Published online 20 February 2016

Nucleic Acids Research, 2016, Vol. 44, No. 10 ***e91***
doi: 10.1093/nar/gkw104

SRAMP: prediction of mammalian N^6-methyladenosine (m^{6}A) sites based on sequence-derived features

Yuan Zhou[1,2,3,*], Pan Zeng[1,2,3], Yan-Hui Li[1,2,3], Ziding Zhang[4] and Qinghua Cui[1,2,3,*]

[1]Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University, Beijing 100191, China, [2]MOE Key Lab of Molecular Cardiovascular Sciences, Peking University, Beijing 100191, China, [3]Center for Noncoding RNA Medicine, Peking University Health Science Center, Beijing 100191, China and [4]State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China

Received October 16, 2015; Revised February 10, 2016; Accepted February 11, 2016

ABSTRACT

N^6-methyladenosine (m^{6}A) is a prevalent RNA methylation modification involved in the regulation of degradation, subcellular localization, splicing and local conformation changes of RNA transcripts. High-throughput experiments have demonstrated that only a small fraction of the m^{6}A consensus motifs in mammalian transcriptomes are modified. Therefore, accurate identification of RNA m^{6}A sites becomes emergently important. For the above purpose, here a computational predictor of mammalian m^{6}A site named SRAMP is established. To depict the sequence context around m^{6}A sites, SRAMP combines three random forest classifiers that exploit the positional nucleotide sequence pattern, the K-nearest neighbor information and the position-independent nucleotide pair spectrum features, respectively. SRAMP uses either genomic sequences or cDNA sequences as its input. With either kind of input sequence, SRAMP achieves competitive performance in both cross-validation tests and rigorous independent benchmarking tests. Analyses of the informative features and overrepresented rules extracted from the random forest classifiers demonstrate that nucleotide usage preferences at the distal positions, in addition to those at the proximal positions, contribute to the classification. As a public prediction server, SRAMP is freely available at http://www.cuilab.cn/sramp/.

INTRODUCTION

With recent advances in genomics and molecular biology, the catalogue and functional importance of RNA modifications are being revealed (1). Among ∼150 types of known RNA modifications (2), N^6-methyladenosine (m^{6}A), the methylation modification on the nitrogen at the 6th position of the adenosine base, stands out due to its prevalent existence and extensive functional impacts (3,4). The prevalence of m^{6}A is two-folded: on the one hand, m^{6}A appears in nearly all kinds of RNA transcripts, whether coding or non-coding (5–7); on the other hand, m^{6}A is enriched near the stop codon (5,6), but also dispreads along all parts of a pre-mRNA, including coding sequence, un-translated regions (UTRs) and introns (8–10). At the same time, as a versatile molecular tag, m^{6}A modification is involved in a variety of important biological processes, including but not limited to RNA localization and degradation (11), RNA structure dynamics (12), alternative splicing (9), primary microRNA processing (7), cell differentiation and reprogramming (13,14) and regulation of circadian clock (15).

Knowledge about the positions of m^{6}A sites plays essential roles in investigating the mechanisms and functions of this modification. Independent evidence has emerged to validate the DRACH (where D = A, G or U; R = A or G; H = A, C or U) consensus motif and the GAC consensus motif surrounding the m^{6}A sites from mammalian and yeast transcriptomes, respectively (5,6,16–18). However, as such short motifs can be frequently observed in one genome, identifying exact positions of m^{6}A sites in transcripts is still challenging. Currently, high-throughput experimental identifications of m^{6}A sites heavily rely on next-generation sequencing-based techniques like MERIP (5) and m^{6}A-seq (6). Such techniques are able to detect tens of thousands of m^{6}A-containing sequence fragments of ∼100 nt length from the transcriptome, but their resolutions are not fully satisfying, i.e. these methods cannot exactly point out which adenosine is methylated (19). As a result, until recently, there was no computational tool available for predicting m^{6}A site from sequences, due to the lack of gold standard datasets. In 2013, Schwartz *et al.* further improved these

*To whom correspondence should be addressed. Tel: +86 10 82801585; Fax: +86 10 82801001; Email: cuiqinghua@hsc.pku.edu.cn
Correspondence may also be addressed to Yuan Zhou. Tel: +86 10 82801585; Fax: +86 10 82801001; Email: soontide6825@163.com

techniques to produce a near single-nucleotide resolution map of m^{6}A sites in yeast genome (16). With this higher resolution data, they proposed a computational method to predict m^{6}A sites using the nucleotide composition, local secondary structure stability and relative position in gene as the input features (16). This method achieves promising performance in cross-validation tests, but no public tool implementing this method has been made available. Subsequently, Chen *et al.* have established two yeast m^{6}A site prediction servers, i.e. m6Apred (20) and iRNA-Methyl (21). Both predictors are support vector machine-based but trained with different sequence encoding scheme: m6Apred considers chemical property of nucleotide and accumulated nucleotide frequency as its input features (20), while iRNA-Methyl represents RNA sequences using pseudo nucleotide composition features (21). It has been shown that both predictors exhibit considerable accuracy in cross-validation tests on yeast datasets, but whether they can predict mammalian m^{6}A sites has not been tested.

More recently, He and co-workers have significantly improved the resolution of m^{6}A detecting techniques by developing the PA-m6A-seq technique (18). Subsequently, Jaffrey and co-workers have devised a novel technique termed miCLIP (17) and provide the single-nucleotide resolution map of the m^{6}A sites across human transcriptome, giving us an unprecedented opportunity to construct a computational m^{6}A site predictor. In this study, we establish a mammalian m^{6}A site predictor named SRAMP (sequence-based RNA adenosine methylation site predictor) under the random forest machine learning framework. As its name implies, SRAMP considers the sequence-derived features only, including the positional binary encoding of nucleotide sequence, the K-nearest neighbor (KNN) encoding and the nucleotide pair spectrum encoding. Nevertheless, SRAMP shows promising performance in both cross-validation tests and independent benchmarking tests. Analyses of the informative features and rules extracted from the random forest classifiers demonstrate that the non-random nucleotide usage, at both the proximal and the distal positions, plays roles in distinguishing m^{6}A sites. In the following sections, we will first describe how the SRAMP was established. The performance assessment and server implementation will be subsequently described.

MATERIALS AND METHODS

Datasets

The positive samples (m^{6}A sites) were extracted from the recently published single-nucleotide resolution maps of mammalian m^{6}A sites (17,22), and only the m^{6}A sites that conform to the DRACH consensus motifs were retained. We further mapped these m^{6}A sites to the human and mouse transcripts recorded by the ENSEMBL database (http://www.ensembl.org, queried in July 2015). If multiple transcripts from the same locus harboured m^{6}A sites, only the longest transcript with the largest number of m^{6}A sites was retained. As for the negative samples (non-m^{6}A sites), the non-methylated adenosines that conform to the DRACH motif were randomly selected from the same set of methylated transcripts. Because the m^{6}A sites are not randomly distributed along the transcripts (3,4,22), to avoid prominent bias, we assigned 10-fold likelihood to be chosen as the negative sample to a non-m^{6}A site near a known m^{6}A site, enabling position-corrected sets of negative samples. Considering the fact that there are much more non-m^{6}A sites than m^{6}A sites, we kept a 1:10 positive-to-negative ratio in our dataset, such a highly unbalanced ratio coordinates the dataset coverage and computational burden. Note that, two prediction modes were built in SRAMP, i.e. the full transcript mode and the mature mRNA mode. The full transcript mode used the genomic sequences as its input, while the mature mRNA mode sequences considered cDNA sequences instead. For either mode, the training samples were extracted from the same 13 500 transcripts (i.e. randomly selected 80% of the total), while samples from the other 3391 transcripts were allocated to the independent testing dataset (see Supplementary Tables S1–4 for these datasets). To test the potential influence of sequence redundancy, we also employed CD-HIT-EST tool (23) to remove the redundant independent testing samples. One testing sample was considered as redundant one if it shares high sequence identity either with a training sample or with another testing sample. Four sequence identity thresholds, i.e. 95, 90, 85 and 80%, were applied, among which the 80% identity is the most rigorous threshold provided by CD-HIT-EST.

