DirectRMDB: a database of post-transcriptional RNA modifications unveiled from direct RNA sequencing technology

Yuxin Zhang1,2,5,†, Jie Jiang2,5,†, Jiongming Ma1,2,5,†, Zhen Wei2,6,*, Yue Wang3,7, Bowen Song3,5, Jia Meng2,4,5, Guifang Jia8, João Pedro de Magalhães6, Daniel J. Rigden5, Daiyun Hang2,7,* and Kunqi Chen1,9,*

1Key Laboratory of Ministry of Education for Gastrointestinal Cancer, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, Fujian 350004, China, 2Department of Biological Sciences, Xi’an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China, 3Department of Mathematical Sciences, Xi’an Jiaotong-Liverpool University, Suzhou, Jiangsu 205123, China, 4AI University Research Centre, Xi’an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China, 5Institute of Systems, Molecular and Integrative Biology, Xi’an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China, 6Institute of Life Course and Medical Sciences, Xi’an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China, 7Department of Computer Science, University of Liverpool, L69 7ZB, Liverpool, UK, 8Synthetic and Functional Biomolecules Center, Beijing National Laboratory for Molecular Sciences, Key Laboratory of Bioorganic Chemistry and Molecular Engineering of Ministry of Education, College of Chemistry and Molecular Engineering, Peking University, Beijing, China and 9Fujian Key Laboratory of Tumor Microbiology, Department of Medical Microbiology, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, Fujian 350004, China

Received August 07, 2022; Revised October 20, 2022; Editorial Decision October 22, 2022; Accepted October 25, 2022

ABSTRACT

With advanced technologies to map RNA modifications, our understanding of them has been revolutionized, and they are seen to be far more widespread and important than previously thought. Current next-generation sequencing (NGS)-based modification profiling methods are blind to RNA modifications and thus require selective chemical treatment or antibody immunoprecipitation methods for particular modification types. They also face the problem of short read length, isoform ambiguities, biases and artifacts. Direct RNA sequencing (DRS) technologies, commercialized by Oxford Nanopore Technologies (ONT), enable the direct interrogation of any given modification present in individual transcripts and promise to address the limitations of previous NGS-based methods. Here, we present the first ONT-based database of quantitative RNA modification profiles, DirectRMDB, which includes 16 types of modification and a total of 904,712 modification sites in 25 species identified from 39 independent studies. In addition to standard functions adopted by existing databases, such as gene annotations and post-transcriptional association analysis, we provide a fresh view of RNA modifications, which enables exploration of the epitranscriptome in an isoform-specific manner. The DirectRMDB database is freely available at: http://www.rnamd.org/directRMDB/.

INTRODUCTION

Conceptually similar to DNA modifications, RNA molecules undergo chemical modifications as well. The first RNA chemical modifications were documented in the 1950s in tRNAs and rRNAs (1). To date, >170 different modification types have been described, including N6-methyladenosine (m6A), pseudouridine (Ψ), N5-methylcytosine (m5C), N1-methyladenosine (m1A), methylation of 2’-O in the four nucleotides (i.e. Am, Tm, Cm and Gm) and N7-methylguanosine (m7G) (2–7). With recent advanced technologies to map these RNA modifications, our understanding of them has been revolutionized, and they are now understood to be far more widespread and important than previously thought. Systematic studies...
of this post-transcriptional regulatory layer have revealed its profound roles in shaping cellular processes, modulating disease risks, and governing cellular fate (8–11). For instance, m⁶A, one of the most prevalent RNA modifications, is proven to regulate cardiac gene expression, cell growth, stress response, stabilize junctional RNA, etc (12–15). Pseudouridine, the first discovered post-transcriptional modification (16), has recently been implicated in tumor development, maintenance, and progression (17).

