m6AVar: a database of functional variants involved in m\(^6\)A modification

Yueyuan Zheng\(^1,2,\dagger\), Peng Nie\(^2,\dagger\), Di Peng\(^2,\dagger\), Zhihao He\(^2\), Mengni Liu\(^2\), Yubin Xie\(^2\), Yanyan Miao\(^2\), Zhixiang Zuo\(^1,*\) and Jian Ren\(^{1,2,3,*}\)

\(^1\)Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou 510060, China, \(^2\)State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China and \(^3\)Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China

Received August 15, 2017; Revised September 17, 2017; Editorial Decision September 21, 2017; Accepted September 23, 2017

ABSTRACT

Identifying disease-causing variants among a large number of single nucleotide variants (SNVs) is still a major challenge. Recently, N\(^6\)-methyladenosine (m\(^6\)A) has become a research hotspot because of its critical roles in many fundamental biological processes and a variety of diseases. Therefore, it is important to evaluate the effect of variants on m\(^6\)A modification, in order to gain a better understanding of them. Here, we report m6AVar (http://m6avar.renlab.org), a comprehensive database of m\(^6\)A-associated variants that potentially influence m\(^6\)A modification, which will help to interpret variants by m\(^6\)A function. The m\(^6\)A-associated variants were derived from three different m\(^6\)A sources including miCLIP/PA-m\(^6\)A-seq experiments (high confidence), MeRIP-Seq experiments (medium confidence) and transcriptome-wide predictions (low confidence). Currently, m6AVar contains 16,132 high, 71,321 medium and 326,915 low confidence level m\(^6\)A-associated variants. We also integrated the RBP-binding regions, miRNA-targets and splicing sites associated with variants to help users investigate the effect of m\(^6\)A-associated variants on post-transcriptional regulation. Because it integrates the data from genome-wide association studies (GWAS) and ClinVar, m6AVar is also a useful resource for investigating the relationship between the m\(^6\)A-associated variants and disease. Overall, m6AVar will serve as a useful resource for annotating variants and identifying disease-causing variants.

INTRODUCTION

Rapid improvement in high-throughput sequencing technology has resulted in the identification of millions of single nucleotide variants (SNVs) across multiple genomes. A major challenge in delineating these variants is to distinguish the functional variants from the rest. In recent years, numerous studies have been undertaken to explore disease-associated nonsynonymous SNVs that alter amino acid at the protein level (1). Nevertheless, there is growing evidence showing many synonymous SNVs, which do not alter the amino acid sequences of proteins and are considered ‘silent’ mutations, also affect the function of genes and cause various diseases, suggesting a role in transcriptional or post-transcriptional regulation (2). Many studies have shown that variants have the capacity to alter the secondary structure of RNA, influence RNA–protein interactions (3), and change the splicing sites of exonic splicing enhancers and silencers (4) as well as genetic information by means of RNA editing (5). We speculated that variants might also influence RNA modification (e.g. m\(^6\)A) by changing the RNA sequences of the target sites or key flanking nucleotides.

N\(^6\)-Methyladenosine (m\(^6\)A) is a pervasive RNA modification in eukaryotes, that is involved in various biological processes such as embryonic development (6), cell apoptosis (7), spermatogenesis (8) and circadian rhythms (9). Recent development of the high-throughput sequencing techniques for m\(^6\)A (known as Methylated RNA Immunoprecipitation Sequencing (MeRIP-Seq), Photo-Crosslinking-Assisted m\(^6\)A Sequencing Strategy (PA-m\(^6\)A-seq) and m\(^6\)A individual-nucleotide-resolution cross-linking and immunoprecipitation sequencing (miCLIP)) has provided thousands of m\(^6\)A sites and deep insights into the m\(^6\)A machinery (10–13), revealing the essential regulatory roles of...
m\(^6\)A in RNA splicing, miRNA function and RNA stability (7,8,14,15).

