RMDisease: a database of genetic variants that affect RNA modifications, with implications for epitranscriptome pathogenesis

Kunqi Chen, Bowen Song, Yujiao Tang, Zhen Wei, Qingru Xu, Jionglong Su, João Pedro de Magalhães, Daniel J. Rigden and Jia Meng

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

Received July 20, 2020; Revised September 08, 2020; Editorial Decision September 09, 2020; Accepted September 11, 2020

ABSTRACT

Deciphering the biological impacts of millions of single nucleotide variants remains a major challenge. Recent studies suggest that RNA modifications play versatile roles in essential biological mechanisms, and are closely related to the progression of various diseases including multiple cancers. To comprehensively unveil the association between disease-associated variants and their epitranscriptome disturbance, we built RMDisease, a database of genetic variants that can affect RNA modifications. By integrating the prediction results of 18 different RNA modification prediction tools and also 303,426 experimentally-validated RNA modification sites, RMDisease identified a total of 202,307 human SNPs that may affect (add or remove) sites of eight types of RNA modifications (m6A, m5C, m1A, m6A, m5U, m5Am, m7G and Nm). These include 4,289 disease-associated variants that may imply disease pathogenesis functioning at the epitranscriptome layer. These SNPs were further annotated with essential information such as post-transcriptional regulations (sites for miRNA binding, interaction with RNA-binding proteins and alternative splicing) revealing putative regulatory circuits. A convenient graphical user interface was constructed to support the query, exploration and download of the relevant information. RMDisease should make a useful resource for studying the epitranscriptome impact of genetic variants via multiple RNA modifications with emphasis on their potential disease relevance. RMDisease is freely accessible at: www.xjtlu.edu.cn/biologicalsciences/rmd.

INTRODUCTION

With the advances in the high-throughput sequencing technique, millions of single nucleotide polymorphisms (SNPs) have been identified from multiple species and in multiple human cancers, suggesting their critical roles in diverse biological functions and human health. However, deciphering if and how SNPs lead to functional changes is still a major challenge. Even synonymous SNPs, which do not change the amino acid sequence and so are sometimes considered ‘silent’ mutations, can still play critical roles during transcriptional and post-transcriptional regulation, such as changing splicing sites, influencing RNA-protein interactions and altering of RNA secondary structures.

Substantial efforts have been made to relate the genetic variants to their immediate biological consequences, including with regard to transcriptional regulation, post-transcriptional protein modification, RNA–protein interaction, calpain cleavage, ceRNA networks, polyadenylation and RNA modifications. Most of these works were based on a widely adopted computational framework, i.e. a machine learning model is firstly trained to capture the characteristics of a specific type of epigenetic mark (or interaction) from gold standard experimental datasets. With the model, it is then possible to assess the probability of the mark being associated with a given (DNA, RNA or protein) sequence, and further, predicting whether a genetic mutation can affect the status of epigenetic mark by comparing the probabilities of the mark being associated with the original and the mutated sequence.
quences. Importantly, this analysis not only explains how a genetic variant regulates an epigenetic mark, but also helps explain the phenotypes associated with the SNP, as those identified from GWAS analysis. For example, if a SNP is known to increase the risk of a disease, and can also destroy a transcription factor binding site, it is often natural to speculate a transcription-related disease mechanism, even though the two may not be directly causal, and additional experimental validation is necessary.

The epitranscriptome has emerged as an important layer for gene expression regulation (19,20). Recent studies suggest that various RNA modifications occur widely in the transcriptome, play versatile roles in essential biological mechanisms, and are closely related to the progression of various diseases (21–24). For example, N6-methyladenosine (m6A) controls the speed of the circadian clock (25), affects RNA stability (26) and regulates the progression of multiple cancers (27–30). N4-acetylcytidine (ac4C) promotes translation efficiency (31), and 2′-O-methylation (Nm) of HIV transcripts help the virus to avoid innate immune sensing (32).

A number of high-throughput approaches have been developed for profiling the transcriptome-wide distribution of different types of RNA modifications, including m5A-seq (or MeRIP-seq) (33,34), PA-m5A-seq (35), miCLIP (36) and m5A-CLIP (37). These approaches have generated a large number of high-quality epitranscriptome datasets, on the basis of which the properties of the modification-carrying RNA sequences can be characterized. This enables the prediction of RNA modification sites from the primary sequences (38,39), and hence allows prediction of the impact of genetic variants on RNA modification by comparing the potential for modification of the original and the mutated sequences. iRNA-methyl (40) and SRAMP (41) are two of the earliest and most widely adopted approaches for predicting RNA methylation sites from the primary RNA sequences. We previously developed a high-accuracy predictor WHISTLE (42) for prediction of RNA methylation sites by taking advantage of conventional sequence features as well as 35 additional genomic features (43).