RNA-seq has become a popular choice for analyzing complex epitranscriptomics. However, next-generation sequencing (NGS) platforms are typically blind to nucleotide modifications and thus need specific protocols to highlight RNA modifications on the molecules. These typically involve three strategies: (i) antibody immunoprecipitation, which specifically recognizes the modified bases with antibodies; (ii) enzyme-digestion, where RNAs are digested with modification-sensitive enzymes; (iii) chemical treatment, using chemical compounds that selectively react with the modified ribonucleotide of interest. Example of immunoprecipitation methods include m⁶A-seq (3), PADS-Seq (18), m⁶A-CLIP-seq (19), miCLIP (20), m⁶A-LAIC-Seq (21), m⁶ACE-Seq (22), Mazter-seq (23), m⁶A-REF-Seq (24), and DART-seq (25) are enzyme digestion-based methods that quantify m⁶A modification with single-nucleotide resolution. Pseudo-seq (6) and AlkAniline-seq (26) are typical chemical-based detection methods. These methods are similar in that they enrich fragments harboring modified ribonucleotides, followed by high-throughput sequencing and bioinformatics analysis to detect these changes.

Although these methods provide invaluable information, they are limited by the availability of high-quality antibodies and the lack of practical chemical reactivities towards a particular RNA modification (27). Thus, only a few of the over 170 known modification types can be accurately and effectively profiled. When selective antibodies or chemical treatments are available, the RNA modification to be studied should be chosen beforehand, and customized protocols must be designed for the chosen type, limiting our ability to characterize the epitranscriptome in a systematic and flexible manner (28). Also, these methods require multiple ligation and amplification steps during the library preparation, introducing undesired biases and artifacts (29,30). Finally, with respect to the sequencing itself, NGS platforms face the problem of short read length. Mapping modifications on highly repetitive splicing isoforms and characterizing the co-occurrence of distant modifications in the same transcripts remain challenges (27). Thus, most of existing NGS-based methods have the isoform-ambiguity issue and they report only genome-coordinate of RNA modifications.

The continuing discoveries of novel classes of RNA modifications in various organisms call for more sensitive, plastic, and convenient modification profiling methods. A promising alternative to NGS technologies is the direct RNA sequencing (DRS) platform developed by Oxford Nanopore Technologies (ONT) (31). Each nucleotide will cause distinct ionic current signals as it passes through a sensitive channel. This platform infers the RNA sequence by deconvoluting the serial electric signal event when the molecule is threading through the sensitive protein channel (32). Natural modifications along the molecule can result in characteristic signals that suggest both the position and identity of modifications (33). Theoretically, direct RNA sequencing allows the real-time and simultaneous detection of any given modification in the native RNA molecule. Additionally, nanopore sequencing offers ultra-long reads that can cover the entire length of the RNA molecule, which benefits the study of RNA modifications on splicing isoforms (34).

ONT sequencing platforms have yielded robust data of reasonable quality, and several pilot studies have detected RNA modifications from the data. For example, EpiNano (m⁶A) (35), ELIGOS (27), DRUMMER (36) and the work of Parker et al. (37) screen RNA modifications by examining the sequencing error profiles. Another body of work, such as xPore (38), m⁶Anet (m⁶A) (39), MINES (m⁶A) (34), nanoM6A (m⁶A) (40), and nanoPsu (pseudouridine) (41) utilized the variation in current signal intensities. These tools were confirmed to have high accuracy for modification detection with single-nucleotide resolution.

To date, various comprehensive databases of RNA modification sites reported by NGS approaches are publicly available, including MODOMICS (42), RMBase (43), REPIC (44), m⁶A-atlas (45), m⁵C-atlas (46), MeT-DB (47), RMVar (48) and M6A2Target (49), which have together provided invaluable information to help decipher the complexities of epitranscriptomes. However, due to the low sensitivity and detection chemistry of NGS-based approaches, a huge proportion of modified sites have not been detected, and the landscape of RNA modifications on the transcriptome is yet to be well-studied (50). To address this gap, we have developed DirectRMDB, the first comprehensive database of RNA modification sites derived from direct RNA sequencing data. In this study, a collection of 16 quantitative modification profiles among 25 species and various cell types or tissues under different conditions were integrated from direct RNA sequencing samples. Data from other studies or techniques were also collected to validate the collected sites. A significant advantage of DirectRMDB is that it provides isoform-level information, including isoform-specific distributions of RNA modifications, isoform expression levels and secondary structure. We constructed a user-friendly web interface for the query, visualization, and sharing of the modification profiles and their association with transcriptional and post-transcriptional regulatory machinery (i.e. RNA binding proteins, miRNAs, splicing events), as well as their potential involvement in pathogenesis. As the first DRS-based database, DirectRMDB is expected to provide new insight into the complex epitranscriptome (Figure 1).