An increasing number of studies have revealed that dysregulation of m\(^6\)A modification may impact various diseases. It has been found that the knockout of METTL3 in human cancer cells decreased the invasion of tumor cells (16). The activation of ALKBH5 in hypoxic breast cancer cells would promote cancer stem cell enrichment (17). In addition, previous studies have suggested that the m\(^6\)A eraser FTO is related to metabolism dysfunction (18) and acts as an oncogenic role in Acute Myeloid Leukemia (19). Furthermore, a previous study in mice indicated that m\(^6\)A might be important in neurodevelopmental processes (10). To further investigate the potential pathogenesis of m\(^6\)A modification, it is necessary to evaluate the effect of variants on m\(^6\)A modification. This will be helpful for both an understanding of the variants’ pathogenic molecular mechanisms and the identification of additional disease-causing variants.

As a result of the intensifying researches and accumulating data on the m\(^6\)A machinery, databases on m\(^6\)A modification have emerged in recent years. In 2015, Liu et al. collected 74 samples from 22 different m\(^6\)A-seq experiments and constructed MeT-DB, the first comprehensive m\(^6\)A database of the mammalian transcriptome (20). Later, Sun et al. developed the RNA modification database called ‘RMBase’ that includes 226,000 m\(^6\)A sites and 10,005 m\(^5\)C sites (21). Although the above databases have greatly aided research of m\(^6\)A functions, there is still no specific resource that would help study influence of variants on m\(^6\)A modification.

In this study, we present m6AVar (http://m6avar.renlab.org), a comprehensive database that allows the annotation, visualization and exploration of m\(^6\)A-associated variants in humans and mice (Figure 1). A great number of the m\(^6\)A-associated variants were derived from millions of germlines and somatic variants as well as three different m\(^6\)A sources that included miCLIP experiments, PA-m\(^6\)A-seq experiments, MeRIP-Seq experiments and transcriptome-wide predictions. We further annotated the m\(^6\)A-associated variants by checking whether they localized in regions with RBP binding sites, as well as miRNA targets and splicing sites. Moreover, disease-associated data from GWAS and ClinVar database were also integrated into m6AVar, which allows users to explore the underlying relationship between the m\(^6\)A machinery and diseases.

MATERIALS AND METHODS

Data resource

Germline and somatic variants were obtained from dbSNP and TCGA, respectively (Supplementary Table S1). We preserved those variants within the exonic regions for subsequent analysis. All of the m\(^6\)A sites were derived from seven miCLIP experiments, two PA-m\(^6\)A-seq experiments, 244 MeRIP-Seq experiments (Supplementary Table S2) and a transcriptome-wide prediction based on Random Forest algorithm. To identify the potential roles of m\(^6\)A-associated variants in post-transcriptome regulation, the RBP binding sites from starBase2 (22) and CLIPdb (23) (Supplementary Table S3), the miRNA–RNA interactions from starBase2 and the canonical splice sites (GT-AG) from Ensembl annotations were collected. In addition, we obtained a large number of disease-associated SNPs from different data sets (GWAS catalog (24), Johnson and O’Donnel (25), dbGAP (26), GAD (27) and ClinVar (28)) to perform disease-association analysis. The detailed description and statistics for these data resources can be found in Supplementary Table S4.

Data preprocessing

As the raw data collected from the diverse databases utilized different data formats, it is essential to unify them under standard procedures. To do this, the genomic coordinates of all of the data resources were converted to GRCh37 for the human and GRGm38 for the mouse using the LiftOver (29). The location of each m\(^6\)A site was then annotated by the transcript structure, including the CDS, 3’ UTR, 5’ UTR, start codon and stop codon etc. All the genomic information on the non-coding genes are from DASHR (30), miRBase (v21) (31), GtRNAdb (32) and piRNABank (33). Furthermore, all of the SNPs were annotated by ANNOVAR (updated to 1 February 2016) in two steps (34). First, we studied the conservation of evolutionary sequence in the m\(^6\)A-associated variants using phastCons 100-way and 60-way gene conservation scores for the human and mouse respectively (35). Second, we measured the deleterious level of each variant by integrating the results from five predictors of variant function (SIFT (36), PolyPhen2 HVAR (37), PolyPhen2 HDIV (37), LRT (38) and FATHMM (39)). Each variant was scored from 0 to 5 scale by counting deleterious levels of the variants obtained by the above five methods according to their thresholds curated by the dbNSFP database (40).