There exist several databases of RNA modifications with different focuses. The MODOMICS database concerns mainly RNA modification pathways. MeT-DB (44), CVm6A (45) and REPIC (46) collect and annotate the transcriptome m6A sites under different experimental conditions; while RMBase is currently the most comprehensive database containing well annotated sites of multiple types of RNA modifications identified in multiple species. m6AVar (17) and m7GDiseaseDB (18) contain disease-associated SNPs that can affect m6A and internal mG RNA modification, respectively. Both were based on their respective customized RNA modification site prediction tools (m6AFinder and m7GFinder). These efforts together greatly facilitated research into RNA modifications. However, to the best of our knowledge, the impact of genetic mutation on most transcriptome modifications such as m3C, Ψ and Nm, has not been studied, and a centralized platform is not yet available for systematically deciphering the association of genetic variants with multiple RNA modification types as predicted by multiple independent tools.

To address this, we present here RMDisease, a database of genetic variants that affect RNA modifications with a focus on their potential disease association. By integrating 303,426 RNA modification sites, 40,915,548 somatic and germline SNPs and 18 prediction tools, RMDisease represents the most comprehensive available mapping from genome variants to their epitranscriptome disturbance. RMDisease contains a total of 202,307 human SNPs that can affect (add or remove) eight types of RNA modifications (m6A, m5C, m1A, m3U, Ψ, m6Am, mG and Nm), including 4,289 disease-associated variants that may imply disease pathogenesis functioning at the epitranscriptome layer. Additionally, the RNA modification-affected SNPs were further annotated with putative post-transcriptional machinery including RNA-binding protein (RBP) binding sites, miRNA targets and splicing sites. A graphical user interface was constructed to support the query, exploration and download of the database. The overall design of RMDisease is summarized in Figure 1.

**MATERIALS AND METHODS**

**Data resource**

We considered in this study only the RNA modifications that widely occur in the transcriptome. Since there is not yet available a relatively complete high-confidence collection of such data, we manually collected from 32 studies the sites of eight types of RNA modifications, including m6A (178 049 sites), m5C (95 391 sites), m1A (16 346 sites), m3U (3696 sites), Ψ (3137 sites), m6Am (2447 sites), mG (2525 sites) and Nm (1835 sites), respectively. These sites were reported by 68 high-throughput sequencing experiments generated by 18 base-resolution technologies, including m6A-REF-seq (47), MAZTER-seq (48), miCLIP (36), m6A-CLIP-seq (37), PA-m6A-seq (35), Ψ-seq (49), Pseudo-seq (50), CeU-Seq (51), RBSeq (52), m1A-MAP (53), m1A-seq (54), Aza-IP (55), RNA-BisSeq (56), FICSeq (57), Nm-seq (58), m7G-seq (59), m7G-miCLIP-Seq (60). Detailed information regarding these sequencing samples is provided in Supplementary Table S1.

We obtained 3 820 716 somatic variants and 37 094 832 germline variants from dbSNP (v151) and TCGA (v15.0), respectively (see Supplementary Table S3). Only the variants located within the mature transcripts were kept for further analysis.

**Derivation of RNA modification-associated variants**

The RNA modification-associated SNPs (RM-SNPs) are defined as SNPs that may lead to the gain or loss of an RNA modification site, as reported by the prediction tools via comparing the modification status of the original and the mutated sequences. RM-SNPs were further classified into 3 sub-groups based on their reliability, including: (i) high: a SNP directly alters the experimentally validated RNA modification site, leading to its loss; (ii) medium: a SNP alters a nucleotide within the 41 bp flanking window of an experimentally validated RNA modification site (but not directly the modifiable nucleotide itself), causing its loss as predicted by a machine learning model; (iii) Low: a SNP alters a nucleotide within the 41 bp flanking window of an...
RNA modification site (may or may not directly the modifiable nucleotide itself), causing significant increase or decrease in the probability of RNA modification. It may be worth noting that the identification of RM-SNPs of high or medium confidence level both require experimentally validated RNA modification sites in the very beginning.

We integrated 18 prediction tools developed for eight RNA modifications to perform comprehensive and independent evaluations, including five deep learning-based methods: DeepPromise (38), DeepM6Aseq (61), DeepMRMD (62), iPseU-CNN (64) and 13 machine learning-based methods: WHISTLE(65), SRAMP(66), iRNA-3type (67), iMRM (68), iRNA-2OM (69), RAMPred (71), iRNA-PseColl (72), iRNA-m7G (73), m7GFinder (74), iRNA-PseU (75), PIANO (76), ISGm1A (77).