**MATERIALS AND METHODS**

**Collection of candidate modification sites**

125 direct RNA sequencing samples for 25 species, including 44 FAST5 and 81 FASTQ samples, were collected from 39 independent studies in the Gene Expression Omnibus (GEO) database (Supplementary Table S1). Eight modification detection tools, namely nanom6A (40), MINES (34), xPore (38), m⁶Anet (39), DRUMMER (36), ELIGOS (35),
Figure 1. The overall design of DirectRMDB. DirectRMDB is the first comprehensive database that integrates quantitative modification profiles determined by direct RNA sequencing. For quality assurance, eight different software tools for mining RNA modifications were rigorously integrated, and additional next-generation sequencing samples were collected for validation. DirectRMDB provides an isoform-specific view of modification sites, including their distribution on individual transcripts and the secondary structure predicted from the RNA primary sequences. The potential involvement of reported sites in pathogenesis and their potential interactions with post-transcriptional machinery can also be queried.

(27), the work of Parker et al. (37), and Nanopsu (41) were used to infer possible modification sites from samples (Table 1). It is worth noting that although direct RNA sequencing allows the detection of RNA modifications with an isoform-level resolution, some tools (e.g. MINES) still rely on genome-level features and thus cannot distinguish between transcripts. Supplementary Figure S1 shows the general workflow for candidate site collection. As the colors indicate, the eight tools can be categorized into three classes in terms of their required input and thus the different pipelines of data pre-processing: (i) Tombo-based (i.e. nanom6A and MINES): the raw data (FAST5) was resquiggled (i.e. a new assignment from current signal level data to the reference sequences was defined) with either transcriptome or genome reference using Tombo ‘resquiggle’ function. Specifically for MINES, the Tombo ‘de novo modification detection’ function was used to infer non-canonical bases from the re-squiggled current signals. This Tombo output was provided to nanom6A and MINES as input and candidate m6A sites returned. (ii) Nanopolish-based (i.e. xPore and m6Anet): the Nanopolish (51) ‘eventalign’ function was used to map the signal events extracted from the raw FAST5 sample to the reference transcriptome. m6Anet and xPore then analyzed the output TXT files to identify possible modifications. It is notable that xPore is a comparative method, which requires modification-free samples as control. (iii) BAM-based (i.e. DRUMMER, ELIGOS, the work of Parker et al., and NanoPsu): FAST5 samples were base called into FASTQ format with Guppy before alignment. Base-called reads, as well as downloaded FASTQ samples, were aligned to either reference genome or transcriptome using Minimap2 (52) with ‘–ax map-ont’ settings. The resulting SAM files were transformed into BAM files, sorted and indexed with Samtools (53). ELIGOS and NanoPsu examine the error distribution profiles from the alignment file directly, while DRUMMER and Simpson’s work requires control samples to perform the modification detection. Samples were analyzed by some or all of the eight tools depending on their data format (i.e. FASTQ or FAST5), the availability of control samples, and authentic reference sequences. References used for each species are summarized in Supplementary Table S2.

