paired sequences from the HMD1 data set was conducted. This 5-fold cross-validation was also used to determine the $\Delta$SPO threshold.
in the SPO algorithm. For each of the five sets of training sequences, the \( \Delta \text{SPO} \) threshold that yielded the optimal \( F \)-measure on the corresponding testing sequences was chosen. The value that corresponded to the highest occurrence of these five thresholds (to five decimal points) was designated as \( T \) for the \( \Delta \text{SPO} \) threshold in the SPO algorithm. Based on this, a cryptic 5' SS competing with its authentic 5' SS was considered activated if its \( \Delta \text{SPO} \) score was greater than \( T \), and the amount of \( \Delta \text{SPO} \) score elevated from \( T \) was used to rank the probability for such activation. If no single highest occurrence appeared from any of these five thresholds, the 5-fold cross-validation was reiterated until such a threshold was obtained.

RESULTS AND DISCUSSION
Identification of activated cryptic 5' SS by scoring methods
An HMD1 data set that contained 490 pairs of human authentic and cryptic 5' splice sequences was used for evaluating the performance of the proposed SPO algorithm (Supplementary Table S3). A threshold of \( T = 1.2214 \), previously obtained from analyzing the HMD1 data set with 5-fold cross-validation, was used to determine whether a splice site was activated. The detailed sensitivity, precision, specificity, false positive rate, accuracy and \( F \)-measure in different \( \Delta \text{SPO} \) score thresholds were shown in Figure 2. Moreover, the other seven

Figure 2. Sensitivity, specificity, precision, false positive rate, accuracy and \( F \)-measure vary with \( \Delta \text{SPO} \) score. (A) Sensitivity and precision vary with \( \Delta \text{SPO} \) score; (B) specificity and false positive rate vary with \( \Delta \text{SPO} \) score; (C) accuracy and \( F \)-measure vary with \( \Delta \text{SPO} \) score.
Table 1. Performance of scoring methods in identifying activated cryptic 5′ SSs based on 490 paired splicing sequences included in the HMD1 data set

| Method | Performance measures | Sensitivity | Specificity | Accuracy | F-measure | Precision | AUC |
|--------|----------------------|-------------|-------------|----------|-----------|-----------|-----|
| SPO    |                      | 0.823       | 0.884       | 0.857    | 0.851     | 0.849     | 0.905|
| MaxEnt |                      | 0.730       | 0.840       | 0.792    | 0.780     | 0.781     | 0.849|
| MDD    |                      | 0.712       | 0.836       | 0.782    | 0.768     | 0.774     | 0.844|
| MM     |                      | 0.744       | 0.818       | 0.786    | 0.778     | 0.762     | 0.828|
| WMM    |                      | 0.665       | 0.720       | 0.696    | 0.691     | 0.650     | 0.734|
| S&S    |                      | 0.740       | 0.695       | 0.714    | 0.714     | 0.655     | 0.782|
| Ri     |                      | 0.730       | 0.647       | 0.706    | 0.707     | 0.687     | 0.772|
| ΔG     |                      | 0.679       | 0.609       | 0.667    | 0.667     | 0.658     | 0.730|

Identification of cryptic 5′ SS of different strengths

To verify that the proposed SPO algorithm can identify cryptic 5′ SSs of various activation strengths, an HMD2 data set containing 52 data pairs from two competition assays (11) was used, including 12 strong, 26 intermediate and 14 weak 5′ SSs, according to various activation levels (11). Based on the comparison for the performance of the other seven methods (Table 3), the SPO algorithm consistently achieved a high accuracy in all of the three groups and yielded the highest accuracy when the three groups of data were pair wisely combined as used in Roca’s study (11).