To compare SRAMP with previously published yeast m^{6}A site predictors m6Apred (20) and iRNA-Methyl (21), we compiled two benchmarking datasets for yeast and mammalian, respectively. As for the yeast benchmarking dataset, we downloaded the m6Apred's independent testing dataset and retained the samples which met two criteria: (i) the sample should map onto one yeast cDNA sequence and (ii) there should be a 51-nt sequence window (25 nt on each side) available surrounding the central adenosine, as required by iRNA-Methyl server. Consequently, the yeast benchmarking dataset contains 370 positive samples and 1750 negative samples (Supplementary Table S5). The mammalian benchmarking dataset was extracted from the independent testing dataset for the mature mRNA mode predictor of SRAMP. We noted that yeast samples conform to the GAC consensus motif surrounding the central adenosine, but mammalian samples allow either the GAC or the AAC consensus motifs. To ensure fair comparison, an additional criterion, i.e. the sample should conform only to the GAC consensus motif, was applied to filter the mammalian samples, resulting in a mammalian benchmarking dataset containing 8378 positive samples and 65 562 negative samples (Supplementary Table S6).

We also tested whether SRAMP can predict the binding sites of YTHDF1 and YTHDF2, two known RNA-binding proteins that preferentially recognize m^{6}A modification sites (11,24). DRACH motifs inside the experimentally identified YTHDF binding regions were assigned as the positive samples. If multiple motifs existed in the same region, we only considered the one with the highest prediction score. Negative samples were randomly picked outside the YTHDF binding region, keeping the 1:10 positive-to-negative ratio (see Supplementary Tables S7 and S8).

Feature encoding scheme

Generally speaking, SRAMP encoded an m^{6}A/non-m^{6}A site by extracting sequence or predicted secondary structure features from a W nt flanking window where the m^{6}A/non-m^{6}A site settled at the central position. The window size W varied between different encoding schemes and different prediction modes, and were optimized through 5-fold cross-validation tests. The optimized window sizes were listed in Supplementary Table S9. Note that, if an m^{6}A/non-m^{6}A site appeared near one terminus of the transcript, the flanking window was truncated at the transcript terminus for the nucleotide pair spectrum encoding, but completed with gaps in the cases of the other encodings to ensure flanking windows of the fixed size. Details about each encoding are described below:

Positional binary encoding of nucleotide sequence (binary encoding). This encoding exactly depicts the nucleotide at each position in the flanking window. The A, C, G, U and the gap character filling the sequence termini were translated as a binary vector of (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1) and (0,0,0,0), respectively. Therefore, the binary encoding of a W nt flanking window should result in a W*4-dimensional feature vector.

K-nearest neighbor encoding (KNN encoding). This encoding depicts how much the flanking window of one query sample resembles those of other m^{6}A sites. Due to the huge size of the training dataset (~0.5 million samples), it was computational prohibitive to compare query samples with every training sample. Instead, the training samples were first grouped according to their 21 nt flanking windows, and 5000 reference positive samples and 50 000 reference negative samples were randomly selected from the training dataset, keeping the fraction of each group. Then the flanking window of the query sample was firstly compared with all reference samples to obtain pair-wise similarity scores:

$$\text{Pair}-\text{wise similarity} = \sum_{i=1}^{W} \text{NUC44}\,(q_i, r_i) \qquad (1)$$

where q_i and r_i are the nucleotides at the ith position of the query sample and the reference sample's flanking windows, respectively. W is the window size. The NUC44 is a common nucleotide similarity scoring matrix given as +5 when matched, -4 when mismatched, -2 when one is terminal gap and -1 when both are terminal gaps. Then, the fraction of positive samples (FoP) in the top K most similar reference samples was taken as the KNN feature. The considered Ks were preliminarily optimized as (50, 100, 150, …, 1350) for the full transcript mode, corresponding to the (1%, 2%, 3%, …, 30%) of the total of positive reference samples. For mature mRNA mode, the considered Ks were (50, 100, 150, …, 1500).

Nucleotide pair spectrum encoding (spectrum encoding). This encoding depicts the sequence context of an m^{6}A/non-m^{6}A site by calculating the frequencies of all possible d-spaced nucleotide pairs (e.g. UxxxG is a three-spaced nucleotide pair) inside a flanking window. That is, this encoding examines throughout the nucleotide pair spectrum in a flanking window. Just like the analogous encoding scheme for the amino acid sequence (25,26), the frequency of a spaced nucleotide pair np_i was calculated as

$$\text{Frequency}\,(np_i) = \frac{C\,(np_i)}{W - d - 1} \qquad (2)$$

where $C(np_i)$ is the count of np_i inside a flanking window, W is the window size and d is the space between two nucleotides, ranging from 0 to $d_{\max}$. Therefore, the nucleotide pair spectrum encoding would denote the flanking window as a 4*4* ($d_{\max}$+1) dimensional vector. The optimized $d_{\max}$ was 3 for both prediction modes.

Predicted secondary structure pattern. For a preliminary test of RNA secondary structure features, we employed this encoding to depict the predicted secondary structure status at each position. The secondary structures were predicted by the RNAfold tool (version 2.1.9) in ViennaRNA package (27) with default parameters. Because it is very time-consuming to predict the secondary structures of full RNA transcripts, we instead extracted a 2001-nt local sequence window (truncated at transcript termini) centred at an m^{6}A/non-m^{6}A site, as the input to RNAfold. The RNAfold tool outputted the predicted secondary structure in dot-bracket notation where unpaired and paired positions were denoted as dots and brackets, respectively. To obtain a more specific description of RNA secondary structures, the secondary structures were further classified into hairpin loop (H), multiple loop (M), interior loop (I), paired (P) and bulged loop (B). Finally, the secondary structure status H, M, I, P, B and the terminal gap were encoded as the binary vectors as (1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), (0,0,0,0,1) and (0,0,0,0,0), respectively. Therefore, the predicted secondary structure pattern of a W nt flanking window constituted a W*5 dimensional feature vector.

Random forest classifier training and performance assessment

In brief, SRAMP integrates multiple random forest classifiers that were trained with different feature encodings (Figure 1). We noted that the positive-to-negative ratio of our training datasets was highly unbalanced (1:10). Such an unbalanced training dataset was an unfavourable choice for machine learning classifiers. Therefore, for each training dataset encoded by one specific encoding scheme, we created 10 subsets of training data with 1:1 positive-to-negative ratio by randomly splitting the negative samples into 10 parts. Subsequently, 10 random forest models (sub-classifiers) were trained and the average output score from these 10 sub-classifiers was taken as the prediction score of the random forest classifier. The random forest sub-classifiers were trained by the *randomForest* package in R (28), and the tree number in each sub-classifier was preliminarily optimized to 300. Finally, the prediction scores of the random forest classifiers trained with different feature encodings were combined using the weighted summing formula below:

