*Bioinformatics*, 36(11), 2020, 3528–3536
doi: 10.1093/bioinformatics/btaa178
Advance Access Publication Date: 12 March 2020
Original Paper



Databases and ontologies

# m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine ($m^7G$) sites in human

**Bowen Song[1,†], Yujiao Tang[1,2,†], Kunqi Chen[1,3,*], Zhen Wei[1,3], Rong Rong[1,3], Zhiliang Lu[1,3], Jionglong Su[4], João Pedro de Magalhães[3], Daniel J. Rigden[2] and Jia Meng [1,3,5]**

[1]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China, [2]Institute of Integrative Biology, University of Liverpool, Liverpool L7 8TX, UK, [3]Institute of Ageing & Chronic Disease, University of Liverpool, Liverpool L7 8TX, UK, [4]Department of Mathematical Sciences and [5]AI University Research Centre (AI-URC), Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Peter Robinson

Received on December 14, 2019; revised on February 7, 2010; editorial decision on March 6, 2020; accepted on March 9, 2020

## Abstract

**Motivation:** Recent progress in N7-methylguanosine ($m^7G$) RNA methylation studies has focused on its internal (rather than capped) presence within mRNAs. Tens of thousands of internal mRNA $m^7G$ sites have been identified within mammalian transcriptomes, and a single resource to best share, annotate and analyze the massive $m^7G$ data generated recently are sorely needed.

**Results:** We report here m7GHub, a comprehensive online platform for deciphering the location, regulation and pathogenesis of internal mRNA $m^7G$. The m7GHub consists of four main components, including: the first internal mRNA $m^7G$ database containing 44 058 experimentally validated internal mRNA $m^7G$ sites, a sequence-based high-accuracy predictor, the first web server for assessing the impact of mutations on $m^7G$ status, and the first database recording 1218 disease-associated genetic mutations that may function through regulation of $m^7G$ methylation. Together, m7GHub will serve as a useful resource for research on internal mRNA $m^7G$ modification.

**Availability and implementation:** m7GHub is freely accessible online at www.xjtlu.edu.cn/biologicalsciences/m7ghub.

**Contact:** kunqi.chen@liverpool.ac.uk

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Over 150 different RNA modifications have been identified in all three kingdoms of life, playing important roles in various cellular processes (Jaffrey, 2014; Zaccara *et al.*, 2019). Among them, N7-methylguanosine ($m^7G$), the most ubiquitous RNA cap modification, is added to the 5′ cap co-transcriptionally during the initial phases of transcription and before other RNA processing events (Cowling, 2010). As a positively charged RNA modification, $m^7G$ capping plays significant roles in gene expression, protein synthesis and transcript stabilization (Furuichi *et al.*, 1977). It has been found that almost every phase of the life cycle of mRNA can be regulated by $m^7G$ cap modification, including transcription (Pei and Shuman, 2002), mRNA splicing (Konarska *et al.*, 1984), nuclear export (Lewis and Izaurflde, 1997) and translation (Muthukrishnan *et al.*, 1975). The $m^7G$ RNA modification was also found in tRNA (Guy and Phizicky, 2014) and rRNA (Sloan *et al.*, 2017), where its presence has been associated with various diseases. For example, mutations in the METTL1-WDR4 may cause a distinct form of microcephalic primordial dwarfism (Shaheen *et al.*, 2015).

Thanks to the advances of high-throughput sequencing approaches developed for transcriptome-wide mapping of internal $m^7G$ modification (Chu *et al.*, 2018; Enroth *et al.*, 2019; Malbec *et al.*, 2019; Marchand *et al.*, 2018; Zhang *et al.*, 2019a), recent studies confirmed the widespread internal existence of $m^7G$ RNA modification on mRNAs, and revealed its conservation (Malbec *et al.*, 2019), regulation and dynamics (Zhang *et al.*, 2019a) as well as its role in translation control. Zhang *et al.* (2019a) invented the

m$^7$G-MeRIP-Seq and m$^7$G-Seq techniques based on antibody immunoprecipitation and termination of reverse transcription, respectively. While m$^7$G-MeRIP-Seq (Zhang *et al.*, 2019a) provides only limited resolution (~100 bp), m$^7$G-Seq (Zhang *et al.*, 2019a) achieved base-resolution in the detection of internal mRNA m$^7$G sites by taking advantage of the misincorporation at m$^7$G sites during reverse transcription. In addition, an alternative approach m$^7$G-miCLIP-Seq (Malbec *et al.*, 2019) was also developed by combining anti-m$^7$G antibody immunoprecipitation enrichment with ultraviolet cross-linking. It provided an improved resolution (~30 bp) than the conventional MeRIP-Seq method, and its resolution may be further narrowed down to base-resolution if combined with motif analysis.

Experimental methods are usually effective but still costly and laborious. To ensure that the massive data related to internal mRNA m$^7$G methylation generated from high-throughput experiments are properly shared, annotated and taken advantage of, it is often beneficial to develop complementary bioinformatics solutions. To date, many in silico efforts have been made to support the study of the epitranscriptome or RNA epigenetics (Chen *et al.*, 2017b, 2019c). For example, the experimentally validated m$^6$A and other RNA modification sites in different species were collected in RMBase and MetDB (Liu *et al.*, 2017; Xuan *et al.*, 2018) along with various functional annotations such as the splicing sites, microRNA targets and RNA protein binding sites. The RNA modification pathways can be queried from MODOMICS (Boccaletto *et al.*, 2017). Dedicated software tools and pipelines were developed for high-throughput sequencing data profiling various RNA modification marks (Cui *et al.*, 2016; Hauenschild *et al.*, 2015; Meng *et al.*, 2013; Rieder *et al.*, 2016; Schmidt *et al.*, 2019; Zhang *et al.*, 2019c), and machine learning approaches such as iRNA-Methyl (Chen *et al.*, 2015), iRNA-m7G (Chen *et al.*, 2019b), BERMP (Huang *et al.*, 2018), SRAMP (Zhou *et al.*, 2016), DeepPromise (Chen *et al.*, 2019c), Gene2Vec (Zou *et al.*, 2018) and WHISTLE (Chen *et al.*, 2019a) were designed for accurate prediction of RNA modification sites. Enzyme-specific RNA modification site predictions were made possible for PSI (He *et al.*, 2018a) and m6A (Song *et al.*, 2019). Annotations related to RNA modifications may be obtained with RNAmod (Liu and Gregory, 2019), RCAS (Uyar *et al.*, 2017) and RNA framework (Incarnato *et al.*, 2018). Meanwhile, the disease and functional association of m6A RNA modification were revealed by m6AVar (Zheng *et al.*, 2017), m6ASNP (Jiang *et al.*, 2018), m6Acomet (Wu *et al.*, 2019), Deepm6A (Zhang *et al.*, 2019b), DRUM (Tang *et al.*, 2019) and FunDMDeep-m6A (Zhang *et al.*, 2019c) via disease-associated genetic variants or gene regulatory network and enrichment analysis. However, to the best of our knowledge, bioinformatics efforts for internal m$^7$G RNA modification are still scarce. None of the existing bioinformatics databases collected the internal mRNA m$^7$G sites, and their disease association has not been systematically inferred.