The association level (AL) between SNP and an RNA modification is calculated as follows:

$$\text{AL} = \begin{cases} 2P_{SNP} - 2\max(0.5, P_{WT}) & \text{for gain} \\ 2P_{WT} - 2\max(0.5, P_{SNP}) & \text{for loss} \end{cases}$$

(1)

where, $P_{WT}$ and $P_{SNP}$ represent the probability of modification for the wild type and mutated sequences, respectively, as obtained from individual prediction tool (or experiment data if available). The association level (AL) ranges from 0 to 1, with 1 indicating the greatest impact on RNA modification. The statistical significance was assessed by comparing to the ALs of all mutations, with which the upper bound of the $P$-value can be calculated. Different prediction tools were employed for the same modification to obtain the association level for their corresponding target modification. As shown in previous studies, the RNA modification prediction tools that integrate both the sequence and genome-derived features outperform those based on sequence only (38,65,74,76–77), and were thus used as the primary method. We retained the RM-SNPs with association level $>0.4$ and $P$-value $<0.05$ as predicted by approaches that took advantage of both sequence and genome-derived features. The results obtained from other methods (based on sequence only) were provided in RMDisease as well for reference purpose (see Supplementary Table S2).

**Additional annotation**

To annotate the basic genomic information of RNA modification-associated variants, the transcript structure from UCSC genome browser including CDS, 3'UTR, 5'UTR, start codon and stop codon, etc. were used, and the genomic conservation were annotated by the phastCons 60-way. In addition, the deleterious level of each RNA modification-associated variant was analyzed by SIFT (78), PolyPhen2 HVAR (79), PolyPhen2HDIV (79), LRT (80) and FATHMM (81) using the ANNOVAR package (82). Furthermore, to provide annotation of putative post-transcriptional regulatory machinery, information on RBP binding sites from POSTAR2 (83), miRNA-RNA interaction from miranda (84) and startBase2 (85), and splicing sites from UCSC annotation with GT-AG role within 100 bp upstream and downstream of RNA modification-associated variants was integrated into the database. To unveil potentially epitranscriptome-related pathogenesis, the association between disease and RNA modification-associated variants was extracted from GWAS catalog (86), ClinVar (87) and Johnson and O’Donnel’s data (88). Additionally, the RNA molecule–drug sensitivity associations from RNAactDrug (89) were obtained to provide a drug suggestion for each RNA modification-associated SNP.

**Database and web interface implementation**

MySQL tables were used for the storage and management of the metadata in RMDisease. Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS) and Hyper-
RESULTS

Database content

We firstly evaluated the potentials of genetic variants to add or remove an RNA modification site directly or indirectly. In the end, a total of 57,622, 23,463, 61,563, 23,875, 24,822, 640, 5047 and 5275 genetic variants were found to be associated with m^6^A, m^1^A, m^C^C, ψ, m^G^G, Nm, m^U^U and m^6^Am, respectively, providing so far the most comprehensive map of genetic factors of epitranscriptome disturbance (Table 1).

We then obtained disease annotations of SNPs from GWAS catalog, Johnson and O’Donnel’s data, and ClinVar, and mapped them to RM-SNPs. These SNPs may link disease pathogenesis and clinical relevance to epitranscriptome regulations (Table 2). We summarized in Table 3 the diseases that are associated with the most RM-SNPs of a specific RNA modification type.

We then asked whether the RM-associated SNPs are more functional relevant to important biological events compared to non-associated SNPs, and used evolutionary conservation as an indicator. For this purpose, the phastCons 100-way conservation scores from UCSC was considered to evaluate the conservation of individual site, which was calculated for human genome derived from genomewide multiple alignments with 99 other vertebrate species. Interestingly, we found that the RM-associated variants were more conserved than non-associated variants (Figure 2), suggesting that the RM-associated SNPs underwent stronger selection pressure than the other variants, and may be related to important biological events that can be regulated at the epitranscriptome layer.

Website interface and usage

The user-friendly web interfaces provided in RMDisease enable the search, browser and download RNA modification associated-SNPs by modification type, gene, disease, chromosome region, RsID, and post-transcriptional regulations. A genome browser was integrated for interactive exploration of genome regions of interest. All data provided in the RMDisease database can be freely downloaded. For the convenience of users, detailed instructions on how to use RMDisease were placed in the ‘help’ page. RMDisease is freely accessible at: www.xjtlu.edu.cn/biologicalsciences/rmd.