The landscape of RNA modifications on transcripts

Nanom6A, xPore, m6Anet, and DRUMMER detect bulk-level RNA modifications by examining either error distribution profiles or current signals distributions along transcripts. Nanom6A maps detected sites to the reference genome and present the results with genome coordinates. Therefore, only xPore, m6Anet and DRUMMER were used to predict modifications in individual transcripts. The work-
flow to run the three tools is shown in Supplementary Figure S1. To compare the isoform-level modification patterns and for the simplicity of results presentation, we converted the transcripts’ coordinates to genomic coordinates while keeping the isoform-level information.

Integration of results and validation by NGS methods

The collection of sites reported by each software could contain a significant proportion of false-positive sites. To ensure reliability, we collated the results from different samples and tools and then applied strict filtration criteria to generate reliable modification profiles for each species. To screen high-confident m6A sites, we searched for its known consensus modification profiles for each species. To screen high-confidence sites, we used the work of Parker et al. (38) and supplementary data of published works. To ensure reliability and save space, unlabeled ELIGOS results were excluded from the final proposed profiles but can be downloaded from DirectRMDB.

To further validate our results, we collected high-confidence modification sites and modification-enriched peaks derived from next generation sequencing samples (Supplementary Table S3). Additionally, multiple modification profiles reported by LC-MS techniques were downloaded from MODOMICS and RMBase. Cross-validation was performed between candidate and NGS/LC-MS-derived sites. Sites confirmed by other techniques were clearly labeled. We also compared our results with sites published by other ONT-based modification detection studies (38,54). Overlapped sites were indicated as well.

Table 1. brief description and comparison of modification calling tools

<table>
<thead>
<tr>
<th>Modification</th>
<th>Input</th>
<th>Isoform-level?</th>
<th>Control sample?</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>nanoPsu</td>
<td>(\psi)</td>
<td>No</td>
<td>No</td>
<td>Sequencing error</td>
</tr>
<tr>
<td>ELIGOS</td>
<td>Mixed</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Drummer</td>
<td>/</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>xspore</td>
<td>Nanopolish output</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>m6Anet</td>
<td>m6A</td>
<td>Yes</td>
<td>No</td>
<td>Current signal</td>
</tr>
<tr>
<td>MINES</td>
<td>Tombo output</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Nanom6A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ‘/’ means that detected modification types depend on the modification-free samples. For example, if an m6A-free sample is used as a control, reported sites are expected to be m6A methylation.

Secondary structure prediction

RNA plays a vital role in the cell, not only as an intermediate product for the transmission of genetic information, but also as a functional element. Single-stranded RNA molecules can fold into specific and stable structures. It is known that there is a strong association between their functions and structures (55). The three-dimensional structure of RNA molecules can only be determined by X-ray crystallography, nuclear magnetic resonance, and other laborious and high-cost methods (56). Therefore, we present the secondary structure of isoforms, which is easier to predict computationally. We use RNAfold (57), a widely used RNA secondary structure prediction software, with default parameter settings, to infer the structure from the RNA primary sequences. The landscapes of RNA modifications on each isoform under different conditions were annotated and highlighted on the predicted structure. For a better view, ultra-long reads (>2001nt) were cut into 2001nt fragments that contained modified bases.

Quantitative profiles of putative modification sites

44 FAST5 samples from nine species were collected to quantify the modification status of high-confident modification sites under different cell lines/tissues and conditions. The Tombo ‘de novo modification detection’ function was used to investigate non-canonical (i.e. modified) bases within individual reads and the fractions of modified reads aligned to each genomic position were output with Tombo ‘text output’ command. The modification fraction is used to quantify the modification status of reported sites. In addition to modification status, the transcripts’ expression profiles were also estimated from the BAM file with transcriptome reference using nanoCount (58), isoform expression level calculation software designed for direct RNA sequencing data.