We present here m7GHub, a centralized online platform for deciphering the location, regulation and pathogenesis of internal mRNA m$^7$G RNA methylation. The m7GHub consists of the following four major components:

i. m7GDB: a database for experimentally validated internal mRNA m$^7$G sites annotated with the post-transcriptional regulations potentially affected.
ii. m7GFinder: a web server for high-accuracy prediction of putative internal mRNA m$^7$G sites from DNA sequences or human genome coordinates.
iii. m7GSNPer: a web server for assessing the epitranscriptome impact of genetic mutations on internal m$^7$G RNA methylation.
iv. m7GDiseaseDB: a database for the disease-associated genetic variants that may lead to the gain or loss of an internal m$^7$G site, with implications for disease pathogenesis involving m$^7$G RNA methylation.

Together, m7GHub serves as a useful online resource for the studies of internal mRNA m$^7$G modification.

## 2 Materials and methods

### 2.1 Internal mRNA m$^7$G sites collected in m7GDB (m$^7$G database)

We collected a total 69 159 internal m$^7$G sites reported from eight experiments in two independent studies (Malbec *et al.*, 2019; Zhang *et al.*, 2019a). The data were generated using three different techniques (m$^7$G-Seq, m$^7$G-MeRIP-Seq and m$^7$G-miCLIP-Seq). In m$^7$G-Seq, a chemical reactivity can induce misincorporation at m$^7$G sites during the process of reverse transcription, and all the known genomic mutation sites from dbSNP were excluded from the results to reveal m$^7$G modification sites at base-resolution. The same data-processing protocol was implemented as the original publication (Zhang *et al.*, 2019a) to reproduce the internal m$^7$G map in HeLa and HepG2 cell lines, respectively, at base-resolution level. For m$^7$G-MeRIP-Seq and m$^7$G-miCLIP-Seq, all the guanosines localized within the reported m$^7$G peaks or clusters (of 30-bp window) were collected. It is worth mentioning that, as m$^7$G-MeRIP-Seq and m$^7$G-miCLIP-Seq are not base-resolution approaches, the G sites extracted from the reported regions by the two techniques should still contain a large proportion of non-m$^7$G sites. In m7GDB, the reliability of the m$^7$G sites reported from these two techniques was further assessed using our customized m$^7$G site predictor m7GFinder (detailed in the following). The datasets collected in m7GDB are summarized in Table 1.

### 2.2 Internal small RNAs m$^7$G sites collected in m7GDB (m$^7$G database)

Besides the internal m$^7$G sites on mRNAs, m7GDB also collected the known internal m$^7$G sites on small RNAs (tRNA and rRNA) reported from m$^7$G-MaP-Seq (Enroth *et al.*, 2019) and MODOMICS (Boccaletto *et al.*, 2018) (see Supplementary Table S1).

**Table 1.** Data collected in m7GDB

| ID | Site no. | Cell line | Technique | Resolution (bp) | Dataset | Source |
|---|---|---|---|---|---|---|
| H1 | 6032 | Hela | m$^7$G-Seq | 1 | | |
| H2 | 3333 | HepG2 | | | | |
| H3 | 17 225 | Hela | m$^7$G-MeRIP-Seq | ~100 | GEO: GSE112276 | Zhang *et al.* (2019a) |
| H4 | 21 577 | HepG2 | | | | |
| H5 | 18 956 | HEK293T | | | | |
| H6 | 517 | Hela | m$^7$G-miCLIP-Seq | ~30 | GSA: CRA001302 | Malbec *et al.* (2019) |
| H7 | 942 | RppH-HEK293T | | | | |
| H8 | 568 | TAP-HEK293T | | | | |
| Total | | | 69 159 record (44 058 unique sites) | | | |

*Note*: m7GDB collected 44 058 unique internal mRNA m$^7$G sites reported by three different sequencing approaches under eight experiment conditions.

### 2.3 Training and testing data for m7GFinder (m$^7$G site predictor)

We developed a customized predictor, m7GFinder, for internal m$^7$G sites. The primary training and testing datasets were generated from the base-resolution m$^7$G profiling technique m$^7$G-Seq (Zhang *et al.*, 2019a). Additionally, m7GFinder is validated on two independent techniques (m$^7$G-MeRIP-Seq and m$^7$G-miCLIP-Seq). For those m$^7$G peaks identified by m$^7$G-MeRIP-Seq (Zhang *et al.*, 2019a), all the guanosines localized within the reported peak ranges were considered as positive sites in the validation. For m$^7$G-miCLIP-Seq (Malbec *et al.*, 2019), we retained only the guanosines located within both the reported 30-bp flanking windows and the claimed m$^7$G motifs AACAAG (Malbec *et al.*, 2019) for performance validation of m7GFinder. To construct the m$^7$G prediction model under the full transcript mode, the human m$^7$G base-resolution sites from m$^7$G-Seq were used as positive data, and the negative m$^7$G sites were randomly collected from unmodified G sites located on the same transcripts as sites used as positive data. As the existing data overwhelmingly relies on polyA selection, it cannot effectively capture intronic RNA fragments and may lead to an over-estimation in the prediction accuracy under the full transcript mode, a mature mRNA mode was also considered as previously described (Chen *et al.*, 2019a). Under the mature mRNA mode, the positive and negative m$^7$G sites were filtered so that only those located on mature mRNAs remained (see Supplementary Fig. S1).

It is worth noting that the full transcript and mature mRNA modes we considered here are different from those implemented in the SRAMP method (Zhou *et al.*, 2016). In the SRAMP method, the two models (full transcript and mature mRNA) describe whether predictive features should be extracted from pre-mRNA or mature mRNA; while in our method, the two models describe whether intronic sites were considered in the training and evaluation process. Due to the polyA selection step in RNA-Seq library construction, intronic signals are likely to be under-represented in experiment data, leading to an over-estimation in performance evaluation under full transcript mode when intronic sites were considered; a mature mRNA mode is thus proposed for more accurate performance evaluation.

For optimal use of the limited number of experimentally validated internal mRNA m$^7$G sites, we collected 10 negative sites for each positive one, and the negative sites were randomly split into 10 subsets to generate 10 separate predictors each with 1:1 positive-to-negative ratio. The same procedure was also applied to generate negative sites for testing data, and the prediction results of the 10 predictors were averaged. For datasets generated from base-resolution technique (H1 and H2), dataset level cross-validation was performed, in which one of the datasets was used as training purpose, while the other one was used for independent testing. Furthermore, the H3–H8 datasets generated from m$^7$G-MeRIP-Seq and m$^7$G-miCLIP-Seq were also used for performance validation.

### 2.4 Predictive features of m7GFinder

To achieve the best possible predictive performance, the m$^7$G site predictor, we constructed considered both sequence and genome-derived features as previously described in similar work (Chen *et al.*, 2017a, 2019a).