A Pearson’s coefficient ($r$ value) was computed between these two variables by using the HMD2 test data (consisting of two competition assays CS-I and CS-II, each of which included 26 authentic and cryptic 5′ SS data pairs) to correlate the strength of cryptic 5′ SS activation with the predicted ΔSPO scores. Table 4 summarizes the resulting $r$ values for ΔSPO, ΔMaxEnt, ΔMDD, ΔMM, ΔWMM, ΔS&S, ΔR, and ΔΔG scores, in which the SPO algorithm displayed a greater degree of correlation than the others. In particular, the SPO algorithm appeared to perform efficiently for both CS-I and CS-II assays; however, all the other seven methods demonstrated relatively inferior performance for CS-I assay than for CS-II assay. It is known that wild types of authentic 5′ SSs were used in CS-I assay, but weakened types of authentic 5′ SSs were used in the CS-II assay (11). In in vitro experiments, the average activation of cryptic 5′ SSs was considerably stronger ($P = 6.13E-07$) in the CS-II assay than in the CS-I assay. Therefore, activation of cryptic 5′ SSs in the CS-II assay is easier than in the CS-I assay. In summary, the SPO algorithm was able to correctly predict the activation of a cryptic 5′ SS as well as to infer the activation level by evaluating the increase of ASPO score from its threshold. With this feature, it is reasonable to verify the cryptic 5′ SS activation by ranking the ΔSPO scores, when a number of splicing pairs were available for consideration. In other words, SPO algorithm can be used to predict novel cryptic 5′ SSs, especially when sequencing data (like RNA-seq data) is not available.

DNA structural profiles as an impact factor in cryptic 5′ SS

To analyze whether DNA structural profiles extracting from the hydroxyl radical cleavage pattern can improve the identification of activated cryptic 5′ SSs, the HMD1 data set and HMD2 data set were used to estimate the effect of structural profiles. First, without the inference from structural profiles, the identification for activated cryptic 5′ SSs from HMD1 data set decreased by 7.9% in sensitivity, 4.4% in specificity, 5.9% in accuracy, 6.2% in F-measure, 6.2% in precision and 5.7% in AUC (corresponding to the result in Table 1). Second, without using structural profiles, the SPO algorithm obtained a lower degree (82%) of correlation between the strength of cryptic 5′ SS activation and ΔSPO score (corresponding to the result in Table 4), and its accuracy decreased to 0.865 for the analysis of the 52 data pairs from HMD2.
Interestingly, the DNA structural profiles can also improve the 2, 2, 2, 4, 5, 6 and 1% degrees of correlation between the strength of cryptic 5′ SS activation and score from MaxEnt (16), MDD (17), MM (15), S&S (13), WMM (14), $R_i$ (10) and $\Delta G$ (32), respectively, for the analysis of the HMD2 data set. The improvement for the seven methods was based on the use of the $Sc$ value as a weight factor to multiply the original scores from these compared approaches. For example, an improved $\Delta$MaxEnt score was defined as the $\Delta$MaxEnt score multiplied by the $Sc$ value. The scores for the other methods were improved by using the same strategy. These results indicate that DNA structural profiles derived from the hydroxyl radical cleavage pattern can

Table 2. Accuracy of scoring methods in different mutant categories

| Mutant type    | SPO   | MaxEnt | MDD   | MM    | WMM   | S&S   | $R_i$ | $\Delta G$ |
|----------------|-------|--------|-------|-------|-------|-------|-------|------------|
| Point mutation | 0.822 | 0.723  | 0.708 | 0.738 | 0.535 | 0.629 | 0.728 | 0.678      |
| Deletion       | 0.889 | 0.889  | 0.778 | 0.889 | 0.778 | 0.778 | 0.778 | 0.667      |
| Duplication    | 1.000 | 1.000  | 1.000 | 1.000 | 0.000 | 0.000 | 1.000 | 1.000      |
| Insertion      | 0.667 | 0.667  | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667      |
| Total          | 0.772 | 0.679  | 0.665 | 0.693 | 0.502 | 0.591 | 0.684 | 0.637      |