Basic annotation for modification sites

Gene annotation files were downloaded from Gencode (Human and Mouse) (59) and NCBI (60) (other species). Those high-confidence sites (genome-wide) were annotated by collected gene annotations and were classified into different gene types and genomic regions using ChIPseeker (61). In addition to basic gene annotation, the potential interactions between modifications and splicing events, miRNA as well as RNA binding proteins (RBPs) were included in
human and mouse. miRNA target sites, RBP binding, and other events information were obtained from starBase (62), POSTAR (63) and the UCSC genome browser database (64), respectively. Since nanopore sensor protein takes a k-mer (4–6nt) as input each time, the presence of non-canonical bases could cause misleading signals thus influencing the deconvolution of adjacent bases. Therefore, for each site, we indicated the presence of other modifications within 5bp upstream and downstream as a warning of false positives.

Potential involvement of individual modification sites in pathogenesis

It is known that RNA modifications are closely related to the progression of diseases. To investigate the contribution of individual modification sites in disease development, we analyzed their positional relationship with potentially disease-associated genetic mutations. Sites that exactly overlapped with mutations were indicated. Collection of single nucleotides polymorphisms (SNPs), including both common variations and clinical mutations for human and mouse, was downloaded from dbSNP (65).

Database and web interface implementation

MySQL was used to store and manage the metadata. Hypertext Markup Language (HTML), Cascading style sheets (CSS) and Hypertext Preprocessor (PHP) were used to build the web interface. Genome browser JBrowse2 (66) was used to provide an integrated view of reference sequences, modification site information, related RBP binding, splicing, miRNA binding event, and associated SNPs.

RESULTS

The eight modification detection tools, applied to 125 direct RNA sequencing samples, suggested more than 16,000,000 candidate modification sites. By manually integrating, evaluating, and filtering the results, a total of 904,712 sites of 16 chemical modifications, namely m6A, ψ, 2′-O-Me, m5C, m7G, m5U, m6Am, D, f5C, Y and ac4c, across 25 species, including Homo sapiens, Mus musculus, Arabidopsis thaliana, Sis. scrofa and Escherichia coli, were confidently identified (Table 2 and Supplementary Table S4). Among these proposed sites, 149,353 human sites and 91,910 mouse sites were further confirmed by other techniques (i.e. NGS techniques and LC–MS) (Supplementary Figure S2). The landscapes of RNA modifications on human and mouse transcriptomes were evaluated. 225,041 sites in 26,039 distinct human transcripts and 228,558 sites in 21,413 mouse transcripts, corresponding respectively to 88,230 and 112,820 bases on the human and mouse genome, were found. We also predicted the secondary structure of isoforms and calculated their expression levels under specific cell lines and conditions.

Quantitative modification profiles (i.e. the fraction of modified reads) for nine species under 44 different cell lines/tissues and conditions were calculated. Gene annotation of 20 species was successfully performed, while annotation of the Bipolaris sorokiniana, Candida nivariensis, Chikungunya virus and influenza A virus failed since no feasible annotation file is available for these species. Since non-canonical bases can influence the deconvolution of adjacent nucleotides due to the nanopore sequencing chemistry, we evaluated the interaction between reported modification sites. A total of 105,581 sites were screened as consecutive modifications (i.e. has other modifications within 5bp up and downstream). For human and mouse, we also investigated the interaction between RNA modification and RNA binding proteins, splicing sites, and miRNA targets. For human, 171 RNA binding proteins, 826 miRNAs, and 101,587 splicing events are suggested to be associated with respectively 275,956, 54,390 and 108,738 modification sites. For mouse, we identified 39 RNA binding proteins, 905 miRNA and 79,010 splicing events that are related to RNA modifications. Also, 80,614 human sites and 563,012 mouse modification sites are documented SNP sites, suggesting their potential involvement in disease development.

Comprehensive atlas of various types of RNA modifications

We constructed DirectRMDB, the first database that integrates direct RNA sequencing data to explore post-transcriptional modifications of RNAs. A user-friendly web interface was provided to search, browse, visualize and download the 16 types of high confidently collected modification sites and their potential relationships with miRNA targets, RBPs, splicing events, and pathogenesis. JBrowse2 was integrated for interactive exploration of individual sites or regions of interest. We also provided isoform-level information, including the landscape of RNA modifications on individual transcripts, annotated secondary structures and transcripts expression levels under particular cell lines, tissues, and conditions. The DirectRMDB database is freely available at: http://www.rnamd.org/directRMDB/ and has a mirror at: www.xjtlu.edu.cn/biologicalsciences/directRMDB.