CONCLUSIONS

An increasing number of biological mechanisms and disease mechanisms have been associated with the epitranscriptome, which consists of more than 100 different types of RNA modifications (91) at tens of thousands of loci in the human transcriptome. To systematically unveil the linkage between genetic factors and their respective epitranscriptome disturbance, we developed RMDisease, a database of genetic variants with potentials to alter eight types of widely spread transcriptome modifications, with emphasis on epitranscriptome disease pathogenesis. RMDisease revealed for the first time the impacts of genetic variants on six types of RNA modifications (m^6^A, m^1^A, m^U^U, ψ, m^Am^Am and Nm), and offered substantial improvements over existing works for m^6^A and m^G^G (17, 18).

RMDisease and m7GDiseaseDB (92) used the same inference method. Both databases were based on m7GFinder (74), which integrates the sequence as well as the genomic features. The main different between them is that, a different scoring system was implemented in RMDisease (see Materials and Methods section), which penalizes direct mutation of a putative m^6^A site and makes the association level (AL) fall within the range of 0 to 1. Additionally, RMDisease integrated the results obtained from another m^G^G predictor iRNA-m7G (93), and use it as an independent reference. There exists major difference between m6AVar (17) and RMDisease for m^6^A-associated SNPs. Besides the aforementioned differences between m7GDiseaseDB and RMDisease, i.e. a different scoring framework and extra independent tools integrated, RMDisease also provides the statistical significance of the predicted associations, and was based on more accurate m^6^A predictor WHISTLE (42) and with more reliable epitranscriptome datasets integrated. A total of 12,325 m^6^A sites that can be affected by SNPs were found to be shared between RMDisease and m6AVar (see Supplementary Table S4 for more details).

Previous studies have shown that there exist snoRNPs that can guide the formation of Nm and psi with the base-pairing mechanism (94–98). To the best of our knowledge, none of the existing prediction approaches for RNA modification explicitly considered this mechanism, which may undermine their prediction capability. Indeed, sequence-based predictors for pseudouridine sites without considering the base-pairing mechanism between target RNAs and snoRNPs yielded very limited accuracy (lower than 80%).
Table 1. RM-SNPs collected in RMDisease

| Modification type | Confidence level | Loss | Gain | All | Loss | Gain | All | Loss | Gain | All |
|------------------|------------------|------|------|-----|------|------|-----|------|------|-----|
| m^6A              | High             | 1405 | -    | 1405| 4276 | -    | 4276| 5681 | -    | 5681|
|                  | Medium           | 13118| 13118| -   | 33666| 33666| -   | 46784| 46784| -   |
|                  | Low              | 692  | 654  | 38  | 161  | 4304 | 4465| 199  | 4958 | 5157|
| m^1A              | High             | 104  | -    | 104 | 61   | -    | 61  | 165  | -    | 165 |
|                  | Medium           | 856  | 856  | -   | 3134 | 3134 | -   | 3990 | -    | 3990|
|                  | Low              | 3720 | 2654 | 1066| 5730 | 9858 | 15588| 6796 | 12512| 19308|
| m^5C              | High             | 596  | -    | 596 | 1518 | -    | 1518| 2114 | -    | 2114|
|                  | Medium           | 11350| 11350| -   | 28682| 28682| -   | 40032| -    | 40032|
|                  | Low              | 1338 | 3862 | 5200| 6448 | 7769 | 14217| 7786 | 11631| 19417|
| ψ                | High             | 3    | -    | 3   | 19   | -    | 19  | 22   | -    | 22  |
|                  | Medium           | 412  | -    | 412 | 1271 | -    | 1271| 1683 | -    | 1683|
|                  | Low              | 1133 | 2726 | 3859| 4324 | 13987| 18311| 5457 | 16713| 22170|
| m^7G              | High             | 10   | -    | 10  | 82   | -    | 82  | 92   | -    | 92  |
|                  | Medium           | 253  | -    | 253 | 922  | -    | 922 | 1175 | -    | 1175|
|                  | Low              | 1511 | 745  | 2256| 9685 | 11614| 21299| 11196| 12359| 23555|
| Nm               | High             | 4    | -    | 4   | 21   | -    | 21  | 25   | -    | 25  |
|                  | Medium           | 85   | -    | 85  | 530  | -    | 530 | 615  | -    | 615 |
|                  | Low              | 0    | 0    | 0   | 0    | 0    | 0   | 0    | 0    | 0   |
| m^3U              | High             | 12   | -    | 12  | 0    | -    | 0   | 12   | -    | 12  |
|                  | Medium           | 14   | -    | 14  | 38   | -    | 38  | 52   | -    | 52  |
|                  | Low              | 350  | 703  | 1053| 433  | 3497 | 3930 | 783  | 4200  | 4983 |
| m^6Am             | High             | 12   | -    | 12  | 2    | -    | 2   | 14   | -    | 14  |
|                  | Medium           | 24   | -    | 24  | 132  | -    | 132 | 156  | -    | 156 |
|                  | Low              | 220  | 1598 | 1818| 82   | 3205 | 3287 | 302  | 4803  | 5105 |