$$S_{\text{combined}} = \sum\nolimits_{1}^{n} \alpha_i \cdot S_i \qquad (3)$$

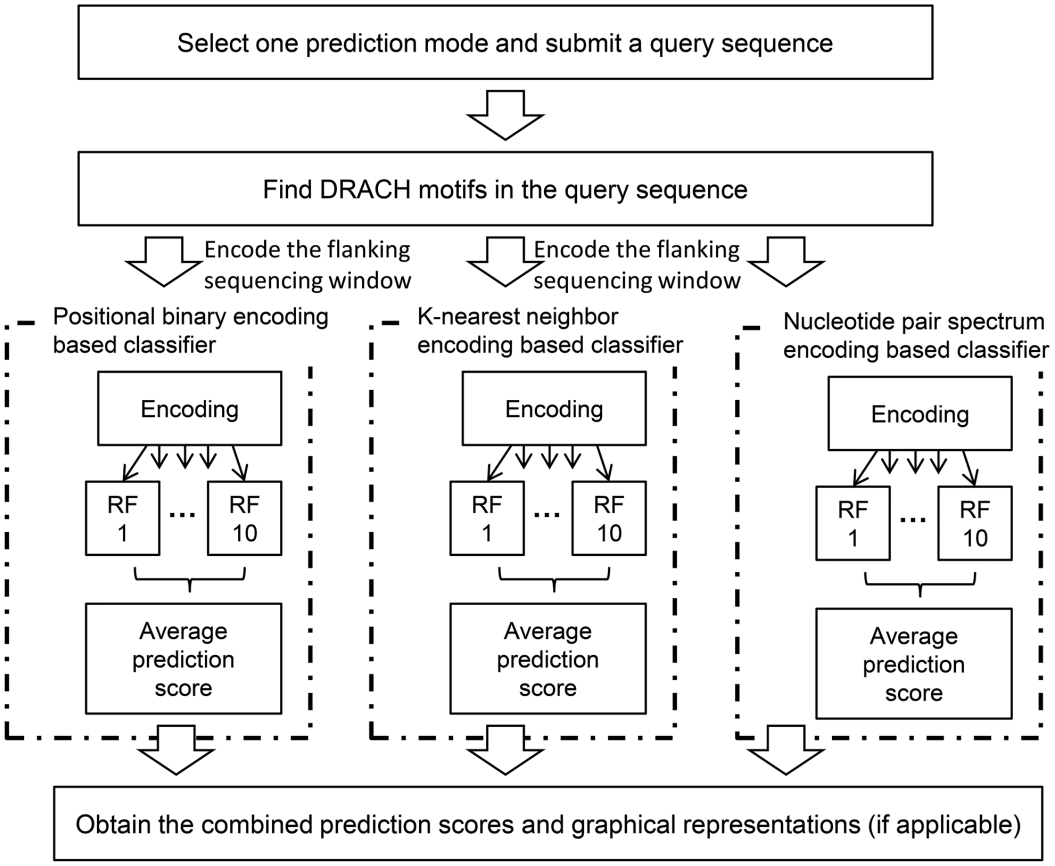


Figure 1. The computational framework of SRAMP. Two prediction modes have been built in SRAMP, i.e. the full transcript mode and the mature mRNA mode. Both prediction modes adopt the same computational framework. First, for a DRACH motif presented in the query sequence, its flanking sequence window is extracted and represented using the three sequence-based encodings. Then the encoded features will be submitted to the corresponding random forest classifiers. Each random forest classifier summarizes the output scores from 10 sub-classifiers, which were trained on all positive samples and a distinct subset of negative samples in the training dataset. Finally, the prediction scores of the random forest classifiers are combined through weighted summing formula. Four stringency thresholds correspond to the 99%, 95%, 90% and 85% specificities in 5-fold cross-validation test that are used to judge the classification and associated confidence. If analysing secondary structure function is enabled, the secondary structure context of the predicted m^6A sites will be also provided.

where the S_i and α_i are the prediction score and the weight for the classifier trained with the ith encoding, respectively. n is the total number of the classifiers taken into account. The optimized weights were also listed in Supplementary Table S9.

Once the random forest classifiers were trained, we employed both 5-fold cross-validation tests on the training dataset, and the independent tests to assess our predictors. We used sensitivity, specificity and Matthews correlation coefficient (MCC) to measure the predictor's performance at certain thresholds. These parameters read

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \tag{6}$$

where TP, TN, FP and FN represent the counts of true positive, true negative, false positive and false negative predictions, respectively. We also plotted the ROC curves (which plot sensitivity against 1-specificity) for the predictors and calculated the area under ROC curve (AUROC) to evaluate the overall performance of the predictors. The AUROC ranges from 0 to 1. An AUROC near 1 implies perfect predictions while the AUROC of random guess is 0.5. Finally, the area under precision-recall curve (AUPRC) was calculated to examine the performance of predictors when restricting low false positive rates. The precision-recall curves plot precision (the fraction of TP in all predicted positives) against recall (sensitivity). This curve is more sensitive to false positives than ROC curve.

Extraction of overrepresented rules and informative features

By definition, a random forest model is comprised by a series of decision trees (29). One decision tree contains single root node, several leaf nodes that denote the final decisions and many intermediate nodes that describe the conditions supporting the final decisions (30). A path from the root node to the leaf node is called a rule. Intuitively, a rule depicts how a set of features collaborate to classify

the samples. We extracted the rules from the random forest models by using the *inTrees* package in R (https://cran.r-project.org/web/packages/inTrees). Due to the limitation of our computational resource, we only considered rules that contained 15 or less intermediate nodes. To obtain a non-redundant, overrepresented set of rules that predict m^6A sites, we first discarded the rules which covered <100 m^6A sites or showed no more than 2-fold enrichment for m^6A sites. Then the significance of a retained rule was evaluated by F1-score given by

$$\mathrm{F1-score} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad (7)$$

Finally, if there were two redundant rules, only the one with higher F1-score was retained. The redundancy between two rules R_1 and R_2 was measured by the Jaccard index (JI) given by

$$\mathrm{JI} = \frac{|\boldsymbol{P}_1 \cup \boldsymbol{P}_2|}{|\boldsymbol{P}_1 \cap \boldsymbol{P}_2|} \quad (8)$$

where $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$ are the subsets of m^6A sites covered by the rules R_1 and R_2, respectively. Any pair of rules shares a JI larger than 0.001 would be considered as redundant ones.

We also extracted the informative features from the spectrum encoding-based random forest classifier. After one random forest sub-classifier was trained, an importance score was assigned to each feature. We used the average importance score from the 10 sub-classifiers to measure how much a feature is informative.

Online server construction

The SRAMP online server was built under the 'Linux+Apache+Django' framework. The visualization of the structural context of the predicted m^6A sites was powered by the VARNA structure visualization tool (31).

RESULTS AND DISCUSSION

Establishment of the predictors

As mentioned in the 'Materials and Methods' section, two prediction modes were built in SRAMP, i.e. the full transcript mode and the mature mRNA mode. We focus on describing the establishment of the predictor for the full transcript mode, as the predictor for the mature mRNA mode was established in the same way.

The training dataset is comprised the experimentally identified m^6A sites and randomly selected non-m^6A sites from the same transcripts. Both m^6A sites and the non-m^6A sites conform to the DRACH consensus motif. To distinguish methylated DRACH motifs from non-methylated ones, an intuitive way is to describe the flanking nucleotide sequence as is. The positional binary encoding of nucleotide sequence (binary encoding) exactly translates the nucleotide at each position into a binary vector and has been widely employed to build predictors for the protein and RNA modification sites (32–35). Here, the binary encoding is introduced to depict the nucleotide sequences in the 61 nt (30 nt on each side of the m^6A/non-m^6A sites) flanking windows. The random forest classifier using the binary encoding achieves encouraging performance in the 5-fold cross-validation test on the training dataset (Figure 2A; AUROC = 0.813). When focusing on the performance at low false positive rate, the performance of binary encoding classifier is also competitive (Figure 2B; AUPRC = 0.374), indicating that the positional sequence pattern is a strong predictor of m^6A site. Indeed, two sample logos that visualize the relatively more or less favoured nucleotide also demonstrate a weak but prevalent nucleotide usage preference around m^6A sites (Supplementary Figure S1A). The binary encoding describes such weak nucleotide usage preference at individual position, and the random forest classifier further exploits their combinations, enabling more powerful classifications. For a straightforward demonstration of the working principle underlying the random forest classifier, we extracted the overrepresented rules from it (see 'Materials and Methods' section for details). Each rule can be interpreted as a specific combination of feature value conditions that discriminate m^6A sites from non-m^6A sites. After removal of redundant rules, 37 overrepresented rules were obtained (Supplementary Table S10). The most profound nucleotide preferences observed in the two sample logos (Supplementary Figure S1A) are also observed among the rules. For example, G at the −2 and −1 positions and U at the +2 position are frequently presented as one of the conditions that constitute a rule. The rules also indicate the favoured combination of nucleotides at different positions. For example, the two sample logos indicate that the A at the -1 position is less favoured when compared with negative samples. Nevertheless, when checking through the proximal sequences that are more favoured by m^6A sites than non-m^6A sites (Supplementary Figure S1B), one can easily point out that A at the -1 position is certainly allowed, especially when a G or A is presented at the −3, −2, +3 or +4 position. Such specific combinations can also be captured by the overrepresented rules (e.g. rules #6, #10, #12, #16, #26 and #30), indicating the prediction capability of our classifier for the AAC sites. Indeed, the binary encoding predictor predicts AAC sites equivalently well as the GAC sites (AUROC = 0.813 and 0.814, respectively). Finally, the rules also clearly demonstrate the contribution of nucleotide preference at the distal positions, as the nucleotide preferences at −29, −28, −23, −21, −7, +13, +16, +20, +26, +29 positions are considered by at least 9 out of 37 overrepresented rules. Generally, the requirements for nucleotide usage at distal positions are relaxed, but exceptions also exist. For example, rule #14 exactly required a U presented at the -28 position and an A at the −14 position. Last but not least, as indicated by the F1-scores, it is not possible for any single rule to accurately predict m^6A sites. It is the random forest model that integrates a large ensemble of rules to achieve the robust performance.