*Sequence-derived features*. The sequence-based features centered on m$^7$G and non-m$^7$G sites within the 41-bp flanking window were encoded by the chemical properties of nucleotides and nucleotide density. The chemical properties of the four types of nucleotides (A, G, C and U) were classified into three categories. The first category focused on the difference of the ring structure, in which adenosine and guanosine have two rings, while cytidine and uridine have one ring; In the second category, hydrogen bonding is considered, in which guanosine and cytidine can form one more hydrogen bond than adenosine and uridine. In the last category, distinction is made in which adenosine and cytidine contain the amino group, whereas it is the keto group in the case of guanosine and uridine. To sum up, a vector $S_i = (x_i, y_i, z_i)$ can represent the $i$-th nucleotide from the sequence:

$$x_i = \begin{cases} 1 & \text{if } s_i \in \{A,G\} \\ 0 & \text{if } s_i \in \{C,U\} \end{cases}, \quad y_i = \begin{cases} 1 & \text{if } s_i \in \{A,C\} \\ 0 & \text{if } s_i \in \{G,U\} \end{cases}, \quad z_i = \begin{cases} 1 & \text{if } s_i \in \{A,U\} \\ 0 & \text{if } s_i \in \{C,\ G\} \end{cases} \quad (1)$$

Therefore, the A, C, G, U can be encoded as a vector including three features (1,1,1), (0,1,0), (1,0,0) and (0,0,1), respectively. Additionally, the cumulative nucleotide frequency of nucleotide in the $i$-th position is calculated for the nucleotide density. To define the density of nucleotide in $i$-th position, the formula $d_i = A_i/i$ is introduced as the sum of the occurrences of the nucleotide $A_i$ before the $i+1$ position divided by its position $i$. If we use a sample sequence 'CGGAUAC' to explain the formula, the cumulative frequency for adenosine at the fourth and sixth position is calculated as 0.25 (1/4) and 0.33 (2/6), respectively; similarly, the frequency for cytidine is 1 (1/1) and 0.29 (2/7) at the first and seventh positions of the sample sequence.

*Genome-derived features*. Besides the conventional sequence-derived features mentioned above, we also integrated 42 additional genome-derived features (Supplementary Table S2) into our m$^7$G site prediction model that may contribute to the prediction accuracy, including 35 features considered previously (Chen *et al.*, 2019a) and 7 new features. Specifically, we first paid attention to the transcript regions within which the guanosine falls: this information was represented by the Genomic Features 1–16 as dummy variable features, as generated by the GenomicFeatures R/Bioconductor package using the transcript annotations hg19 TxDb package (Lawrence *et al.*, 2013). We only extracted the transcript sub-regions on the primary (longest) transcripts of each gene, helping to eliminate isoform ambiguity from our analysis. For Genomic Features 17–20, we considered the relative position of the transcript regions (3′-UTR, 5′-UTR, CDS and whole transcript) as encoded by a real value, such as the distance from the guanosine to the 3′ end divided by the width of the region. If a site does not belong to one specific region, the value is set to zero. The length of the transcript region where m$^7$G modification sites fall was represented by Genomic Features 21–25. For Genomic Features 26–27, the distance from the guanosine sites to the 5′ end or 3′ end of the splicing junctions is considered. The Phast-Cons (Siepel, 2005) score and the fitness consequence (Gulko *et al.*, 2015) scores are used to measure the conservation degree, which are shown in features 28–31 calculated for the guanosine sites and its flanking regions. The RNA secondary structures around the guanosine site are predicted using RNAfold from the Vienna RNA package (Lorenz *et al.*, 2011) and shown in features 32–33. Genomic properties of transcripts where m$^7$G sites are located were represented in features 34–38. Last but not least, the attributes of genes or transcripts were represented in features 39–42, such as microRNA targeted genes (Chou *et al.*, 2018) and HNRNPC binding sites (ENCODE Project Consortium, 2012). Please refer to Supplementary Table S2 for the details of all the 42 genomic features considered in m7GFinder.

### 2.5 Machine learning approach and performance evaluation of m7GFinder

Support vector machine (SVM) has been shown to be a quite effective machine learning algorithm in the field of computation biology and achieved good performance previously in prediction of m$^6$A RNA methylation sites (Chen *et al.*, 2017a). We applied the R language interface of LIBSVM (Chang and Lin, 2011) to construct our m$^7$G site prediction model, and the radial basis function was set as kernel following the default setting for other parameters. A 5-fold cross-validation was performed on training dataset, and the final performance of the m7GFinder was evaluated by independent testing datasets including those generated from the same or different techniques, as described previously. The prediction accuracy was represented by the ROC (receiver operating characteristic curve) (sensitivity against 1-specificity), and the area under ROC curve (AUROC) was calculated as the main performance evaluation metric. Only the m$^7$G sites not previously applied to training data were considered in the performance testing, so that the performance directly reflects the capability of our approach in discovering less prominent (or condition-specific) previously unknown m$^7$G sites, rather than well-established (or house-keeping) internal m$^7$G sites that are

robustly detected in different m$^7$G profiling experiments, as previously implemented (Chen *et al.*, 2019a).

When comparing the performances of m7GFinder and a previous developed m$^7$G site predictor iRNA-m7G, since the iRNA-m7G web server reports only the putative m$^7$G sites with scores above its cutoff level, we cannot calculate the AUROC as previously; instead, the sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthews correlation coefficient (MCC) were presented for performance evaluation, specifically:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{4}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

where TP represents the true positive, while TN represents the true negative, FP is the number of false-positive and FN the number of false-negative.

### 2.6 Output of m7GFinder

To convey whether a guanosine site is more likely to be an m$^7$G RNA methylation site or not, the likelihood ratio (LR) is calculated as the formula below, and reported in m7GFinder as its output:

$$\mathrm{LR} = \frac{P(\mathrm{observation} \mid \mathrm{m^7G})}{P(\mathrm{observation} \mid \mathrm{G})} \tag{6}$$

A larger LR value means that the site is more likely to be an m$^7$G RNA methylation site. The upper bound of a *P*-value can be inferred from the LRs of all the transcriptome G sites. It suggests how extreme is the observed LR among all the transcriptome G sites, and can be used to assess the statistical significance of a LR value. This is also reported by m7GFinder.

### 2.7 Annotation of post-transcriptional regulations affected by internal mRNA m$^7$G

To unveil the potential impact of m$^7$G in post-transcriptional regulations, we found the intersection of all the internal m$^7$G sites with RBP regions and miRNA targets. Notably, although the binding information of METTL1, a known m$^7$G methyltransferase, is not available from existing RBP database. This information has been manually added to m7GHub (Bao *et al.*, 2018). Furthermore, we obtained the regions of splicing sites within 100 bp upstream and downstream, the m$^7$G sites localized on this regions were also collected for analysis. These functional annotations are available in both m7GDB and m7GFinder (when the inputs are genome coordinates).