Note: RM-SNPs are further classified into two categories: (i) Direct: a SNP directly alters the modifiable nucleotide, leading to the loss of a known or predicted RNA modification site, or alters a non-modifiable nucleotide into one that can be modified. (ii) Indirect (within 41 bp): a SNP alters a nucleotide within the 41 bp flanking window of an RNA modification site (but not directly the modifiable nucleotide itself), causing significant increase or decrease in the probability of RNA modification. We considered only SNPs within the 41bp window for possible indirect effects. This is because most existing RNA modification prediction methods chose to be based on 41bp sequence or less (67–71,73,108). Increasing the length considered here may not help improve the completeness of the results but add substantially the computation load. Additionally, the Nm-SNPs of low confidence were not predicted due to the tremendous search space (Nm can happen to all nucleotide) and its relatively low abundance in the human transcriptome.

Table 2. Disease-associated RM-SNPs collected in RMDisease

| Modification type | SNP source | Total RM-SNP | SNP | Disease | Gene | SNP | Disease | Gene |
|------------------|------------|--------------|-----|---------|------|-----|---------|------|
| m^6A             | dbSNP151   | 15215        | 989 | 400     | 453  | 148 | 77      | 117 |
|                  | TCGA       | 42407        | 332 | 208     | 247  | 29  | 27      | 29  |
| m^1A             | dbSNP151   | 4680         | 217 | 175     | 139  | 0   | 0       | 0   |
|                  | TCGA       | 18783        | 326 | 208     | 247  | 29  | 27      | 29  |
| m^5C             | dbSNP151   | 17146        | 217 | 175     | 139  | 0   | 0       | 0   |
|                  | TCGA       | 44417        | 326 | 208     | 247  | 29  | 27      | 29  |
| ψ                | dbSNP151   | 4274         | 51   | 63      | 41   | 0   | 0       | 0   |
|                  | TCGA       | 19601        | 51   | 63      | 41   | 0   | 0       | 0   |
| m^7G             | dbSNP151   | 2519         | 183 | 141     | 166  | 25  | 22      | 25  |
|                  | TCGA       | 22303        | 41   | 63      | 33   | 0   | 0       | 0   |
| Nm               | dbSNP151   | 89           | 5    | 5       | 5    | 0   | 0       | 0   |
|                  | TCGA       | 551          | 2    | 18      | 2    | 0   | 0       | 0   |
| m^3U             | dbSNP151   | 1079         | 40   | 40      | 39   | 8   | 8       | 8   |
|                  | TCGA       | 3968         | 40   | 40      | 39   | 8   | 8       | 8   |
| m^6Am            | dbSNP151   | 1854         | 83   | 75      | 77   | 4   | 4       | 4   |
|                  | TCGA       | 3421         | 31   | 16      | 27   | 0   | 0       | 0   |

(61,99), and there exists speculation that the sequence features of pseudouridine sites may not exist at all (100). Incorporating base-pairing information into the prediction models are likely to further improve the prediction performance. Nevertheless, the primary approaches implemented in RMDisease are based on both sequence and genomic features. For pseudouridylation, the PIANO method (101), which was used in RMDisease as the primary prediction approach, has achieved substantially better performance than those based on sequence features only (92,101). One possible explanation is that, since many biological features are correlated, the base-pairing information may be indirectly and inexplicitly captured by the model after including additional genomic features, such as, secondary structure of RNA and genomic conservation. Additionally, our previous studies showed that, including additional ge-
Table 3. Diseases associated with the most RM-SNPs

| Disease name                              | ClinVar study accession | MedGen identifier | Type | #SNP |
|-------------------------------------------|-------------------------|------------------|------|------|
| Hereditary cancer-predisposing syndrome   | RCV000129430.4          | C0027672         | m^6^A| 58   |
| Hereditary cancer-predisposing syndrome   | RCV000129430.4          | C0027672         | m^1^A| 14   |
| Hereditary pancreatitis (PCTT)            | RCV000468581.1          | C0238339         | m^5^C| 88   |
| Hereditary cancer-predisposing syndrome   | RCV000129430.4          | C0027672         | Psi  | 12   |
| Cardiovascular phenotype                  | RCV000249770.1          | CN230736         | m^3^G| 7    |
| Adenocarcinoma of lung                    | RCV000439229.1          | C0152013         | Nm   | 1    |
| Adenocarcinoma of lung                    | RCV000439229.1          | C0152013         | m^5^U| 8    |
| Leigh syndrome (LS)                       | RCV000268982.1          | C0023264         | m^6^Am| 4   |

Figure 2. Comparing the PhastCons scores of RNA modification-associated and non-associated SNPs. The sites where RNA modification-associated variants localized were more conserved than non-associated variants for all the eight transcriptome modifications considered in RMDisease.