The above analysis of non-redundant overrepresented rules also indicates that the m^6A sites tend to form diverse clusters among which the similar sequence pattern is followed. The binary encoding is not straightforward enough to demonstrate the sequence similarity between the m^6A sites. Therefore, we introduced the KNN encoding, which has been successfully applied for predicting protein phos-

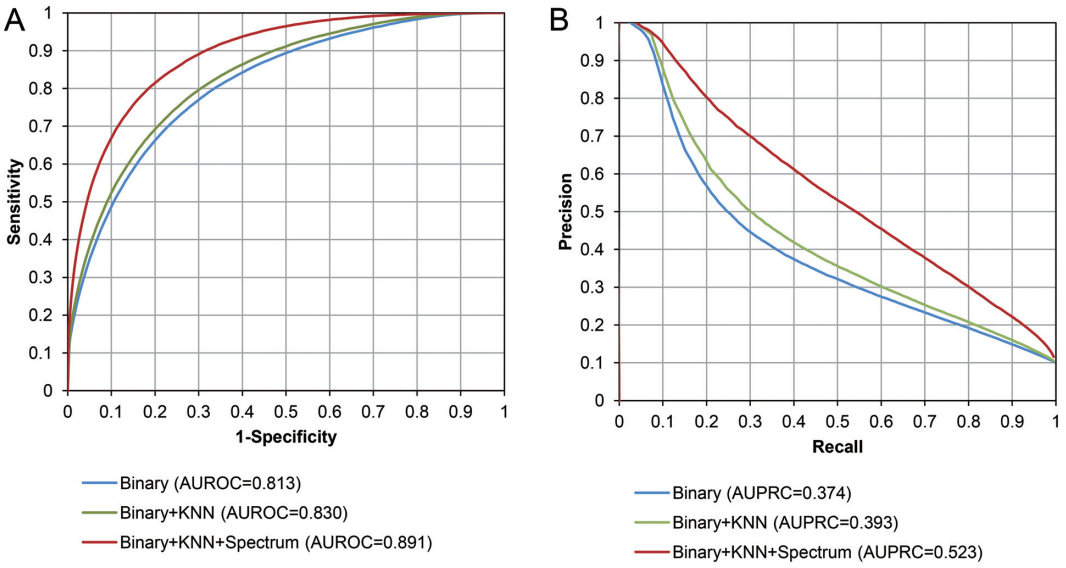


Figure 2. The overall performances of the full transcript mode classifiers based on the results from 5-fold cross-validation tests. The performances are illustrated by the ROC curves (**A**) and the precision-recall curves (**B**).

phorylation and ubiquitination sites (36,37). For a query sample, the KNN feature depicts the FoP among its *K* most similar reference samples, in other words, the FoP among its *K*-nearest neighbors in the reference dataset (a representative subset of training dataset, see 'Materials and Methods' section for details). As *K* increases, the KNN encoding can reflect weak but non-random sequence similarity between the m^{6}A sites at different levels. We adopted a series of *K*s from 50 to 1350, which corresponds to 1–27% of the positive reference samples. The random forest classifier trained with the KNN encoding exhibits a competitive performance in cross-validation (AUROC = 0.781, AUPRC = 0.297). Incorporation of the KNN encoding-based classifier also results in performance improvement (Figure 2; AUROC = 0.830, AUPRC = 0.393). We further analysed the overrepresented rules extracted from this classifier (Supplementary Table S11). Because the positive-to-negative ratio is 1:10 in our datasets, for a query m^{6}A site, FoP around $1/11 = 0.091$ can be expected if this site exhibits random similarities to both positive samples and negative samples. Supplementary Table S11 clearly demonstrates the prevalent requirement of FoP larger than 0.091, indicating the non-random sequence similarity between m^{6}A sites. However, for a query m^{6}A site, the enrichment of positive training samples among its nearest neighbors is not always necessary. Sixteen out of 44 rules require prominent enrichment of positive samples when $K =$ 50 or 100, implying tightly clustered positive samples. But for other rules, the higher FoPs are only required when *K* is large, indicating many positive samples are loosely clustered. Direct clustering of positive samples may ignore such loosely clustered positive samples, but such information can be recognized by the KNN features with larger *K*s. In all, by reflecting the (often weak) clustering tendencies between positive samples, the KNN encoding can distinguish the m^{6}A sites from random background.

Both the binary encoding and the KNN encoding exploit positional sequence patterns. Nevertheless, position-independent sequence information may also be helpful. For example, an RNA motif may be presented in the flanking window of an m^{6}A site, but its relative position to the m^{6}A sites may not be fixed. Exhaustive sampling of the overrepresented motifs around m^{6}A sites is time-consuming and sensitive to noise. An alternative solution is to use the frequency of short nucleotide words as a cryptic representation of position-independent sequence pattern. Here, the frequencies of *d*-spaced nucleotide pairs constitute the spectrum of possible nucleotide words, where any nucleotide pair spaced by *d* nucleotides in-between can be considered. Our preliminary test has indicated that considering *d* from 0 to 3 is enough to achieve stable prediction performance. The random forest predictor based on this spectrum encoding achieves an overall performance as good as the binary encoding (AUROC = 0.812, AUPRC = 0.340). After integrating the spectrum encoding, a major augment of performance is observed (Figure 2; AUROC = 0.891, AUPRC = 0.523), indicating this position-independent sequence encoding indeed supplements the position-dependent encodings. For a more straightforward illustration, we picked the top 20 informative features from the spectrum encoding-based random forest classifier and mapped them onto the flanking window of m^{6}A sites. As illustrated in Supplementary Figure S2, most of the informative features do not show strong positional bias along the flanking windows. Even when focused on more proximal positions, only the spaced nucleotide pairs that constitute the DRACH consensus motif (e.g. GA, AC) exhibit prominent enrichment near the central position. These results suggest that the spectrum encoding, at least partly, reflects the positional independent features of m^{6}A sites, and therefore could enhance the positional sequence pattern-based classifiers.

Finally, it has been indicated that the (predicted) RNA secondary structures can serve as informative features for predicting RNA functional sites like those targeted by microRNA (38,39) or vulnerable to deleterious mutations

(40). Moreover, it has been also suggested that some m^{6}A sites favour specific secondary structure context to exert their regulatory roles (9). Therefore, we also trained a random forest classifier based on the positional patterns of predicted RNA secondary structures. However, this classifier shows only weak accuracy in 5-fold cross-validation test (AUROC = 0.618), and cannot improve the predictor. Thus, current SRAMP does not consider such secondary structure features. Nevertheless, analysing the overrepresented rules from this random forest classifier could provide interesting suggestions (Supplementary Table S12). On the one hand, these rules clearly demonstrate the contribution of the secondary structure preference at the distal positions, in addition to those at proximal positions. On the other hand, the proximal secondary structure pattern also exhibits an interesting property: while the central m^{6}A sites could be in loops, one proximal position should be paired (as observed in rules #11, #12 and #13), indicating that some m^{6}A sites tend to locate near the boundary of stem-loop transition. Therefore, to facilitate the investigation of relationship between m^{6}A sites and RNA structural elements, SRAMP also allows users to analyse the local structural context of the predicted m^{6}A sites. Finally, Spitale *et al.* have recently demonstrated that the RNA structural imprints as a powerful predictor of m^{6}A sites (41). We found their structural data could not cover our dataset sufficiently, and such data remained difficult to be handled by non-specialist. Thus, at current stage, SRAMP does not consider experimental RNA structural imprints, though such data should be promising with augmented coverage and convenience in the future.

Though the above shown performance of the full transcript mode predictor is encouraging, there is also one profound limitation: the genomic sequences are required by this prediction mode. Unlike cDNA or mRNA sequences, the genomic sequence is usually not available in public nucleotide sequence database. To facilitate the users, we have also established a mature mRNA mode predictor, which considers cDNA or mRNA sequences as its input. The predictor of mature mRNA mode was established under an analogous framework to that of the full transcript mode. It is noteworthy that accurate prediction m^{6}A sites in cDNA or mRNA sequences are more challenging. First, there is evidence supporting the idea that (at least a considerable fraction of) RNA m^{6}A modification events are occurred at the pre-mRNA level (8,9,17). Discarding all introns may disrupt the original sequence context of an m^{6}A site and therefore reduce the discriminative capability of the m^{6}A site predictor. Second, the distance between an m^{6}A site and a non-m^{6}A site generally becomes closer in a cDNA sequence compared with that in the corresponding genomic sequence. As a result, the sequence features of m^{6}A sites and non-m^{6}A sites become less distinguishable because of the more overlapped flanking sequence windows. As expected, the predictors of mature mRNA prediction mode show a decreased performance in 5-fold cross-validation (Supplementary Figure S3, AUROC = 0.797, AUPRC = 0.312). Nevertheless, such performance is still competitive and acceptable for users who prefer cDNA or mRNA sequences as the input.