### 2.8 Assessment of the impact of genetic mutations on internal mRNA m$^7$G methylation status by m7GSNPer

The web server m7GSNPer was designed to evaluate the epitranscriptome impact of genetic mutations on internal m$^7$G RNA methylation status. A variant is defined as m$^7$G-associated variant if it can cause the alteration of methylation status of an internal mRNA m$^7$G site, including two scenarios: (i) a mutation directly alters G to another base, leading to the loss of an experimentally validated or computationally predicted m$^7$G site, or alters another nucleotide to G, leading to the gain of a computationally predicted m$^7$G site; (ii) a mutation alters the nucleotide within the 41-bp flanking window of an experimentally validated or computationally predicted m$^7$G site, causing significant increase or decrease in the probability of m$^7$G methylation, as is reported by our customized m$^7$G site predictor m7GFinder.

The m$^7$G sites considered in m7GSNPer were classified into three confidence levels. The high confidence level involves experimentally validated m$^7$G sites reported by base-resolution sequencing approach m$^7$G-Seq. The medium level involves m$^7$G sites identified from non-base-resolution approaches (m$^7$G-MeRIP-Seq and m$^7$G-miCLIP-Seq). In addition, m7GFinder was applied transcriptome-wide to identify all the putative m$^7$G sites, which are defined as m$^7$G sites of low confidence level. The complete dataset of 37 094 832 germline variants (dbSNP151) from dbSNP (Sherry, 2001) and 3 820 716 somatic variants (TCGA v15.0) from TCGA (Tomczak *et al.*, 2015) was collected as inputs to decipher the applications of m$^7$GSNP. Only the variants localized on exons were considered in the analysis.

### 2.9 Association analysis between m$^7$G and various diseases (m7GDiseaseDB)

The m7GDiseaseDB was developed to explore the potential association between m$^7$G sites and disease-associated genetic mutations, which might implicate possible disease pathogenesis involving m$^7$G RNA methylation. In this analysis, the disease-associated variants act as a potential bridge to link m$^7$G RNA modification to known diseases. In order to unveil the potential impact of m$^7$G modification on diseases, disease-associated SNPs (tagSNPs) were derived from different resources, including GWAS catalog (Buniello *et al.*, 2019), Johnson and O'Donnell (2009) and ClinVar (Landrum *et al.*, 2016), we then mapped all m$^7$G-associated variants to the collected tagSNPs. To annotate m$^7$G sites and m$^7$G-associated variants, the transcript structure from UCSC (Lawrence *et al.*, 2013) was used, and the evolutionary conservation of sequence was extracted from phastCons 20-way (Siepel, 2005). In addition, the deleterious level of each m$^7$G-SNPs was analyzed by SIFT (Kumar *et al.*, 2009), PolyPhen2 HVAR (Adzhubei *et al.*, 2010), PolyPhen2HDIV (Adzhubei *et al.*, 2010), LRT (Chun and Fay, 2009) and FATHMM (Shihab *et al.*, 2013) using ANNOVAR package (Wang *et al.*, 2010).

### 2.10 Website construction

MySQL tables were exploited for the storage and management of the metadata in m7GHub. Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) and Hypertext Preprocessor (PHP) were used to construct the web interface. The multiple statistical diagrams were presented by EChars, and Jbrowse genome browser (Skinner *et al.*, 2009) was employed for interactive exploration and visualization of relevant genome coordinate-based records.

## 3 Results

### 3.1 Collection of internal mRNA m$^7$G sites in m7GDB

A total of 44 058 internal m$^7$G sites were collected from data generated under eight samples profiled with three different techniques in two independent studies (see Table 1). These sites were annotated with post-transcriptional regulations such as miRNA target sites, alternative splicing sites and RNA binding protein target sites, which may be potentially regulated by internal mRNA m$^7$G methylation. The reliability of internal m$^7$G sites extracted from non-base-resolution techniques was also re-evaluated using our customized m$^7$G predictor m7GFinder. To the best of our knowledge, m7GDB is the first and only database for internal mRNA m$^7$G sites. Besides, m7GDB also collected internal m$^7$G sites detected in small RNAs, for example, tRNA and rRNA, as well as the m$^7$G sites collected in MODOMICS database (see Supplementary Table S1), making m7GDB the most comprehensive collection of internal RNA m$^7$G sites.

**Table 2.** Performance evaluation of m7GFinder (AUROC)

| Mode | Testing method | Encoding method | Base-resolution technique ($m^7G$-Seq) | | |
|---|---|---|---|---|---|
| | | | Hela | HepG2 | Average |
| Full transcript | Cross-validation | m7GFinder | 0.977 | 0.976 | 0.977 |
| | | PseKNC2 | 0.750 | 0.721 | 0.736 |
| | | EIIP | 0.786 | 0.785 | 0.786 |
| | | PSNP | 0.844 | 0.837 | 0.841 |
| | | Composition | 0.785 | 0.783 | 0.784 |
| | | MethyRNA | 0.824 | 0.772 | 0.798 |
| | | AutoCorrelation | 0.700 | 0.640 | 0.670 |
| | Independent dataset | m7GFinder | 0.973 | 0.974 | 0.974 |
| | | PseKNC2 | 0.697 | 0.705 | 0.701 |
| | | EIIP | 0.737 | 0.783 | 0.760 |
| | | PSNP | 0.808 | 0.811 | 0.810 |
| | | Composition | 0.737 | 0.781 | 0.759 |
| | | MethyRNA | 0.728 | 0.759 | 0.744 |
| | | AutoCorrelation | 0.673 | 0.639 | 0.656 |
| Mature mRNA | Cross-validation | m7GFinder | 0.903 | 0.891 | 0.897 |
| | | PseKNC2 | 0.673 | 0.647 | 0.660 |
| | | EIIP | 0.717 | 0.709 | 0.713 |
| | | PSNP | 0.788 | 0.785 | 0.787 |
| | | Composition | 0.717 | 0.708 | 0.713 |
| | | MethyRNA | 0.753 | 0.724 | 0.739 |
| | | AutoCorrelation | 0.553 | 0.528 | 0.541 |
| | Independent dataset | m7GFinder | 0.904 | 0.874 | 0.889 |
| | | PseKNC2 | 0.591 | 0.577 | 0.584 |
| | | EIIP | 0.651 | 0.614 | 0.633 |
| | | PSNP | 0.688 | 0.649 | 0.669 |
| | | Composition | 0.649 | 0.613 | 0.631 |
| | | MethyRNA | 0.730 | 0.718 | 0.724 |
| | | AutoCorrelation | 0.520 | 0.511 | 0.516 |

*Note*: About 93.9% of $m^7G$ sites reported by $m^7G$-Seq can be identified by m7GFinder under full transcript mode (sensitivity: 0.939) at the cut-off of 0.5 in prediction probability.