It is also worth noting that substantial discrepancy has been observed among different epitranscriptome profiling approaches, which can capture different bias (104–107). To minimize the impact of technology usage, special efforts have been made in this study to obtain the most comprehensive collection of the RNA modification sites including those generated from different technologies (Supplementary Table S1). Multiple machine learning models were trained with these datasets for each RNA modification to produce the most reliable results out of the data that is available.

In summary, RMDisease will serve as a useful resource for studies of genetic factors concerning the epitranscriptome regulatory circuits and their potential roles in pathogenesis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author’s contribution: Z.W. conceived the idea; K.C. and B.S. collected and processed the data; Y.T. built the website; K.C. and B.S. drafted the manuscript. All authors read, critically revised and approved the final manuscript.
FUNDING

National Natural Science Foundation of China [31671373]; XJTLU Key Program Special Fund [KSF-E-51]; AI University Research Centre through XJTLU Key Programme Special Fund [KSF-P-02]. Funding for open access charge: National Natural Science Foundation of China [31671373]; XJTLU Key Program Special Fund [KSF-E-51]; AI University Research Centre through XJTLU Key Programme Special Fund [KSF-P-02].

Conflict of interest statement. D.J.R. is Executive Editor of NAR.

REFERENCES

1. Sauna, Z.E. and Kimchi-Sarfaty, C. (2011) Understanding the contribution of synonymous mutations to human disease. Nat. Rev. Genet., 12, 683–691.

2. Wu, X. and Hurst, L.D. (2016) Determinants of the usage of splice-associated motifs predict the distribution of human pathogenic SNPs. Mol. Biol. Evol., 33, 518–529.

3. Mao, F., Xiao, L., Li, X., Liang, J., Teng, H., Cai, W. and Sun, Z.S. (2018) In silico detection of sequence variations modifying transcriptional regulation. PloS Comput. Biol., 4, e5.

4. Ryu, G.M., Song, P., Kim, K.W., Oh, K.S., Park, K.J. and Kim, J.H. (2009) Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. Nucleic Acids Res., 37, 1297–1307.

5. Ren, J., Jiang, C., Gao, X., Liu, Z., Yuan, Z., Jin, C., Wen, L., Zhang, Z., Xue, Y. and Yao, X. (2010) PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. Mol. Cell. Proteomics, 9, 623–634.

6. Kim, Y., Kang, C., Min, B. and Yi, G.S. (2015) Detection and analysis of disease-associated single nucleotide polymorphism influencing post-translational modification. BMC Med Genomics, 8, 57.

7. Reimand, J. and Badar, G.D. (2015) MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. Nat. Methods, 12, 531–533.

8. Xu, H.-D., Shi, S.-P., Chen, X. and Qiu, J.-D. (2015) Systematic effect of genetic variants on alternative polyadenylation in human cancer. Nucleic Acids Res., 43, D139–D145.

9. Song, B., Tang, Y., Chen, K., Wei, Z., Rong, R., Lu, Z., Su, J., de Magalhaes, P.J., Rigden, D.J. and Meng, J. (2020) m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human. Bioinformatics, 36, 3528–3536.

10. He, C. (2010) Grand challenge commentary: RNA epigenetics. Nat. Chem. Biol., 6, 863–865.

11. Saletore, Y., Meyer, K., Korphal, J., Vilfan, I.D., Jaffrey, S. and Mason, C.E. (2012) The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol., 13, 175.

12. McCown, P.J., Ruszkowska, A., Kunkler, C.N., Breger, K., Hulewicz, J.P., Wang, M.C., Springer, N.A. and Brown, J.A. (2020) Naturally occurring modified ribonucleosides. WIREs RNA, n/a, e1595.

13. Jones, J.D., Monroe, J. and Koutmou, K.S. (2020) A molecular-level perspective on the frequency, distribution, and consequences of messenger RNA modifications. WIREs RNA, n/a, e1586.

14. Esteve-Puig, R., Bueno-Casta, A. and Esteller, M. (2020) Writers, readers and erasers of RNA modifications in cancer. Cancer Lett., 485, 127–137.