Table 3. Accuracy of scoring methods in competition assays based on 52 (12 strong, 26 intermediate and 14 weak) paired splicing sequences in the HMD2 data set

| Data type       | SPO   | MaxEnt | MDD   | MM    | WMM   | S&S   | $R_i$ | $\Delta G$ |
|-----------------|-------|--------|-------|-------|-------|-------|-------|------------|
| Strong          | 1.000 | 0.833  | 1.000 | 0.917 | 0.917 | 0.917 | 1.000 | 0.917      |
| Intermediate    | 0.769 | 0.654  | 0.692 | 0.692 | 0.615 | 0.731 | 0.654 | 0.654      |
| Weak            | 1.000 | 1.000  | 0.929 | 1.000 | 0.929 | 1.000 | 0.714 | 0.929      |
| Strong and inter | 0.842 | 0.711  | 0.789 | 0.763 | 0.711 | 0.789 | 0.763 | 0.737      |
| Strong and weak | 0.850 | 0.775  | 0.775 | 0.800 | 0.825 | 0.825 | 0.675 | 0.750      |
| Total           | 0.885 | 0.788  | 0.827 | 0.827 | 0.769 | 0.846 | 0.750 | 0.788      |

Figure 3. Comparison of predictive accuracy of the scoring methods for identifying activated cryptic 5′ SSs. (A) Sensitivity versus 1 − specificity for the scoring methods; (B) false positive rate versus false negative rate for the scoring methods.
improve the identification of activated cryptic 5′ SSs in human mutation cases.

Although the effect of DNA structural profiles was useful for identifying activated cryptic 5′ SSs, the detailed relationship between the DNA structural profiles and the cryptic 5′ SSs is unclear. One possible explanation could be that the changes of the DNA structural profiles at either the cryptic 5′ SS or the corresponding authentic 5′ SS may respond to the strength of cryptic 5′ SS activation. On the other hand, the changes of DNA structural profiles may be involved in non-intronic splicing mechanism when mutation occurs on the DNA level. Some non-intronic splicing information was assumed to play a vital role in shaping the split structure of eukaryote genes (7). Consequently, the DNA structural profiles may improve the identification of cryptic 5′ SSs in eukaryote genes.

CONCLUSION

This study proposes the SPO algorithm that combined structural profiles with odds measure to obtain the ΔSPO score for identifying the activated cryptic 5′ SSs. Based on the results, the SPO algorithm yields a superior identification of cryptic 5′ SSs than that by the other seven methods, and its ΔSPO score also provides information to estimate the inherent strength of 5′ SSs in human mutation data. In practical application, the SPO algorithm can be used as a powerful tool for designing mutagenesis experiments of various splicing events and can be used to study the influences of activated cryptic 5′ SSs in the field of amino acid changes in human hereditary diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–4.

ACKNOWLEDGEMENTS

We thank for the comments from the anonymous reviewers. The experimental data provided by Roca, Rogan and Buratti’s studies are also appreciated.