Assessment of the SRAMP's performance with independent datasets

To rigorously evaluate our predictors, we further tested our method on the independent testing datasets. The results from the independent tests generally agreed well with those from the cross-validation tests. When all three sequence-based random forest classifiers are combined, the full transcript mode and mature mRNA mode predictors achieve AUROCs of 0.871 and 0.794, respectively (Supplementary Figure S4A and B). The combined predictors predict GAC and AAC sites equivalently well (AUROC = 0.861 and 0.784 for GAC sites, AUROC = 0.873 and 0.790 for AAC sites). We noted that the sequence redundancy may result in overestimations of the prediction performance. To address this issue, we employed CD-HIT-EST tool (23) to remove the redundant testing samples. As shown in Supplementary Figure S5, for both predictors, the performances are largely stable after different identity thresholds have been applied. Even when the most rigorous threshold provided by CD-HIT-EST is applied (i.e. 80% sequence identity), the full transcript mode predictor and the mature mRNA predictor exhibit only ∼0.02 and 0.01 decrease of AUROC respectively, indicating that these predictors are robust to sequence redundancy. In terms of the AUPRC, the performances of SRAMP also remain acceptable (AUPRC = 0.449 and 0.321, respectively; Supplementary Figure S4C and D). To assess the performances more precisely, we have applied four stringency thresholds corresponding to the 99%, 95%, 90% and 85% specificities in the cross-validation tests, respectively. In line with the intuitive observation from the ROC curves and precision-recall curves, the predictors also exhibit competitive performance with different false positive rate control, as indicated by stable MCCs for most thresholds (Table 1). Moreover, we also compare the distribution of the experimentally identified m^{6}A sites and that of the predicted m^{6}A sites (Supplementary Figure S4E and F). On pre-mRNAs, the experimentally identified m^{6}A sites show a prominent enrichment near the stop codon and a weak enrichment near the start codon, similar to the previous observations (3). On mature mRNAs, the experimentally identified m^{6}A sites also exhibit strong tendency to be located near the stop codon. The predicted m^{6}A sites could largely recapitulate these biased distributions. Since no site position or topology information has been considered, these results indicate that SRAMP could recognize the specific sequence features of the m^{6}A-enriched regions and provide reasonable prediction results. Finally, during the independent tests, we also observed that the full transcript mode and the mature mRNA mode predictors tend to predict highly overlapped but not identical subsets of m^{6}A sites (Supplementary Figure S6). This observation, at least partly, supports the above-mentioned idea that both confounding sequence features among positive and negative samples, and the altered sequence context in mRNAs contributed to the performance discrepancy between two prediction modes of SRAMP.

We further assessed our predictors by external datasets. YTHDF1 and YTHDF2 are two RNA-binding proteins that selectively recognize m^{6}A-modified mRNAs, where YTHDF1 promotes protein translation and YTHDF2

Table 1. Performance of SRAMP predictors on the independent testing dataset at various stringency thresholds

Predictor	Stringency	Sensitivity	Specificity	MCC
Full transcript mode	Very high	25.7%	98.7%	0.373
	High	50.3%	93.7%	0.414
	Moderate	64.5%	88.1%	0.405
	Low	72.8%	83.0%	0.385
Mature mRNA mode	Very high	11.0%	99.1%	0.211
	High	29.6%	95.0%	0.273
	Moderate	44.0%	90.0%	0.293
	Low	54.2%	85.3%	0.294

The very high, high, moderate and low stringency thresholds correspond to the 99%, 95%, 90% and 85% specificities in 5-fold cross-validation tests, respectively.

modulates mRNA stability (11,24). If SRAMP could predict *bona fide* m^{6}A sites, it should at least partly recognize the DRACH motifs inside the YTHDF protein binding sites. The full transcript mode predictor clearly discriminates YTHDF binding sites from DRACH motifs outside the binding sites (Supplementary Figure S7; AUROC = 0.855, AUPRC = 0.485), validating its prediction capability. The mature mRNA mode can also recognize YTHDF binding site with medium accuracy (Supplementary Figure S7; AUROC = 0.720, AUPRC = 0.251). Though the performance of the mature mRNA mode predictor on the YTHDF binding site dataset is not quite satisfactory, it does not imply that the mature mRNA mode predictor cannot predict *bona fide* m^{6}A sites. To validate the accuracy of the mature mRNA mode, we benchmarked it on a golden standard dataset, from which the methylated/non-methylated status of each site was rigorously examined by the SCARLET method (42). As shown in Figure 3A, the predictor correctly identified all of the *bona fide* m^{6}A sites at the very high or high confidence thresholds. We noted that only 9, 14, 12 and 3 m^{6}A sites were predicted as m^{6}A sites above the high confidence threshold along the transcripts of *MALAT1*, *TUG1*, *TPT1* and *BSG1*, respectively, indicating that the predictor could find *bona fide* m^{6}A sites with acceptable false positive rate. Indeed, there are totally three false positive predictions at the high confidence threshold within the golden standard negative sites (Figure 3A), suggesting the false positive rate is well controlled. Nevertheless, the predictor usually predicts twice to triple more m^{6}A sites when using the low confidence threshold. Though no more false positive predictions within the golden standard dataset are produced after relaxing the threshold, there should be higher false positive rates with the relaxed thresholds. Therefore, for users who wish to predict m^{6}A sites with high reliability, only predictions above the high confidence threshold should be considered.

Recently, Chen *et al.* have proposed two yeast m^{6}A site predictors: m6Apred (20) and iRNA-Methyl (21). It is therefore interesting to interrogate the cross-species performance of SRAMP and these yeast-centric predictors. Because the yeast predictors were trained on the yeast mRNA m^{6}A dataset (16), for a fair comparison, we then only compared these predictors with the mature mRNA mode predictor of SRAMP. Moreover, we built a mammalian benchmarking dataset by filtering against the samples that cannot be processed by yeast predictors. Especially, all mammalian samples in the independent testing dataset that conform to the mammalian-specific AAC consensus motif have been removed (see 'Materials and Methods' section for details). On the filtered mammalian benchmarking dataset, SRAMP still exhibits robust performance (Figure 3B and C; AUROC = 0.784, AUPRC = 0.342). By contrast, yeast-centric predictors cannot effectively predict mammalian m^{6}A sites (AUROC = 0.649 and 0.597, AUPRC = 0.192 and 0.158 for m6Apred and iRNA-Methyl, respectively). These results confirm that the construction of a mammalian-specific predictor is necessary and crucial.

On the other hand, SRAMP did not accurately predict yeast m^{6}A sites as well. We curated a yeast benchmarking dataset from the independent dataset of m6Apred (see 'Materials and Methods' section for details). On this dataset, the m6Apred ranks the best (Figure 3D and E; AUROC = 0.684, AUPRC = 0.329). The overall performances of iRNA-Methyl and SRAMP are comparable (AUROC = 0.633 and 0.612, AUPRC = 0.253 and 0.243, respectively). Given the fact that there is no universally best predictor that accurately predict mammalian and yeast m^{6}A sites at the same time, we would like to nominate SRAMP as a mammalian-centric m^{6}A site predictor. For users who are interested in predicting yeast m^{6}A sites, the yeast-centric predictor like m6Apred should be their first choice.

At last, since the m^{6}A sites of the SRAMP's training dataset were identified from the five tissues (HEK293 cell, CD8+ T cell, A549 cell, brain and liver), it is interesting to check to what extent a predictor trained with data from one tissue recognizes the m^{6}A sites from another tissue. For either prediction mode, we have trained five tissue-specific predictors using tissue-specific m^{6}A sites and tested them on the independent datasets from other tissues. The intra- and cross-tissue independent test performances are summarized in Supplementary Figure S8. The cross-tissue prediction performance is generally acceptable, indicating that the SRAMP's generic predictors that exploit all m^{6}A data could robustly predict m^{6}A sites across different tissues. On the other hand, the intra-tissue performances are superior to cross-tissue performances for the most cases. Therefore, for better prediction of m^{6}A from a specific tissue, the tissue-specific predictors have also been made accessible at our online SRAMP server.