Derivation of the m\(^6\)A sites

The m\(^6\)A sites in m6AVar were derived using three different strategies with confidence levels ranging from high to low as illustrated below:

1. The m\(^6\)A sites having a high confidence level were extracted from the published single-nucleotide resolution m\(^6\)A sites in the miCLIP experiments. Besides, we also obtained m\(^6\)A sites that conformed to the DRACH (where D = A, G or U; R = G or A; H = A, C or U) motif from PA-m\(^6\)A-seq experiments (10,11,41).

2. The m\(^6\)A sites having a medium confidence level were predicted from the previously published MeRIP-seq data. We first downloaded all the MeRIP-Seq samples from the GEO database as raw data. Quality control was performed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and the sequencing adaptors were removed using Trimomatic (v0.33) (42). A minimum of 25 nucleotides was required for unambiguous alignment. All qualified reads were mapped to reference genomes (GRCh37 for human and GRGm38 for mouse) by TopHat (v2.1.1) using default parameters (43). We applied three peak callers (MACS2 (44), MetPeak (45) and Meyer’s method (10)) to identify the m\(^6\)A peaks separately. MSPC (46) was then ap-
Figure 1. Overall design and construction of m6AVar.

### Derivation of the m6A-associated variants

We defined a variant as an m6A-associated variant by evaluating whether it has the potential to alter the DRACH motif or other sequence features essential for m6A modification. According to various levels of confidence, we extracted the corresponding m6A-associated variants as follows.

1. For m6A sites having a high confidence level, we retained the variants that located nearby the m6A sites and then looked for the variants that disrupt DRACH motif around the m6A sites, such as changing from D(A/G/U) to C, R(G/A) to C/T, A to C/G/U, C to G/A/U, H (A/C/U) to G.

2. For m6A sites with a medium confidence level, the m6A-associated variants were derived from the intersection between the variants and the m6A sites generated from MeRIP-Seq experiments. The Random Forest prediction model was subsequently applied to find the variants in the m6A site region that change the DRACH motif or other sequence features.

3. For m6A sites with a low confidence level, we separately predicted the m6A status for the sequence around the variants in both the reference sequence and mutant sequence compared to reference sequence were defined as m6A-loss variants. In the opposite case, they were defined as m6A-gain variants.

### Post-transcriptional regulation association analysis

First, m6A-associated variants were intersected with RNA-binding proteins (RBPs) regions for the same sample. In terms of miRNA targets, we matched all of the m6A-associated variants with miRNA targets to obtain the m6A-associated variants which potentially impacted miRNA-target interactions. Additionally, we extracted 100 bp upstream from the 5’ splicing sites and 100 bp downstream from the 3’ splicing sites. Subsequently we matched all of the m6A-associated variants with these regions to obtain the splicing sites affected by the m6A-associated variants.
**RESULTS**

**Database content**

m6AVar contains three different confidence levels of m⁶A-associated variants for human and mouse (Table 1). The m⁶A-associated variants with high confidence level were derived from miCLIP or PA-m⁶A-Seq experiments. For human, there are 13 703 and 144 534 m⁶A-associated variants from miCLIP and PA-m⁶A-Seq experiments, respectively. For mouse, there are 54 222 and 54 222 m⁶A-associated variants from MeRIP-Seq and Prediction experiments, respectively. Moreover, m6AVar contains many associated data, such as RBPs, miRNA targets, splicing sites, and disease-related variants. As a result, we obtained 296 933 and 29 982 low confidence level m⁶A-associated variants in human and mouse, respectively.