### 3.2 Feature selection and performance evaluation of m7GFinder

For $m^7G$ site prediction, both full transcript mode and mature mRNA mode were constructed. The $m^7G$-Seq introduced by Zhang *et al.* (2019a) performed polyA selection in the step of RNA-Seq library preparation. Therefore, the mature mRNA mode was considered to reduce the potential over-estimation of accuracy. Feature selection was implemented to identify the most important subset of genomic features, which in return avoids the over-fitting issue. We implemented the Perturb method (Gevrey *et al.*, 2003) to evaluate the relative importance of each genomic feature under the two modes. The ranking of each genome-derived feature can be found in Supplementary Figure S2. To achieve the most robust performance, we used the top 19 genomic features to construct the predictor under full transcript mode, and the top 22 for mature mRNA mode.

The performance of the newly constructed $m^7G$ site predictor (m7GFinder) was evaluated by 5-fold cross-validation, independent testing, and compared with other sequence-derived encoding methods, including PseKNC2 (Liu *et al.*, 2015a, b), EIIP (He *et al.*, 2018b; He *et al.*, 2019), PSNP (He *et al.*, 2018b; He *et al.*, 2019), Composition (Zhou *et al.*, 2016), MethyRNA (Chen *et al.*, 2017a) and AutoCorrelation (Liu *et al.*, 2015a, b) (see Table 2). When testing on independent datasets generated from two cell lines, m7GFinder achieved an average AUROC of 0.974 and 0.889 under full transcript and mature mRNA modes, respectively, which was superior to the other sequence encoding schemes as well as the newly developed iRNA-m7G method (AUROC of 0.946 under full transcript mode) (Chen *et al.*, 2019b).

It was shown previously that technological preference may significantly affect the results of epitranscriptome profiling (Adachi *et al.*, 2018; Hussain *et al.*, 2013; Zaringhalam and Papavasiliou, 2016). As m7GFinder was trained on base-resolution $m^7G$-Seq data, and the performance can be significantly over-estimated under the full transcript mode, we further validated its performance on datasets generated from two other $m^7G$ profiling techniques ($m^7G$-MeRIP-Seq and $m^7G$-miCLIP-Seq) under mature mRNA mode. Consistent with previous results, the newly developed m7GFinder approach substantially outperformed other encoding schemes (Supplementary Table S3) on datasets generated by both $m^7G$-MeRIP-Seq and $m^7G$-miCLIP-Seq techniques, with AUROC = 0.753 and 0.855, respectively. Taken together, these results suggest that m7GFinder should be a reliable tool for identifying putative internal mRNA $m^7G$ sites. It is worth noting that, although different overall patterns of internal mRNA $m^7G$ sites were reported previously in $m^7G$-miCLIP-Seq (showing enrichment of internal mRNA $m^7G$ in 5′-UTRs) and $m^7G$-Seq (showing enrichment of $m^7G$ in 3′-UTRs) (Malbec *et al.*, 2019; Zhang *et al.*, 2019a), the prediction results of $m^7G$-Seq-trained predictor agreed well with the $m^7G$-miCLIP-Seq data (AUROC = 0.855), suggesting that the positive sites captured by the two techniques share something significant in common, and that these features were successfully captured by our predictor m7GFinder. Meanwhile, as explained previously, $m^7G$-MeRIP-Seq is not a base-resolution technique, and there exists a large number of unmodified G sites under the $m^7G$ peaks called from $m^7G$-MeRIP-Seq data; nevertheless, the results of m7GFinder and $m^7G$-MeRIP-Seq data were coherent (AUROC = 0.753), suggesting consistent patterns were captured among them.

In addition, we further compared the performance of m7GFinder with iRNA-m7G, which, to our knowledge, is so far the only computational model published for internal mRNA $m^7G$ site

**Table 3.** Performance comparison of different methods tested on independent dataset generated by m$^7$G-miCLIP-Seq

| Encoding method | m$^7$G-miCLIP-Seq | | | |
|---|---|---|---|---|
| | Sn | Sp | ACC | MCC |
| m7GFinder | 0.842 | 0.710 | 0.760 | 0.536 |
| iRNA-m7G | 0.866 | 0.469 | 0.667 | 0.364 |
| PseKNC2 | 0.571 | 0.564 | 0.567 | 0.134 |
| EIIP | 0.634 | 0.620 | 0.626 | 0.253 |
| PSNP | 0.688 | 0.790 | 0.728 | 0.467 |
| Composition | 0.635 | 0.622 | 0.628 | 0.257 |
| MethyRNA | 0.684 | 0.664 | 0.673 | 0.347 |
| AutoCorrelation | 0.553 | 0.556 | 0.554 | 0.108 |

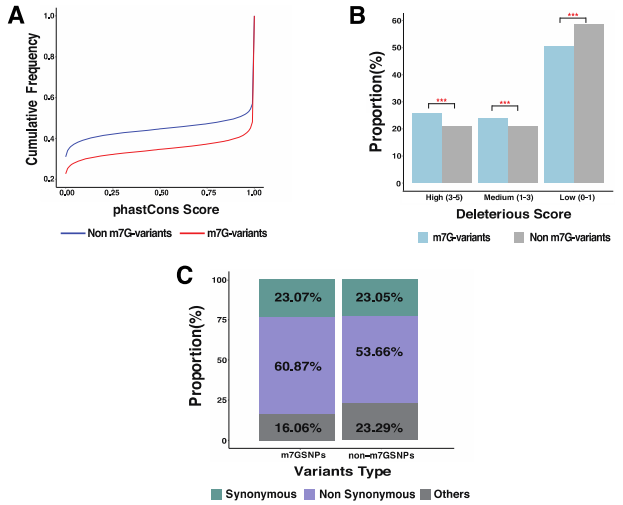


**Fig. 1.** The comparison between m$^7$G-associated variants and non-m$^7$G-associated variants. (**A**) The cumulative distribution function of differences between phastCons score of m$^7$G-associated variants and non-m$^7$G variants, m$^7$G-associated variants were more conservative than non-m$^7$G variants. (**B**) Proportional distribution of the m$^7$G-associated variants and non-m$^7$G variants at high, medium and low deleterious levels. The deleterious level was analyzed by SIFT (Kumar *et al.*, 2009), PolyPhen2 HVAR (Adzhubei *et al.*, 2010), PolyPhen2HDIV (Adzhubei *et al.*, 2010), LRT (Chun and Fay, 2009) and FATHMM (Shihab *et al.*, 2013), a high level indicated that the variant was considered deleterious in at least three out of the five above-listed methods. (**C**) Proportional distribution of the m$^7$G-associated variants and non-m$^7$G variants at different variant types, nonsynonymous type constitutes the majority of the m$^7$G-associates variants

prediction. As the AUROC cannot be calculated from the output of the iRNA-m7G web server of, we instead calculated the Sn, Sp, ACC and MCC for performance evaluation. Meanwhile, since the training data used for iRNA-m7G were generated from m$^7$G-Seq, to avoid over-fitting, we applied the dataset from another technique m$^7$G-miCLIP-Seq for an independent testing. As is shown in Table 3, our newly proposed model m7GFinder obtained the highest accuracy of 0.760, which is ~9.3% higher than that of iRNA-m7G method.