The SRAMP server

To facilitate the community, the SRAMP predictors have been made freely available as an online server (http://www.cuilab.cn/sramp/). The prediction webpage of SRAMP is

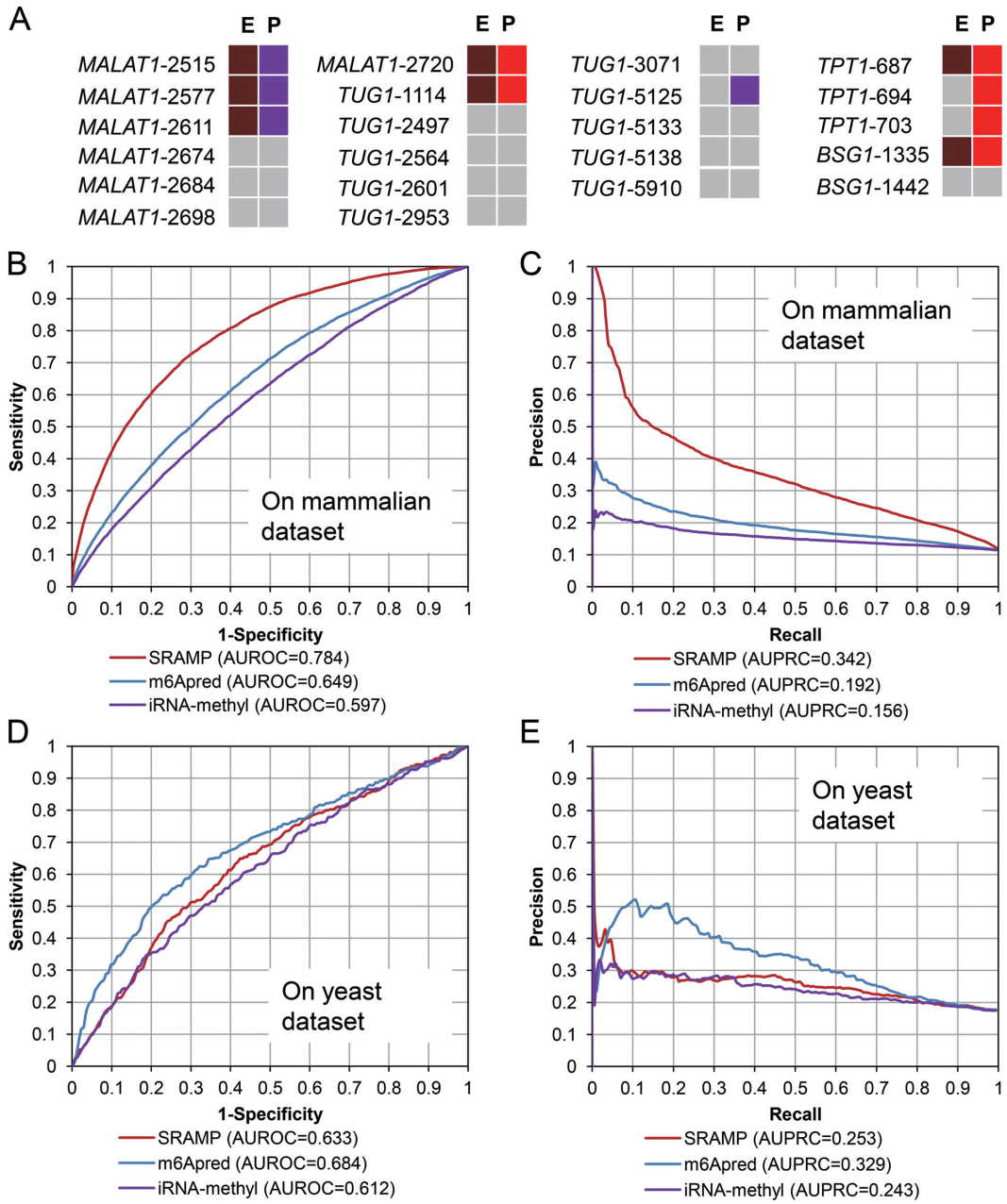
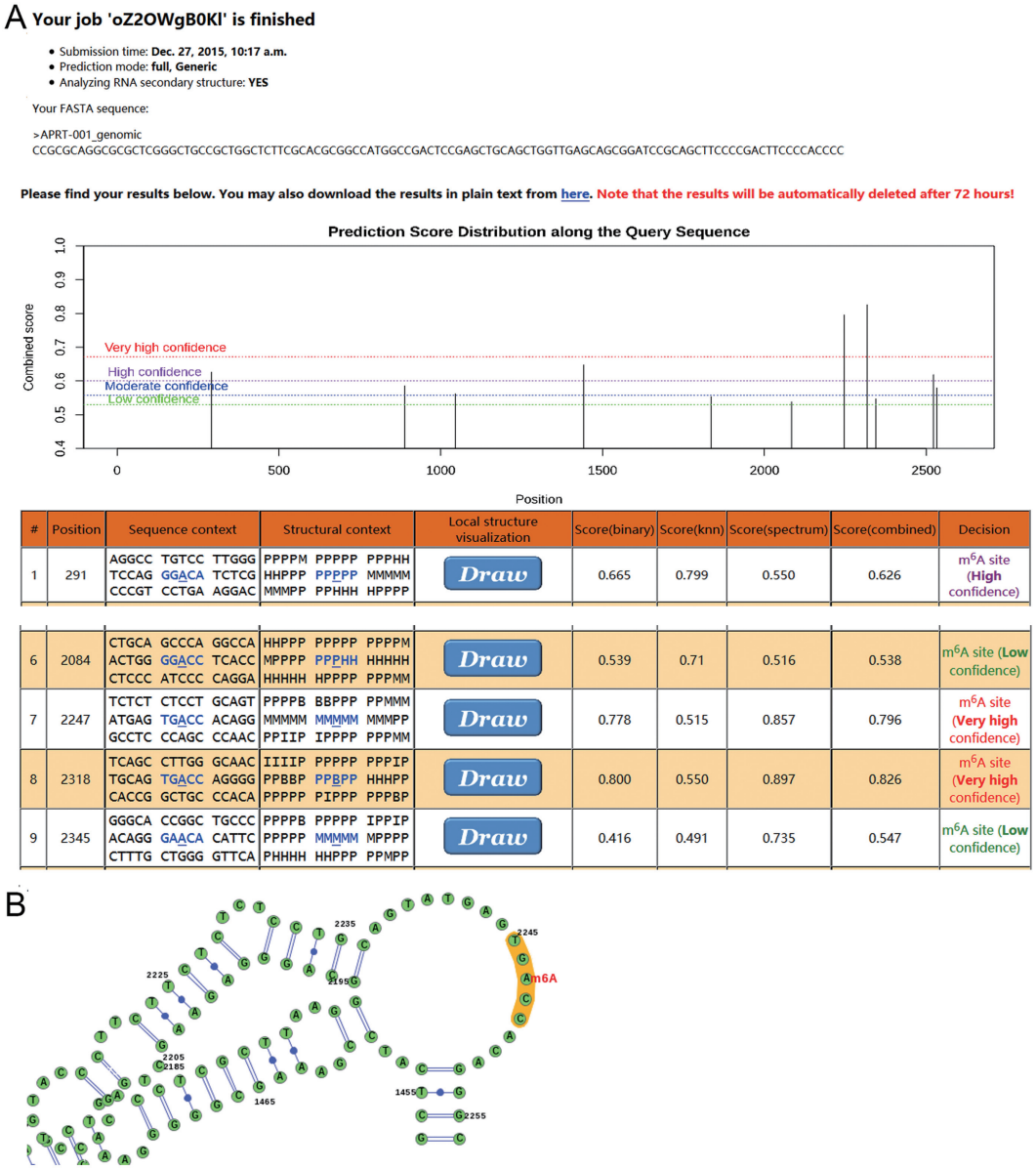


Figure 3. The performances of different m^{6}A site predictors on the gold standard dataset and the benchmarking datasets. (**A**) Prediction results on the golden standard dataset. The gene identifiers and site positions are in lines with the original publication by Liu *et al.* (42). Experimental reference sites and predicted sites are denoted in the E and P columns, respectively. Experimentally verified m^{6}A sites and non-m^{6}A sites are indicated by deep red and grey boxes, respectively. Predicted very high confidence m^{6}A sites, high confidence m^{6}A sites and non-m^{6}A sites are indicated by red, purple and grey boxes, respectively. (**B**) The ROC curve illustrating the performances on the mammalian benchmarking dataset. (**C**) The precision-recall curve illustrating the performances on the mammalian benchmarking dataset. (**D**) The ROC curve illustrating the performances on the yeast benchmarking dataset. (**E**) The precision-recall curve illustrating the performances on the yeast benchmarking dataset.

shown in Supplementary Figure S9. SRAMP only requires nucleotide sequences for running a prediction. Users can select either the full transcript mode or the mature mRNA mode, depending on if they have the genomic or the cDNA sequence at hand, and if they are interested in the intronic m^{6}A sites. Users can also decide whether the RNA secondary structure should be analysed or not. Analysis of RNA secondary structures provides text and graphical representation of the local structure around the predicted m^{6}A site (see Figure 4 as an example), but also consumes much more time. For an intuitive evaluation, SRAMP finished the prediction task on a 1000-nt RNA sequence in 90 s without analysing secondary structure, but took about 4 min when secondary structures were considered. For a quicker prediction, the 'analysing RNA secondary structure' option is by default disabled in SRAMP server. But this option can be easily enabled when submitting new prediction task.