Table 2. The data statistics for DirectRMDB

<table>
<thead>
<tr>
<th>Species</th>
<th>m6A</th>
<th>ψ</th>
<th>2′-O-Me</th>
<th>m5C</th>
<th>m7G</th>
<th>m5U</th>
<th>m6Am</th>
<th>2′-O-Me</th>
<th>m5C</th>
<th>m7G</th>
<th>m5U</th>
<th>m6Am</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>195,871</td>
<td>134,834</td>
<td>1,506</td>
<td>26,033</td>
<td>2,979</td>
<td>3,803</td>
<td>3,650,26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>186,175</td>
<td>45,397</td>
<td></td>
<td>970</td>
<td>693</td>
<td>/</td>
<td>233,265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yeast</td>
<td>148</td>
<td>59</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>19</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>266</td>
<td></td>
</tr>
<tr>
<td>22 other species</td>
<td>203,973</td>
<td>102,241</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>11</td>
<td>306</td>
<td>225</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The numbers in the table indicate the total count of each modification type. In human, ‘Other’ refers to m7G, m5U, m6Am, and AtoI modifications, while in yeast, ‘Other’ refers to ac4c, D, Y and f5C. Please refer to Supplementary Table S4 for more details.
Figure 2. Case study on protein-coding gene RNF138. (A) Searching by gene name. (B) 101 sites of four modification types on the RNF gene. (C) Number of modifications detected by different software. (D) Pie chart of the number of modifications detected in each cell line. (E) Basic information of the example site with ID of ‘directRMDB_HomoSapiens_114258’. (F) Conditions involved in the example site and fraction of modified reads under different conditions. (G) Genome browser view of the example site and its relationships with RNA binding proteins, miRNAs and pathogenesis.
Case study on protein-coding gene RNF138

Ring finger protein 138 (RNF138) is a ubiquitin ligase belonging to the E3 ligase family, which harbors a ring finger protein domain, three zinc-finger-like domains, and a ubiquitin-interacting motif (67,68). It promotes cell survival via counteracting apoptotic signaling or directly influencing genome stability. Emerging evidence has linked the RNF138 protein with tumorigenesis, neurodegenerative disorders, and chronic inflammatory conditions (69,70). By searching through the *Homo sapiens* repository from directRMDB with the gene name RNF138 (Figure 2A), 101 entries, with one AtoI, one m6Am, 45 m6A and 54 /Psi1 sites, detected by MINES, m6Anet, Nanopsu, nanom6A, and ELIGOS from 7 different cell lines were returned (Figure 2C and D). Among the results, ELIGOS screened the majority of them (i.e. one AtoI, one m6Am, 25 m6A and 54 /Psi1 sites), suggesting its high sensitivity in screening modified bases. MINES only contributes to one m6A site, which can be explained by its strict evaluation criteria. Entries without available RBP binding, miRNA, splicing site, SNP, and transcriptomic information or not confirmed by NGS methods can be removed by clicking the corresponding buttons in the top filters box. Users can also retrieve sites of specific modification types from certain cell lines, tools or RNA types (e.g. mRNA, rRNA and tRNA) of interest. Detailed information on individual sites, including interaction with RNA binding proteins, miRNAs, and other sites, can be acquired by clicking the site ID.

Case study on lncRNA MALAT1

Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1), a long non-coding RNA (lncRNA) that has been confirmed to influence cancer development and metastasis (71). 73 modification sites from the DirectRMDB homo sapiens sets, including m6A, m5C, pseudouridine, and 2'-O-Me, were found on MALAT1 transcripts. Similarly, detailed information for individual modification sites, including interaction with RNA binding proteins, miRNAs, and other sites, can be acquired by clicking the site ID.