The performance evaluation of predictor on the binding regions of enzymes related to m$^6$A RNA modification has been applied in previously study (Zhou *et al.*, 2016). In our study, we also tested whether the newly purposed approach can predict the binding sites of METTL1, which is a known m$^7$G methyltransferase. The motifs inside the experimentally identified METTL1 binding regions (Bao *et al.*, 2018) were used as the positive data. While the negative sites were randomly selected outside the METTL1 binding region, keeping the 1:10 positive-to-negative ratio. Consistent with previous results, m7GFinder substantially outperformed other encoding schemes under mature mRNA model (see Supplementary Table S4), suggesting again the reliability of our method from a different perspective. Besides, the comparison between different algorithms indicated that SVM was a quite effective machine learning approach and achieved the best performance in our study.

## 3.3 Assessing the impact of mutations on internal m$^7$G methylation (m7GSNPer)

The m7GSNPer web server was developed to evaluate the impact of genetic variants on internal m$^7$G RNA methylation based on both our customized high-accuracy m$^7$G site predictor (m7GFinder) and the collection of experimentally validated internal mRNA m$^7$G sites. It critically assesses the changes (in the probability) of m$^7$G methylation induced by an arbitrary genetic mutation, unraveling the potential functional machinery of the mutation via the epitranscriptome regulation. The m7GSNPer server also systemically annotates the m$^7$G-associated variants with disease analysis and various post-transcriptional regulations as with m7GDB. To our knowledge, it is the first of its kind developed for assessing the impact of mutations on internal m$^7$G RNA modification.

## 3.4 Comparing the m$^7$G- and non-m$^7$G-associated variants (m7GDiseaseDB)

With m7GSNPer, we systematically evaluated the potential relationship between the methylation status of internal mRNA m$^7$G sites and all the known genetic variants around it. In total, we found 57 769 m$^7$G-associated SNPs, which may cause the gain or loss of an m$^7$G site in human (see Supplementary Table S5). A total of 735 and 12 800 genetic variants may cause the loss of an experimentally validated m$^7$G site at high and medium confidence level, respectively. We observed that the m$^7$G-associated SNPs were enriched in coding DNA sequence (a total of 50 970 m$^7$G-associated SNPs, 90.54%), and especially for the predicted level (40 275 SNPs, 92.75%). The distribution characteristics of m$^7$G-associated SNPs and non-m$^7$G SNPs in different transcript structures were summarized in Supplementary Table S6. We then asked that if m$^7$G-associated variants differ from those non-m$^7$G-associated variants (non-m$^7$G variants) in some biological meaning ways. PhastCons score was considered to evaluate the conservation degree between the two categories of variants. We found that the m$^7$G-associated variants were more conserved than non-m$^7$G variants (see Fig. 1A), suggesting that the sites where m$^7$G-associated variants localized may undergo stronger selection pressure than that of non-m$^7$G variants, and the genetic mutations on those more conservative sites may relate to relatively more important biological functions (e.g. the change of m$^7$G methylation). Besides, the m$^7$G-associated variants were predicted to have a higher proportion in both high-deleterious (14 589 variants, 25.56%; $P < 0.001$, $\chi^2$ test) and medium-deleterious (13 615 variants, 23.85%; $P < 0.001$, $\chi^2$ test) levels, compared with non-m$^7$G variants (Fig. 1B). Moreover, the proportion of nonsynonymous variants in m$^7$G-associated SNPs is higher than non-m$^7$G SNPs (Fig. 1C, $P < 0.001$, two-tailed population test), revealing the variants that affect m$^7$G methylation were also more likely to alter the amino acids in a protein sequence. We also observed that m$^7$G-associated SNPs occurred more frequently in binding regions of METTL1 than the non-m$^7$G SNPs with a *P*-value <0.001 (Supplementary Fig. S3). To sum up, these results suggested that the m$^7$G-associated variants can be distinguished from the majority of passenger variants, and may have important roles in human genomes.

## 3.5 Association of disease and internal mRNA m$^7$G sites (m7GDiseaseDB)

The identified m$^7$G-associated variants were also annotated with disease information and various post-transcriptional regulations (see Supplementary Table S7). For RBP binding regions, 6863 and 22 078 m$^7$G-associated variants from dbSNP and TCGA are related to 166 and 170 RBPs. For disease association analysis, 1218 m$^7$G variants localized on 716 genes were found to be associated with 681 diseases, which highlights the potential pathogenic role of the

**Table 4.** Disease types most enriched with $m^7G$ variants

| Name | No. | Database | Study accession | Clinical significance | Identifiers |
|---|---|---|---|---|---|
| Hereditary cancer-predisposing syndrome | 33 | ClinVar study | RCV000130572.2 | Uncertain significance | MedGen: C0027672 |
| Cardiovascular phenotype | 16 | ClinVar study | RCV000249377.1 | Likely benign | MedGen: CN230736 |
| Primary ciliary dyskinesia (PCD) | 16 | ClinVar study | RCV000462181.1 | Benign | OMIM: PS244400 |

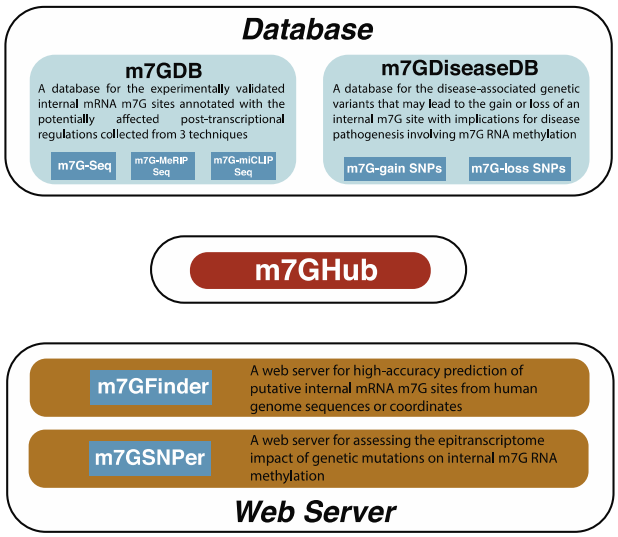


**Fig. 2.** The overall design of m7GHub. The m7GHub consists of m7GDB, m7GFinder, m7GSNPer and m7GDiseaseDB for deciphering the location, regulation and disease pathogenesis of internal mRNA $m^7G$ modification

disease-related genetic mutations via the regulation of internal $m^7G$ RNA methylation functioning at the epitranscriptome layer.

We then identified the disease phenotypes that are most enriched with $m^7G$ variants. Among them, 33 variants (2.71%) were related to hereditary cancer-predisposing syndrome (ClinVar study, Accession: RCV000130572.2), followed by 16 variants (1.31%) related to cardiovascular phenotype and 16 variants (1.31%) in primary ciliary dyskinesia (see Table 4).