15. Zhang, Z., Luo, K., Zou, Z., Qiu, M., Tian, J., Sieh, L., Shi, H., Zou, Y., Wang, G., Morrison, J. and et al. (2020) Genetic analyses support the contribution of mRNA N6-methyladenosine (m6A) modification to human disease heritability. Nat. Genet., 52, 939–949.

16. Fustin, J.M., Moi, D., Yamaguchi, Y., Hida, N., Nishimura, S., Yoshida, M., Isagawa, T., Morioka, M.S., Kakeya, H., Manabe, L. and et al. (2013) RNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell, 155, 793–806.

17. Wang, X., Lu, Z., Gomez, A., Hong, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G. and et al. (2014) N6-methyladenosine-dependent regulation of messenger RNA stability. Nature, 505, 117–120.

18. Yang, S., Wei, J., Cui, Y.H., Park, G., Shah, P., Deng, Y., Aplin, A.E., Lu, Z., Hwang, S., He, C. and et al. (2019) m(6)A RNA demethylase FTO regulates melanoma tumorigenicity and response to anti-PD-1 blockade. Nat. Commun., 10, 2782.

19. Niu, Y., Lin, Z., Wan, A., Chen, H., Liang, H., Sun, L., Wang, Y., Li, X., Xiong, X.F., Wei, B. and et al. (2019) RNA N6-methyladenosine demethylase FTO promotes breast tumor progression through inhibiting BNP3. Mol. Cancer, 18, 46.

20. Lin, X., Chai, G., Wu, Y., Li, J., Chen, F., Liu, J., Luo, G., Taylor, J., Du, J., Lin, S. and et al. (2019) RNA m(6)A methylation regulates the epithelial mesenchymal transition of cancer cells and translation of Snail. Nat. Commun., 10, 2065.

21. Lee, H., Bao, S., Qian, Y., Geula, S., Leslie, J., Zhang, C., Hannah, J.H. and Ding, L. (2019) Stage-specific requirement for Mettl3-dependent m(6)A mRNA methylation during haematopoietic stem cell differentiation. Nat. Cell Biol., 21, 700–709.

22. Arango, D., Sturgill, D., Alhusaini, D., Dillman, A.A., Sweet, T.J., Hanson, G., Hosogane, M., Sinclair, W.R., Nanan, K.K., Mandler, M.D. and et al. (2018) Acetylation of Cytidine in mRNA promotes translation efficiency. Cell, 175, 1872–1886.

23. Ringeard, M., Marchand, V., Decroly, E., Motorin, Y. and Benasser, Y. (2019) FTSJ3 is an RNA 2′-O-methyltransferase recruited by HIV to avoid innate immune sensing. Nature, 565, 500–504.

24. 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.

25. Dominissini, D., Moskitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amargiño, N., Kupiec, M. and et al. (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature, 485, 201–206.

26. Chen, K., Lu, Z., Wang, X., Fu, Y., Luo, G.Z., Liu, N., Han, D., Dominissini, D., Dai, Q., Pan, T. and 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.
76. Song,B., Tang,Y., Wei,Z., Liu,G., Su,J., Meng,J. and Chen,K. (2020) PIANO: a web server for pseudouridine-site (Ψ) identification and functional annotation. *Front. Genet.*, 11, 88.

77. Lian,L., Lei,X., Meng,J. and Wei,Z. (2020) ISGm1A: integration of sequence features and genomic features to improve the prediction of human m1A RNA methylation sites. *IEEE Access*, 8, 81971–81977.

78. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, 4, 1073–1081.

79. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, 7, 248–249.

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

81. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L., Edwards,K.J., Day,I.N. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, 34, 57–65.

82. Wang,K., LIM. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38, e164.

83. Zhu,Y., Xu,G., Yang,Y.T., Xu,Z., Chen,X., Shi,B., Xie,D., Lu,Z.J. and Wang,P. (2018) POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.*, 47, D203–D211.

84. Agarwal,V., Bell,G.W., Nam,J.-W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4, e05005.

85. Li,J.-H., Liu,S., Zhou,H., Qu,L.-H. and Yang,J.-H. (2013) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and e-miRNA regulatory networks. *Nucleic Acids Res.*, 42, D92–D97.

86. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,A., Morales,J., Mountjoy,E., Sollis,E. et al. (2018) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted and array-based association statistics 2019. *Nucleic Acids Res.*, 47, D1005–D1012.

87. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. et al. (2015) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, 44, D682–D688.