Moreover, m6AVar contains many associated data, such as RBPs, miRNA targets, splicing sites, and disease-related variants. As a result, we obtained 296 933 and 29 982 low confidence level m⁶A-associated variants in human and mouse, respectively.

**Disease association analysis**

LD analysis was performed for each GWAS disease-associated SNP. We used Haploview to obtain its LD mutations with a parameter of $r^2 > 0.8$ in at least one of the four populations from CHB, CEU, JPT and TSI (19). Then we selected all the m⁶A-associated variants by mapping them with GWAS disease-associated SNPs and their LD mutations. Moreover, we also collected ClinVar data in order to annotate the m⁶A-associated variants with specific functions.

**Database and web interface implementation**

All the metadata in m6AVar were stored and managed in MySQL tables. The web interfaces were implemented in Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS) and Hypertext Preprocessor (PHP). In order to provide visualization of all the analysis results, multiple statistical diagrams were shown by EChars and genome browser was implemented using Jbrowser (47).

**Web interface and usage**

m6AVar provides user-friendly web interfaces that enable users to browse, search and download all of the m⁶A-associated variants in the database.

**Search.** m6AVar provides four modes to query the database, i.e. by RsID, Gene, Chromosome region and Disease. Here, we illustrate an example to show how to utilize m6AVar by search function (Figure 2). We sought to undertake an investigation of m⁶A modification in breast cancer using m6AVar. Through the ‘RsID’ search mode, users can check whether a variant of interest functionally affects the m⁶A status. In addition, m⁶A-associated variants in known breast cancer-related genes may be obtained by using the ‘Gene’ mode (Figure 2A). Taking the human tumor suppressor gene *BRCA1* as an example, 102 m⁶A-associated variants in *BRCA1* are presented as a table in the search results page (Figure 2B). Among them, 11, 26 and 65 m⁶A-associated variants were derived from miCLIP (high confidence), MeRIP-Seq (medium confidence) and prediction (low confidence), respectively (Figure 2C). A statistical plot shows the number of germline and somatic m⁶A-associated variants (Figure 2D). Users may obtain the related RBPs, miRNA targets, splicing sites and diseases from...
A schematic workflow of the search interface in m6AVar. (A) m6AVar provides the four search modes of Rs ID, Gene, Region and Disease. (B) Snapshot of search results for ‘BRCA1’ using the ‘Gene’ search mode. Basic information on all of the m6A-associated variants located in the ‘BRCA1’ output presented as a table. (C and D) Distribution of m6A-associated variants in the different sources and databases. (E) Detailed information on m6A-associated variants related to post-transcriptional regulation and disease. (F) Visualization of specific m6A-associated variants with JBrowse.

Figure 2

The detailed information on each variant (Figure 2E). Furthermore, m6AVar also allows users to find more disease-related variants directly through ‘Disease’ mode. In order to facilitate follow-up experimental studies, it allows users to customize results with the advanced search and to sort the table by clicking on the column names. Furthermore, we applied the JBrowse Genome Browser to visualize every m6A-associated variant. Users can select the tracks of interest to be shown, such as gene information, SNP site, m6A site, RBP binding regions, miRNA targets and the MeRIP-Seq peak level from the different samples (Figure 2F).

Browse. The ‘Browse’ page displays: (i) Summary of m6A-associated variants from three m6A sources (with a high, medium and low confidence level) (Supplementary Figure S2). (ii) Statistical graphs showing the overall functional gain and loss variants’ frequency distribution in a circular layout (Supplementary Figure S3), and m6A-associated variants’ distribution in gene regions and gene types as well as other databases (Supplementary Figure S4). (iii) Browse m6A-associated variants by gene types. To retrieve data more efficiently, various filters, such as gene types, associations and confidence levels are provided (Supplementary Figure S5).