In the previous disease-relevant studies, synonymous variants were often being neglected by their property of not altering the amino acids sequence of a protein. As more evidences have been found to support the effects of synonymous variants on various diseases (Sauna and Kimchi-Sarfaty, 2011), m7GSNPer and m7GDiseaseDB were designed to classify the predicted variants into synonymous and nonsynonymous groups. Take rs158921 as an example, this synonymous variant alters guanine to adenine at position 60241142 of positive strand on chromosome 15, and is related to Cockayne syndrome (ClinVar study, accession: RCV000278856.1). We also observed an $m^7G$ methylation site at this position by $m^7G$-Seq, and speculated that the dysregulation of $m^7G$ modification may relate to Cockayne syndrome. Together, the disease-relevant information provided in m7GDiseaseDB is particularly valuable for deciphering the disease mechanisms involving internal mRNA $m^7G$ methylation.

### 3.6 The m7GHub website

As a comprehensive online platform, m7GHub consists of four major components m7GDB, m7GFinder, m7GSNPer and m7GDiseaseDB, as previously described, to support studies related to internal mRNA $m^7G$ methylation in human (Fig. 2).

All the components of m7GHub can be easily accessed through the homepage of m7GHub (www.xjtlu.edu.cn/biologicalsciences/m7ghub) with simple and clear guidance (Fig. 3). In m7GDB, all the experimentally validated internal mRNA $m^7G$ sites are classified by their sequencing techniques ($m^7G$-Seq, $m^7G$-MeRIP-Seq and $m^7G$-miCLIP-Seq), together with the detailed annotation of potentially affected post-transcriptional regulations. m7GFinder accepts either FASTA format or a simply tab-delimited txt format containing genome coordinates as the input file, and returns as results a report of the identified putative $m^7G$ sites with statistical summary and a location map. For any particular genetic mutations that researchers may be interested in, the web server m7GSNPer enables the users to upload their genetic variant files for analysis. A comprehensive report containing the epitranscriptome impact of the mutations on $m^7G$ RNA methylation with disease relevance annotations will be returned together with the statistical summary and explanation of each returned results for the users to explore. For m7GDiseaseDB, the details of disease-associated genetic variants and their affected $m^7G$ sites are provided. Users can further filter the variants for information related to ClinVar or GWAS database. In addition, m7GHub provides four search modes to quickly query the databases (m7GDB and m7GDiseaseDB): by Gene, RsID, Disease and Chromosome region. The JBrowse Genome Browser is also available for exploring a genomic region of interest. Lastly, m7GHub also provides a detailed help document, and all the materials presented in database and web server can be freely downloaded.

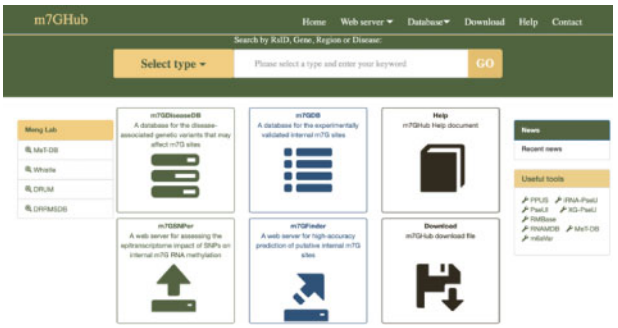


**Fig. 3.** Homepage of m7GHub. The four main components can be easily accessed from the homepage. m7GHub also provides a search bar for quick query of the database contents by Gene, RsID, Disease and Chromosome region. The m7GHub website also features with detailed help documents, and all the contents can be freely downloaded

## 4 Conclusion

With recent advances in high-throughput sequencing techniques, widespread occurrence of internal mRNA $m^7G$ modification has been revealed (Chu *et al.*, 2018; Enroth *et al.*, 2019; Malbec *et al.*, 2019; Marchand *et al.*, 2018 ). We present here, m7GHub, a comprehensive platform for deciphering the location, regulation and pathogenesis of internal mRNA $m^7G$ methylation. The platform provided the first collection of 44 058 previously reported internal mRNA $m^7G$ sites identified under different conditions (m7GDB) by different techniques; a newly developed high-accuracy predictor of internal mRNA $m^7G$ sites that outperformed existing methods (m7GFinder); the first web server for evaluating the impact of genetic mutations on the $m^7G$ methylation status (m7GSNPer) and the first database documenting the inferred 1218 associations between 681 diseases and the $m^7G$ methylation sites located on 716 genes unveiled via disease-associated genetic mutations (m7GDiseaseDB). We also provided the website with rich functional annotations, user-friendly interfaces and detailed documentation. In summary,

m7GHub (www.xjtlu.edu.cn/biologicalsciences/m7ghub) will serve as a useful resource for studies of the internal mRNA $m^7G$ modification in human.

## Funding

This work has been supported by the National Natural Science Foundation of China [31671373] and XJTLU Key Program Special Fund [KSF-T-01]. This work is partially supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund [KSF-P-02].

## Author contributions

K.C. conceived the idea and initialized the project. B.S., K.C. and J.M. designed the research. Z.W. constructed the genomic features considered in site prediction. K.C. processed the raw data and constructed the $m^7G$ site prediction model. B.S. performed analysis related to site prediction, mutation, annotation and disease association. Y.T. built the website. B.S. drafted the article. All authors read, critically revised and approved the final article.

*Conflict of Interest*: none declared.

## References

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Adachi,H. *et al*. (2018) Post-transcriptional pseudouridylation in mRNA as well as in some major types of noncoding RNAs. *Biochim. Biophys. Acta Gene Regul. Mech*., **1862**, 230–239.

Adzhubei,I.A. *et al*. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Bao,X. *et al*. (2018) Capturing the interactome of newly transcribed RNA. *Nat. Methods*, **15**, 213–220.

Boccaletto,P. *et al*. (2017) MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res*., 46, D303–D307.

Boccaletto,P. *et al*. (2018) MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res*., **46**, D303–D307.

Buniello,A. *et al*. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*., **47**, D1005–D1012.

Chang,C.-C., and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol*, **2**, 1–27.

Chen,K. *et al*. (2019a) WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res*., **47**, e41.

Chen,W. *et al*. (2015) iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem*., **490**, 26–33.

Chen,W. *et al*. (2017a) MethyRNA: a web server for identification of N6-methyladenosine sites. *J. Biomol. Struct. Dyn*., **35**, 683–687.

Chen,W. *et al*. (2019b) iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol. Ther. Nucleic Acids*, **18**, 269–274.

Chen,X. *et al*. (2017b) RNA methylation and diseases: experimental results, databases, web servers and computational models. *Brief. Bioinform*., 20, 896–917.

Chen,Z. *et al*. (2019c) Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief. Bioinform*., pii: bbz112.

Chou,C.-H. *et al*. (2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res*., **46**, D296–D302.

Chu,J.-M. *et al*. (2018) Existence of internal N 7-methylguanosine modification in mRNA determined by differential enzyme treatment coupled with mass spectrometry analysis. *ACS Chem. Biol*., **13**, 3243–3250.