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

89. Dong,Q., Li,F., Xu,Y., Xiao,J., Xu,Y., Shang,D., Zhang,C., Yang,H., Tian,Z. and Mi,K. et al. (2019) RNAAcTDrug: a comprehensive database of RNAs associated with drug sensitivity from multi-omics data. *Brief. Bioinform.*, bbz142.

90. Bues,R., Yao,E., Díeck,M., Hayes,R.D. and Holmes,I.H. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, 17, 66.

91. Boccaletto,P., Machnicka,M.A., Purta,E., Piątkowski,P., Baginski,B., Wirecki,T.K., de Crécy-Lagard,V., Ross,R., Limbach,P.A., Kotter,A. et al. (2017) MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.*, 46, D303–D307.

92. Song,B., Chen,K., Tang,Y., Ma,J., Meng,J. and Wei,Z. (2020) PSI-MOUSE: predicting mouse pseudouridine sites from sequence and genome-derived features. *Evolutionary Bioinformatics*, 16, 1176934320297525.

93. Chen,W., Peng,P., Song,X., Lv,H. and Lin,H. (2019) iRNA-m7G: identifying N7-methylguanosine sites by fusing multiple features. *Mol. Ther. - Nucleic Acids*, 18, 269–274.

94. Kiss-Laszlo,Z., Henry,Y., Bachelerie,J.-P., Caizergues-Ferrer,M. and Kiss,T. (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, 85, 1077–1088.

95. Watkins,N.J., Gottschalk,A., Neubauer,G., Kastner,B., Fabrizio,P., Mann,M. and Lührmann,R. (1998) Chb5p, a potential pseudouridine synthase, and Nhp2p, a putative RNA-binding protein, are present together with Gar1p in all H BOX/ACA-motif snoRNPs and constitute a common bipartite structure. *RNA*, 4, 1549–1568.

96. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, 283, 1168–1171.

97. Badis,F., Fromont-Racine,M. and Jacquier,A. (2003) A snoRNA that guides the two most conserved pseudouridine modifications within RNA confers a growth advantage in yeast. *RNA*, 9, 771–779.

98. Motorin,Y. and Helm,M. (2011) RNA nucleotide methylation. *Wiley Interdiscip. Rev.: RNA*, 2, 611–631.

99. Chiu,Y., Xu,G., Yang,Y.T., Xu,Z., Chen,X., Shi,B., Xie,D., Lu,Z.J. and Wang,P. (2018) POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.*, 47, D203–D211.

100. Liu,L., Lei,X., Fang,Z., Tang,Y., Meng,J. and Wei,Z. (2020) LIHOPFIE: improving lncRNA methylation site prediction using an ensemble predictor. *Frontiers in Genetics*, 11, 545.

101. Liu,L., Lei,X., Meng,J. and Wei,Z. (2020) WITMSG: large-scale prediction of human intronic m6A RNA methylation sites from sequence and genomic features. *Cure. Genomics*, 21, 67–76.

102. Adachi,H., De Zoya,M.D. and Yu,Y. (2018) Post-transcriptional pseudouridylation in mRNA as well as in some major types of noncoding RNAs. *Biochim. Biophys. Acta (BBA) - Gene Regul. Mech.*, 1862, 230–239.

103. Zaringhalam,M. and Papavasiliou,F.N. (2016) Pseudouridylation meets next-generation sequencing. *Methods*, 107, 63–72.

104. Hussain,S., Alekis,J., Blanco,S., Dietmann,S. and Frye,M. (2013) Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol.*, 14, 215.

105. Capitanchik,C., Toolan-Kerr,P., Luscombe,N.M. and Ule,J. (2020) LITHOPHONE: improving lncRNA methylation site prediction using an ensemble predictor. *Frontiers in Genetics*, 11, 545.

106. Liu,L., Lei,X., Meng,J. and Wei,Z. (2020) WITMSG: large-scale prediction of human intronic m6A RNA methylation sites from sequence and genomic features. *Cure. Genomics*, 21, 67–76.

107. Zaninghalam,M. and Papavasiliou,F.N. (2016) Pseudouridylation meets next-generation sequencing. *Methods*, 107, 63–72.

108. Hussain,S., Alekis,J., Blanco,S., Dietmann,S. and Frye,M. (2013) Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol.*, 14, 215.

109. Capitanchik,C., Toolan-Kerr,P., Luscombe,N.M. and Ule,J. (2020) LITHOPHONE: improving lncRNA methylation site prediction using an ensemble predictor. *Frontiers in Genetics*, 11, 545.

110. Liu,L., Lei,X., Meng,J. and Wei,Z. (2020) WITMSG: large-scale prediction of human intronic m6A RNA methylation sites from sequence and genomic features. *Cure. Genomics*, 21, 67–76.