Case study: isoform level exploration of RNA modifications

TXNDC12 (chr1:52020131..52056171, GRCh38.p14 assembly) and KTI12 (chr1:52042103..52033810) are two genes that share common regions on chromosome 1. For an RNA modification site located within the shared regions, the fraction of modified reads on different samples shows the presence of other non-canonical bases nearby may cause misleading signals and thus influence the analysis. Although the example site is reported in multiple cases and was supported by other techniques, it is possible that the site is false positive as the result of adjacent misleading signals.
it can be difficult with NGS epitranscriptomics profiling methods to decide which genes or transcripts it belongs to. Fortunately, direct RNA sequencing technologies offer a solution to this isoform ambiguity problem thanks to the longer reads. By more precisely aligning reads to transcriptome references, modifications can be confidently located in an isoform-specific manner (Figure 4A). Here, the m6A site with ID of ‘directRMDB_HomoSapiens_142249’ is taken as an example (Figure 4B). From a genome-wide view, it is located on the shared region of TXNDC12 and KTI12 and was wrongly assigned to TXNDC12 by ChipSeeker in a random manner. On the directRMDB details page, in contrast, we can see that m6Anet unambiguously assigned this site to ENST00000371614, an isoform for the KTI12 gene, under
four different conditions. Also, expression levels of KTI12 isoforms under the four conditions are displayed (Figure 4C). The picture of the predicted RNA secondary structure with highlighted modified bases can be queried by clicking the ‘show’ button in the ‘secondary structure’ column (Figure 4D).

Conclusion

Maps of various RNA modifications have been constructed by coupling antibody immunoprecipitation or chemical probing with high-throughput sequencing. However, customized protocols are required for each RNA modification type, thus limiting our ability to characterize the plasticity of the whole epitranscriptomics systematically and in an unbiased fashion. Fortunately, the development of direct RNA sequencing platforms enables the mapping of diverse RNA modification types simultaneously and detection of any given modification present in native RNA molecules. With the rapid accumulation of direct RNA sequencing data and designed ONT tools, we constructed DirectRMDB, the first database of multiple RNA modifications unveiled by direct RNA sequencing technologies. By taking advantage of direct RNA sequencing technologies, DirectRMDB offers several novel features compared with existing epitranscriptomics databases: (i) since ONT direct sequencing generates ultra-long reads and is less vulnerable to isoform ambiguity, we confidently presented isoform-specific distribution of RNA modification sites. (ii) we provided transcript expression levels under different conditions. (iii) we integrated novel modifications sites that have not been detected by NGS-methods. Also, a user-friendly graphical interface integrated with a genome browser was constructed to facilitate the query, visualization, and analysis of this novel, fine-grained epitranscriptomics data. Due to the nature of ONT direct sequencing, the results might contain some false positive site. Therefore, we clearly indicated the tools, samples, and other techniques (i.e. NGS techniques and LC–MS) that support each site. Users could filter, select and use sites based on their understanding and knowledge. Overall, DirectRMDB provides a fresh view of the epitranscriptome. We will continue to update and improve the database by integrating the latest sequencing data and advanced tools to ensure that it remains a valuable resource for the research community.

DATA AVAILABILITY

No new data were generated or analysed in support of this research. The DirectRMDB database is freely available at: http://www.rrnmd.org/directRMDB/, and has a mirror at: www.xjtlu.edu.cn/biologicalsciences/directRMDB.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [32100519 and 31671373]; XJTU Key Program Special Fund [KSF-E-51 and KSF-P-02]; Scientific Research Foundation for Advanced Talents of Fujian Medical University [XRCZX202109]. Funding for open access charge: Scientific Research Foundation for Advanced Talents of Fujian Medical University [XRCZX202109].

Conflict of interest statement. None declared.

REFERENCES


Downloaded from https://academic.oup.com/nar/article/advance-article/doi/10.1093/nar/gkac1061/6830668 by guest on 20 November 2022