Download. all data in the database can be downloaded from the ‘Download’ page, and a detailed introduction of m6AVar database as well as tutorial are available on the ‘Help’ page.

DISCUSSION

m6AVar is a comprehensive database of the m6A-associated variants that localize in the vicinity of m6A sites and potentially influence m6A modification in human and mouse. Currently, m6AVar holds ~352 000 m6A-associated germline variants and ~62 000 m6A-associated somatic variants, most of them were enriched in protein-coding genes (dbSNP147, 95.77%; dbSNP146, 92.12% and TCGA, 98.89%). The m6A-associated variants that can potentially affect RBP-binding regions, miRNA-targets and splicing sites were discovered by systematic association analyses. Furthermore, disease-related variants from GWAS and ClinVar have been intersected with the m6A-associated variants to identify the pathogenic variations contributing to dysregulation of m6A modification.

m6AVar has the following advantages in comparison with MeT-DB and RMBase. (i) m6AVar is a specific database dedicated to the investigation of the functional association between variants and m6A modification. (ii) m6AVar integrates somatic variants of 34 cancers from TCGA, which will help to reveal the potential mechanisms of m6A in cancer. (iii) m6AVar provides detail annotations and genomic coordinates for each variant and related m6A site. This will help biologists determine its relevant biological features. (iv) m6AVar integrates the results from association analyses with RBP-binding regions, miRNA-targets and splicing sites, revealing the potential relationship among variants, m6A modification and other post-transcriptional reg-
ulation. (v) More than 2000 disease-related variants have been identified by linking the m^6^A-associated variants with GWAS and ClinVar data, which may assist the community in identifying the functional disease-causing variants. (vi) m^6^AVar is a user-friendly database with multiple statistical diagrams and genome browser through which users can browse all of the m^6^A-associated variants and search interested data by various criteria.

In conclusion, m^6^AVar provides useful information on m^6^A-associated variants to help experimental biologists interpret the disease-related variants by m^6^A function and explore the molecular mechanism of m^6^A modification. m^6^AVar will be continually updated whenever new high-throughput m^6^A sites data and variants data are made available in public databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [31771462, 81772614, 31471252, U1611261]; National Key Research and Development Program [2017YFA0106700]; Guangdong Natural Science Foundation [2014QTC1387, 2014A030313181]; Science and Technology Program of Guangzhou, China [201604020003]. Funding for open access charge: National Natural Science Foundation of China [31471252].

Conflict of interest statement. None declared.
30. Leung, Y.Y., Kuksa, P.P., Amlie-Wolf, A., Valladares, O., Ungar, I. H., Kannan, S., Gregory, B.D. and Wang, L.S. (2016) DASHR: database of small human noncoding RNAs. *Nucleic Acids Res.*, 44, D216–D222.

31. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, 42, D68–D73.

32. Chan, P.P. and Lowe, T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, 44, D184–D189.

33. Sai Lakshmi, S. and Agrawal, S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.*, 36, D173–D177.

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

35. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15, 1034–1050.

36. 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.

37. 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.

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

39. 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.

40. Liu, X., Wu, C., Li, C. and Boerwinkle, E. (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, 37, 235–241.

41. Schwartz, S., Mumbach, M.R., Jovanovic, M., Wang, T., Maciag, K., Bushkin, G.G., Mertins, P., Ter-Ovanesyan, D., Habib, N., Cacchiarelli, D. et al. (2014) Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5’ sites. *Cell Rep.*, 8, 284–296.

42. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

43. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14, R36.

44. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9, R137.

45. Cui, X., Meng, J., Zhang, S., Chen, Y. and Huang, Y. (2016) A novel algorithm for calling mRNA m6A peaks by modeling biological variances in MeRIP-seq data. *Bioinformatics*, 32, i378–i385.

46. Jalili, V., Matteucci, M., Masson, M. and Morelli, M.J. (2015) Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics*, 31, 2761–2769.

47. Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. * Genome Res.*, 19, 1630–1638.