#	Position	Sequence context	Structural context	Local structure visualization	Score(binary)	Score(knn)	Score(spectrum)	Score(combined)	Decision
1	291	AGGCC TGTCC TTGGG TCCAG GGACA TCTCG CCCGT CCTGA AGGAC	PPPPM PPPPP PPPHH HHPPP PPPPP MMMMM MMMPP PPHHH HPPPP	Draw	0.665	0.799	0.550	0.626	m^{6}A site (**High** confidence)
6	2084	CTGCA GCCCA GGCCA ACTGG GGACC TCACC CTCCC ATCCC CAGGA	HHPPP PPPPP PPPPM MPPPP PPPHH HHHHH HHHHH HPPPP PPPMM	Draw	0.539	0.71	0.516	0.538	m^{6}A site (**Low** confidence)
7	2247	TCTCT CTCCT GCAGT ATGAG TGACC ACAGG GCCTC CCAGC CCAAC	PPPPB BBPPP PPMMM MMMMM MMMMM MMMPP PPIIP IPPPP PPPMM	Draw	0.778	0.515	0.857	0.796	m^{6}A site (**Very high** confidence)
8	2318	TCAGC CTTGG GCAAC TGCAG TGACC AGGGG CACCG GCTGC CCACA	IIIIP PPPPP PPPIP PPBBP PPBPP HHHPP PPPPP PIPPP PPPBP	Draw	0.800	0.550	0.897	0.826	m^{6}A site (**Very high** confidence)
9	2345	GGGCA CCGGC TGCCC ACAGG GAACA CATTC CTTTG CTGGG GTTCA	PPPPB PPPPP IPPIP PPPPP MMMMM MPPPP PHHHH HHPPP PPMPP	Draw	0.416	0.491	0.735	0.547	m^{6}A site (**Low** confidence)

Figure 4. A sample SRAMP prediction result. The genomic sequence of a representative *APRT* transcript (ENST00000378364) was used as the input to the full transcript mode predictor, and the analysis of secondary structures was enabled. (**A**) The overview of result page. The exhibition of the query sequences is truncated, and only the detailed results for the first predicted m^{6}A site and those near the pathogenic mutation site (G2246->C) are shown. The H, M, I, B, P in the secondary structure strings mean hairpin loop, multiple loop, interior loop, bulged loop and paired residues, respectively. In addition to such string, a graphical representation of the local secondary structure will be generated when click on the 'draw' button. (**B**) A graphical representation of the local secondary structure context around the mutation site. This graphical representation was generated by SRAMP server exploiting the VARNA structural visualization tool. We focused on the local secondary structure in proximal to the mutation site for clarity.

After the query sequence is submitted to SRAMP, the user will be redirected to the 'processing' webpage which is automatically refreshed in each 30 s to check if the prediction is finished. Users can also bookmark the 'processing` webpage and check the progress later. Once the prediction task is finished, the result page will be automatically presented in the same window of the 'processing' webpage. Figure 4 provides a sample screenshot of the result webpage using the genomic sequence of *APRT* transcript (ENST00000378364) as the input. The result page consists of three sections. In the first section, basic information about the prediction task is listed. The second section contains a link to download the prediction results (in the tab-delimited text format) and a plot illustrating the distribution of the predicted m^{6}A sites along the query sequence. The third section is a table showing the detailed results about each predicted m^{6}A site in the query sequence. The index of predicted DRACH motif, position in the query sequence, the flanking sequence and the prediction scores are shown. When the 'analysing RNA secondary structure' option is enabled, a string describing the secondary structure context and a 'draw' button to generate graphical rep-

resentation of the local secondary structure are also available. SRAMP predicts totally 2, 3, 3 and 3 m^{6}A sites at the very high, high, moderate and low confidence thresholds, respectively. Here we only focused on the discussion about the prediction result of position 2247. It was reported that a mutation of its upstream G to C (rs387906584) results in APRT deficiency which makes patients vulnerable to kidney and urinary tract stone (43,44). According to the record from the ClinVar database (44), the most prominent result of this mutation is the loss of stop codon. Interestingly, however, it has been also found that this mutation significantly reduces *in vivo APRT* mRNA level, in addition to the disruptions of APRT proteins (43). Here, we note that position 2247 has been predicted as an m^{6}A site with very high confidence (Figure 4A), and mutation of its upstream G to C will eliminate the DRACH consensus motif of this potential m^{6}A site. So the pathogenic mechanism may also be related to the m^{6}A modification. Moreover, as indicated by the local secondary structure information provided by SRAMP server (Figure 4B), this site settles inside an RNA junction. Given the expected roles of m^{6}A modification in regulating RNA structures and RNA stability (9,11,12), abolishment of this m^{6}A site may destroy the local structure in 3′ UTR and/or alter the stability of *APRT* mRNA. This hypothesis would be further experimentally tested in the future.

Current limitations and future perspectives

We have presented SRAMP as a mammalian m^{6}A site prediction server. To ensure the wider applicability of this tool, only the sequence-derived features are considered and no external -omics data is required. On the one hand, sequence-based predictors show encouraging prediction performance. On the other hand, it is clear that the currently omitted -omics data could also be helpful for m^{6}A site prediction. For example, it has been demonstrated that the m^{6}A sites are enriched near the stop codon and in the longest exons, and overrepresented in the regions targeted by microRNAs and specific RNA-binding proteins (6,9,13,14). Incorporating such -omics features will be the next step to improve the predictors' performance.

Moreover, it is well known that the m^{6}A modification is reversible, dynamic and sometimes tissue- or condition-specific (3,4,19). Currently, SRAMP exploits only a few set of high-throughput data from five tissues (though of the highest resolution), and it is therefore inevitably biased toward m^{6}A sites presented under specific conditions. For example, SRAMP showed decreased performance when predicting YTHDF-binding m^{6}A sites identified from HeLa cell (11,24), a tissue currently not covered by SRAMP (Supplementary Figure S7). In addition to tissue or condition specificity, the m^{6}A sites identified by different experimental techniques would also vary from each other (16–18), and it is known that the m^{6}A antibodies used in these high-throughput experiments are somewhat biased (16). Therefore, more high-resolution m^{6}A site maps are still urgently demanded. On the one hand, such data will calibrate the m^{6}A site predictors. On the other hand, they will also be helpful to recognize the m^{6}A sites from other tissues for better interpretations of their biological functions.

Finally, according to our benchmarking tests, a yeast m^{6}A site predictor usually fails to accurately predict mammalian m^{6}A sites and vice versa (Figure 3). Nevertheless, the cross-species performance of both the m6Apred (built for yeast) and our SRAMP (built for mammalian) is clearly better than random guess (Figure 3B and D, AUROC > 0.6), indicating that there are still chances to build a universally applicable m^{6}A site predictor. Specially, we note that the best performance on the yeast benchmarking dataset is not as impressive as that on the mammalian benchmarking dataset, indicating the sequence pattern around the yeast m^{6}A site has not been fully described. This is largely due to the lack of single-nucleotide resolution map of yeast m^{6}A sites, which will enable unambiguous representation of the sequence context of yeast m^{6}A sites. It can be expected that a more powerful yeast m^{6}A site predictor and ultimately a universally applicable m^{6}A site predictor will be established when more high-resolution data becomes available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Wei Ma at Department of Biomedical Informatics, Peking University for his assistance in server configuration. We also thank Prof. You-Liang Peng's lab at China Agricultural University for kindly enabling us the temporary access to their computational resource.