FUNDING

National Science Council of Taiwan (Grant No: NSC99-2627-M-001-005-MY3; 99-2621-B-001-005-MY2). Funding for open access charge: Biodiversity Research Center, Academia Sinica, Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. Baralle.D. and Baralle,M. (2005) Splicing in action: assessing disease causing sequence changes. J. Med. Genet., 42, 737–748.
2. Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum. Genet., 90, 41–54.
3. Nakai,K. and Sakamoto,H. (1994) Construction of a novel database containing aberrant splicing mutations of mammalian genes. Gene, 141, 171–177.
4. Chen,X., Truong,T.T., Weaver,J., Bove,B.A., Cattie,K., Armstrong,B.A., Daly,M.B. and Godwin,A.K. (2006) Intrinsic alterations in BRCA1 and BRCA2: effect on mRNA splicing fidelity and expression. Hum. Mutat., 27, 427–435.
5. Guimarães,C.P., Lemos,M., Menezes,I., Coelho,T., Sá-Miranda,C. and Azevedo,J.E. (2001) Characterisation of two mutations in the ABCD1 gene leading to low levels of normal ALDP. Hum. Genet., 109, 616–622.
6. Roca,X., Sachidanandam,R. and Kainer,A.R. (2003) Intrinsic differences between authentic and cryptic 5′ splice sites. Nucleic Acids Res., 31, 6321–6333.
7. Kapustin,Y., Chan,E., Sarkar,R., Wong,F., Vorechovsky,I., Winston,R.M., Tatusova,T. and Dibb,N.J. (2011) Cryptic splice sites and split genes. Nucleic Acids Res., 39, 5837–5844.
8. Ragan,P.K., Faux,B.M. and Schneider,T.D. (1998) Information analysis of human splice site mutations. Hum. Mutat., 12, 153–171.
9. Nallà,V.K. and Rogan,P.K. (2005) Automated splicing mutation analysis by information theory. Hum. Mutat., 25, 334–342.
10. Sahashi,K., Masuda,A., Matsuura,T., Shimj,J., Zhang,Z., Takeshima,Y., Matsuo,M., Sobue,G. and Ohno,K. (2007) In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5′ splice sites. Nucleic Acids Res., 35, 5995–6003.
11. Roca,X., Sachidanandam,R. and Kainer,A.R. (2005) Determinants of the inherent strength of human 5′ splice sites. RNA, 11, 683–698.
12. Buratti,E., Chivers,M., Královičová,J., Romano,M., Baralle,M., Kainer,A.R. and Vorechovsky,J. (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.
13. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of human genes: causes and consequences. Hum. Genet., 73, 41–54.
14. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res., 12, 505–519.
15. Salzberg,S.L. (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. Comput. Appl. Biosci., 13, 365–376.
16. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol., 11, 377–394.
17. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268, 78–94.
18. Tullius,T.D. and Dombroski,B.A. (1985) Iron(II) EDTA used to measure the helical twist along any DNA molecule. Science, 230, 679–681.
19. Shcherbakova,I. and Brenowitz,M. (2008) Monitoring structural changes in nucleic acids with single residue spatial and millisecond time resolution by quantitative hydroxyl radical footprinting. Nat. Protoc., 3, 288–302.
20. Shafer, G.E., Price, M.A. and Tullius, T.D. (1989) Use of the hydroxyl radical and gel electrophoresis to study DNA structure. *Electrophoresis*, **10**, 397–404.

21. Price, M.A. and Tullius, T.D. (1992) Using hydroxyl radical to probe DNA structure. *Methods Enzymol.*, **212**, 194–219.

22. Jain, S.S. and Tullius, T.D. (2008) Footprinting protein-DNA complexes using the hydroxyl radical. *Nat. Protoc.*, **3**, 1092–1100.

23. Tullius, T.D. and Dombroski, B.A. (1986) Hydroxyl radical “footprinting”: high-resolution information about DNA-protein contacts and application to lambda repressor and Cro protein. *Proc. Natl Acad. Sci. USA*, **83**, 5469–5473.

24. Viola, I.L. and Gonzalez, D.H. (2011) Footprinting and missing nucleoside analysis of transcription factor-DNA complexes. *Methods Mol. Biol.*, **754**, 259–275.

25. Stoltzfus, A. (2008) Evidence for a predominant role of oxidative damage in germline mutation in mammals. *Mutat. Res.*, **644**, 71–73.

26. Nakken, S., Rødland, E.A. and Hovig, E. (2010) Impact of DNA physical properties on local sequence bias of human mutation. *Hum. Mutat.*, **31**, 1316–1325.

27. Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D. and Margulies, E.H. (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, **324**, 389–392.

28. Greenbaum, J.A., Pang, B. and Tullius, T.D. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res.*, **17**, 947–953.

29. Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J. and Cooper, D.N. (2007) Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.*, **28**, 150–158.

30. Divina, P., Kvítkovcová, 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.

31. Dogan, R.I., Getoor, L., Wilbur, W.J. and Mount, S.M. (2007) SplicePort—an interactive splice-site analysis tool. *Nucleic Acids Res.*, **35**, W285–W291.

32. Markham, N.R. and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–W581.