Chun,S., and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res*., **19**, 1553–1561.

Cowling,V.H. (2010) Regulation of mRNA cap methylation. *Biochem. J*., **425**, 295–302.

Cui,X. *et al*. (2016) A novel algorithm for calling mRNA m 6 A peaks by modeling biological variances in MeRIP-seq data. *Bioinformatics*, **32**, i378–i385.

Enroth,C. *et al*. (2019) Detection of internal N7-methylguanosine (m7G) RNA modifications by mutational profiling sequencing. *Nucleic Acids Res*., **47**, e126.

Furuichi,Y. *et al*. (1977) 5'-Terminal structure and mRNA stability. *Nature*, **266**, 235–239.

Gevrey,M. *et al*. (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model*., **160**, 249–264.

Gulko,B. *et al*. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet*., **47**, 276–283.

Guy,M.P., and Phizicky,E.M. (2014) Two-subunit enzymes involved in eukaryotic post-transcriptional tRNA modification. *RNA Biol*., **11**, 1608–1618.

Hauenschild,R. *et al*. (2015) The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res*., **43**, 9950–9964.

He,J. *et al*. (2018a) PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics*, **19**, 306.

He,W. *et al*. (2018b) 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol*., **12(Suppl 4)**, 44.

He,W. *et al*. (2019) 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics*, **35**, 593–601.

Huang,Y. *et al*. (2018) BERMP: a cross-species classifier for predicting m(6)A sites by integrating a deep learning algorithm and a random forest approach. *Int. J. Biol. Sci*., **14**, 1669–1677.

Hussain,S. *et al*. (2013) Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol*., **14**, 215.

Incarnato,D. *et al*. (2018) RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic Acids Res*., 46, e97.

Jaffrey,S.R. (2014) An expanding universe of mRNA modifications. *Nat. Struct. Mol. Biol*., **21**, 945–946.

Jiang,S. *et al*. (2018) m6ASNP: a tool for annotating genetic variants by m6A function. *GigaScience*, **7, giy035.**

Johnson,A.D., and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet*., **10**, 6.

Konarska,M.M. *et al*. (1984) Recognition of cap structure in splicing in vitro of mRNA precursors. *Cell*, **38**, 731–736.

Kumar,P. *et al*. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc*., **4**, 1073–1081.

Landrum,M.J. *et al*. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*., **44**, D862–D868.

Lawrence,M. *et al*. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol*., **9**, e1003118.

Lewis,J.D., and Izaurflde,E. (1997) The role of the cap structure in RNA processing and nuclear export. *Eur. J. Biochem*., **247**, 461–469.

Liu,B. *et al*. (2015a) repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, **31**, 1307–1309.

Liu,B. *et al*. (2015b) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*., **43**, W65–W71.

Liu,H. *et al*. (2017) MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res*., 46, D281–D287.

Liu,Q., and Gregory,R.I. (2019) RNAmod: an integrated system for the annotation of mRNA modifications. *Nucleic Acids Res*., 47, W548–W555.

Lorenz,R. *et al*. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol*., **6**, 26.

Malbec,L. *et al*. (2019) Dynamic methylome of internal mRNA N(7)-methylguanosine and its regulatory role in translation. *Cell Res*., **29**, 927–941.

Marchand,V. *et al*. (2018) AlkAniline-Seq: profiling of m(7) G and m(3) C RNA modifications at single nucleotide resolution. *Angew. Chem. Int. Ed. Engl*., **57**, 16785–16790.

Meng,J. *et al*. (2013) Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics*, **29**, 1565–1567.

Muthukrishnan,S. *et al*. (1975) 5'-Terminal 7-methylguanosine in eukaryotic mRNA is required for translation. *Nature*, **255**, 33–37.

Pei,Y., and Shuman,S. (2002) Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5. *J. Biol. Chem*., **277**, 19639–19648.

Rieder,D. *et al*. (2016) meRanTK: methylated RNA analysis ToolKit. *Bioinformatics*, **32**, 782–785.
Sauna,Z.E., and Kimchi-Sarfaty,C. (2011) Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet*., **12**, 683–691.
Schmidt,L. *et al*. (2019) Graphical workflow system for modification calling by machine learning of reverse transcription signatures. *Front. Genet*., **10**, 876.
Shaheen,R. *et al*. (2015) Mutation in WDR4 impairs tRNA m(7)G46 methylation and causes a distinct form of microcephalic primordial dwarfism. *Genome Biol*., **16**, 210.
Sherry,S.T. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
Shihab,H.A. *et al*. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat*., **34**, 57–65.
Siepel,A. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.
Skinner,M.E. *et al*. (2009) JBrowse: a next-generation genome browser. *Genome Res*., **19**, 1630–1638.
Sloan,K.E. *et al*. (2017) Tuning the ribosome: the influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biol*., **14**, 1138–1152.
Song,Y. *et al*. (2019) Predict epitranscriptome targets and regulatory functions of N (6)-methyladenosine (m(6)A) writers and erasers. *Evol. Bioinform. Online*, **15**, 117693431987129.
Tang,Y. *et al*. (2019) DRUM: inference of disease-associated m6A RNA methylation sites from a multi-layer heterogeneous network. *Front. Genet*., **10**, 266.
Tomczak,K. *et al*. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol*., **19**, A68.
Uyar,B. *et al*. (2017) RCAS: an RNA centric annotation system for transcriptome-wide regions of interest. *Nucleic Acids Res*., **45**, e91.
Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*., **38**, e164.
Wu,X. *et al*. (2019) m6Acomet: large-scale functional prediction of individual m(6)A RNA methylation sites from an RNA co-methylation network. *BMC Bioinformatics*, **20**, 223.
Xuan,J.-J. *et al*. (2018) RMBase v2. 0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res*., **46**, D327–D334.
Zaccara,S. *et al*. (2019) Reading, writing and erasing mRNA methylation. *Nat. Rev. Mol. Cell Biol*., **20**, 608–624.
Zaringhalam,M., and Papavasiliou,F.N. (2016) Pseudouridylation meets next-generation sequencing. *Methods*, **107**, 63–72.
Zhang,L.S. *et al*. (2019a) Transcriptome-wide mapping of internal N(7)-methylguanosine methylome in mammalian mRNA. *Mol. Cell*, **74**, 1304–1316.e8.
Zhang,S.-Y. *et al*. (2019b) Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods. *PLoS Comput. Biol*., **15**, e1006663.
Zhang,S.-Y. *et al*. (2019c) FunDMDeep-m6A: identification and prioritization of functional differential m6A methylation genes. *Bioinformatics*, **35**, i90–i98.
Zheng,Y. *et al*. (2017) m6AVar: a database of functional variants involved in m6A modification. *Nucleic Acids Res*., 46, D139–D145.
Zhou,Y. *et al*. (2016) SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res*., **44**, e91.
Zou,Q. Sr *et al*. (2018) Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* **25**, 205–218.