FUNDING

National Basic Research Program of China [2012CB517506 to Q.C.]; National High Technology Research and Development Program of China [2014AA021102 to Q.C.]; National Natural Science Foundation of China [91339106 to Q.C., 81422006 to Q.C., 31471249 to Z.Z.]. Funding for open access charge: National High Technology Research and Development Program of China [2014AA021102].
Conflict of interest statement. None declared.

REFERENCES

1. Li,S. and Mason,C.E. (2014) The pivotal regulatory landscape of RNA modifications. *Annu. Rev. Genomics Hum. Genet.*, **15**, 127–150.
2. Machnicka,M.A., Milanowska,K., Osman Oglou,O., Purta,E., Kurkowska,M., Olchowik,A., Januszewski,W., Kalinowski,S., Dunin-Horkawicz,S., Rother,K.M. *et al.* (2013) MODOMICS: a database of RNA modification pathways–2013 update. *Nucleic Acids Res.*, **41**, D262–D267.
3. Meyer,K.D. and Jaffrey,S.R. (2014) The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.*, **15**, 313–326.
4. Fu,Y., Dominissini,D., Rechavi,G. and He,C. (2014) Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet*, **15**, 293–306.
5. Meyer,K.D., Saletore,Y., Zumbo,P., Elemento,O., Mason,C.E. and Jaffrey,S.R. (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell*, **149**, 1635–1646.
6. Dominissini,D., Moshitch-Moshkovitz,S., Schwartz,S., Salmon-Divon,M., Ungar,L., Osenberg,S., Cesarkas,K., Jacob-Hirsch,J., Amariglio,N., Kupiec,M. *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.

7. Alarcon,C.R., Lee,H., Goodarzi,H., Halberg,N. and Tavazoie,S.F. (2015) N6-methyladenosine marks primary microRNAs for processing. *Nature*, **519**, 482–485.
8. Liu,J., Yue,Y., Han,D., Wang,X., Fu,Y., Zhang,L., Jia,G., Yu,M., Lu,Z., Deng,X. *et al.* (2014) A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.*, **10**, 93–95.
9. Liu,N., Dai,Q., Zheng,G., He,C., Parisien,M. and Pan,T. (2015) N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature*, **518**, 560–564.
10. Schwartz,S., Mumbach,M.R., Jovanovic,M., Wang,T., Maciag,K., Bushkin,G.G., Mertins,P., Ter-Ovanesyan,D., Habib,N., Cacchiarelli,D. *et al.* (2014) Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5′ sites. *Cell Rep.*, **8**, 284–296.
11. Wang,X., Lu,Z., Gomez,A., Hon,G.C., Yue,Y., Han,D., Fu,Y., Parisien,M., Dai,Q., Jia,G. *et al.* (2014) N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, **505**, 117–120.
12. Roost,C., Lynch,S.R., Batista,P.J., Qu,K., Chang,H.Y. and Kool,E.T. (2015) Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. *J. Am. Chem. Soc.*, **137**, 2107–2115.
13. Chen,T., Hao,Y.J., Zhang,Y., Li,M.M., Wang,M., Han,W., Wu,Y., Lv,Y., Hao,J., Wang,L. *et al.* (2015) m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell*, **16**, 289–301.
14. Geula,S., Moshitch-Moshkovitz,S., Dominissini,D., Mansour,A.A., Kol,N., Salmon-Divon,M., Hershkovitz,V., Peer,E., Mor,N., Manor,Y.S. *et al.* (2015) Stem cells. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science*, **347**, 1002–1006.
15. Fustin,J.M., Doi,M., Yamaguchi,Y., Hida,H., Nishimura,S., Yoshida,M., Isagawa,T., Morioka,M.S., Kakeya,H., Manabe,I. *et al.* (2013) RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell*, **155**, 793–806.
16. Schwartz,S., Agarwala,S.D., Mumbach,M.R., Jovanovic,M., Mertins,P., Shishkin,A., Tabach,Y., Mikkelsen,T.S., Satija,R., Ruvkun,G. *et al.* (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, **155**, 1409–1421.
17. Linder,B., Grozhik,A.V., Olarerin-George,A.O., Meydan,C., Mason,C.E. and Jaffrey,S.R. (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods*, **12**, 767–772.
18. Chen,K., Lu,Z., Wang,X., Fu,Y., Luo,G.Z., Liu,N., Han,D., Dominissini,D., Dai,Q., Pan,T. *et al.* (2015) High-resolution N(6) -methyladenosine (m(6) A) map using photo-crosslinking-assisted m(6) A sequencing. *Angew. Chem. Int. Ed. Engl.*, **54**, 1587–1590.
19. Liu,H., Flores,M.A., Meng,J., Zhang,L., Zhao,X., Rao,M.K., Chen,Y. and Huang,Y. (2015) MeT-DB: a database of transcriptome methylation in mammalian cells. *Nucleic Acids Res.*, **43**, D197–D203.
20. Chen,W., Tran,H., Liang,Z., Lin,H. and Zhang,L. (2015) Identification and analysis of the N(6)-methyladenosine in the Saccharomyces cerevisiae transcriptome. *Sci. Rep.*, **5**, 13859.
21. Chen,W., Feng,P., Ding,H., Lin,H. and Chou,K.C. (2015) iRNA-Methyl: Identifying N-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33.
22. Ke,S., Alemu,E.A., Mertens,C., Gantman,E.C., Fak,J.J., Mele,A., Haripal,B., Zucker-Scharff,I., Moore,M.J., Park,C.Y. *et al.* (2015) A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev.*, **29**, 2037–2053.
23. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
24. Wang,X., Zhao,B.S., Roundtree,I.A., Lu,Z., Han,D., Ma,H., Weng,X., Chen,K., Shi,H. and He,C. (2015) N(6)-methyladenosine modulates messenger RNA translation efficiency. *Cell*, **161**, 1388–1399.
25. Chen,K., Kurgan,L.A. and Ruan,J. (2007) Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct. Biol.*, **7**, 25.
26. Chen,Z., Chen,Y.Z., Wang,X.F., Wang,C., Yan,R.X. and Zhang,Z. (2011) Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One*, **6**, e22930.
27. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
28. Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.
29. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
30. Friedman,J.H. and Popescu,B.E. (2008) Predictive learning via rule ensembles. *Ann. Appl. Stat.*, **2**, 916–954.
31. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
32. Li,S., Liu,B., Zeng,R., Cai,Y. and Li,Y. (2006) Predicting O-glycosylation sites in mammalian proteins by using SVMs. *Comput. Biol. Chem.*, **30**, 203–208.
33. Chen,Z., Zhou,Y., Song,J. and Zhang,Z. (2013) hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim. Biophys. Acta*, **1834**, 1461–1467.
34. Song,J., Tan,H., Shen,H., Mahmood,K., Boyd,S.E., Webb,G.I., Akutsu,T. and Whisstock,J.C. (2010) Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics*, **26**, 752–760.
35. Li,Y.H., Zhang,G. and Cui,Q. (2015) PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics*, **31**, 3362–3364.
36. Gao,J., Thelen,J.J., Dunker,A.K. and Xu,D. (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell Proteomics*, **9**, 2586–2600.
37. Chen,X., Qiu,J.D., Shi,S.P., Suo,S.B., Huang,S.Y. and Liang,R.P. (2013) Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*, **29**, 1614–1622.
38. Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
39. Sturm,M., Hackenberg,M., Langenberger,D. and Frishman,D. (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, **11**, 292.
40. Barash,D. (2003) Deleterious mutation prediction in the secondary structure of RNAs. *Nucleic Acids Res.*, **31**, 6578–6584.
41. Spitale,R.C., Flynn,R.A., Zhang,Q.C., Crisalli,P., Lee,B., Jung,J.W., Kuchelmeister,H.Y., Batista,P.J., Torre,E.A., Kool,E.T. *et al.* (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519**, 486–490.
42. Liu,N., Parisien,M., Dai,Q., Zheng,G., He,C. and Pan,T. (2013) Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA*, **19**, 1848–1856.
43. Taniguchi,A., Hakoda,M., Yamanaka,H., Terai,C., Hikiji,K., Kawaguchi,R., Konishi,N., Kashiwazaki,S. and Kamatani,N. (1998) A germline mutation abolishing the original stop codon of the human adenine phosphoribosyltransferase (APRT) gene leads to complete loss of the enzyme protein. *Hum. Genet.*, **102**, 197–202.
44. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.