Genome Biology

METHOD

Open Access

DEMINING: A deep learning model embedded framework to distinguish RNA editing from DNA mutations in RNA sequencing data

Zhi-Can Fu[1,2†], Bao-Qing Gao[1,2†], Fang Nan[1], Xu-Kai Ma[1] and Li Yang[1*]

†Zhi-Can Fu and Bao-Qing Gao contributed equally to this work.

*Correspondence: liyang_fudan@fudan.edu.cn

[1] Center for Molecular Medicine, Children's Hospital of Fudan University and Shanghai Key Laboratory of Medical Epigenetics, International Laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China
[2] Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Abstract

Precise calling of promiscuous adenosine-to-inosine RNA editing sites from transcriptomic datasets is hindered by DNA mutations and sequencing/mapping errors. Here, we present a stepwise computational framework, called DEMINING, to *d*istinguish RNA *e*diting and DNA *m*utations d*i*rectly from *R*NA sequen*cing* datasets, with an embedded deep learning model named DeepDDR. After transfer learning, DEMINING can also classify RNA editing sites and DNA mutations from non-primate sequencing samples. When applied in samples from acute myeloid leukemia patients, DEMINING uncovers previously underappreciated DNA mutation and RNA editing sites; some associated with the upregulated expression of host genes or the production of neoantigens.

Keywords: DNA mutation, RNA editing, RNA-seq, Deep learning, Transfer learning, Neoantigens, AML, IDR

Background

A myriad of single-nucleotide DNA mutations (DMs) and RNA variants, mainly adenosine-to-inosine (A-to-I) RNA-editing sites (REs) catalyzed by adenosine deaminase acting on RNA (ADAR) enzymes [1], have been identified by analyzing high-throughput RNA-sequencing (RNA-seq) datasets [2, 3]. Other than directly called from high-quality whole-genome sequencing (WGS) datasets [2], DMs could be also interpreted in RNA-seq datasets [4, 5], but were generally treated as noises during the profiling of A-to-I REs [1]. Given the resulting I preferentially pairs with cytidine (C), A-to-I RE is thus read as A-to-G (guanosine) mutation by the cellular machinery [6] or during sequencing analyses after reverse transcription [7–9]. In this scenario, A-to-I RE affects downstream RNA processing and function, such as recoding amino acid and altering alternative splicing.

Although precise identification of A-to-I REs has been long desired for fully understanding their biological roles [10], calls in A-to-I REs from RNA-seq datasets are fraught with

expressed DMs embedded in transcriptomic datasets [4, 5]. As the number of expressed DMs could be several times more than that of A-to-I REs [11], even a small proportion of misclassification of expressed DMs can result in a disproportionately high rate of false RE calling. To solve this problem, stepwise bioinformatic tools have been developed to precisely identify A-to-I REs from RNA-seq datasets by using parallel WGS and/or prior knowledge for the removal of DMs [11]. Alternatively, efficient A-to-I calling could be also achieved without the requirement of matched WGS but by analyzing multiple RNA-seq datasets for common/constitutive A-to-I or taking advantage of variable allelic linkage between DM and A-to-I RE from reads mapped to the same regions [12–14]. So far, a large number of A-to-I REs have been widely identified from massive human transcriptomes, mainly located in primate-specific *Alu* elements [11–13]. Compared to all 12 types of DMs identified in human genomic DNA, only A-to-I REs are predominantly examined in human transcriptomic RNA-seq datasets [15], and the detection of other non-A-to-I variants in RNA was suggested to be artifacts of sequencing and mapping errors [16–19]. Of note, distinct to the large number of human A-to-I REs and their predominant distribution in primate-specific *Alu* regions, the number of A-to-I REs in mouse is much smaller, possibly due to the lack of *Alu* elements in the mouse genome [20]. In this case, pipelines used for human A-to-I prediction could not be readily used for the similar prediction in mouse.

Other than canonical computational pipelines, machine learning and deep learning models have been recently applied for A-to-I prediction. However, the application of these models in A-to-I calling also faces challenges. On the one hand, the manual and knowledge-based nature of feature extraction used by machine learning may overlook certain inherent features for A-to-I formation. For example, recent findings highlight the impact of RNA secondary structure on RNA editing [21], which however was hardly included in the existing machine learning models, such as RED-ML [22] and RDDpred [23]. On the other hand, deep learning models that were trained only with reference genome sequences inevitably neglect pattern differences of DMs and REs beyond primary sequence, such as mutation frequency and allelic linkage, which may lead to non-negligible false positives [21]. Furthermore, published machine learning and deep learning methods for A-to-I prediction were usually confined to specific species and might not be applied to other species, hindering their broader application in RE calling. Computational tools that are capable of efficiently separating A-to-I REs from DMs are still desired.

Here, we developed a stepwise computational framework, named DEMINING, to *d*istinguish RNA *e*diting and DNA *m*utations d*i*rectly from R*N*A sequenc*ing* datasets, by taking advantage of an incorporated deep learning model, named DeepDDR. After transfer learning, the fine-tuned DEMINING framework could also achieve successful classification of REs and DMs from aligned RNA-seq reads in other species, suggesting its broad application of detecting REs and expressed DMs from massive transcriptomic datasets.

Results

Development of DEMINING to detect A-to-I REs and genomic DMs from RNA-seq datasets

We develop a two-step computational framework, called DEMINING, to directly detect A-to-I REs and genomic DMs from RNA-seq datasets only. In the first step, after quality control, read mapping, and piling, we applied stringent cutoffs to remove sequencing and mapping errors for the identification of high-confidence mutations, including

both DMs and REs. In the second step, we constructed a deep learning model, DeepDDR, to efficiently and precisely subgroup these high-confidence mutations as A-to-I REs or DMs (Fig. 1a and Methods). DeepDDR was trained with reliable sets of DMs and REs obtained from paired human WGS and RNA-seq datasets of 403 donors (Additional file 1: Fig. S1a and Fig. S1b, Additional file 2: Table S1), retrieved from the 1000 Genomes Project [2] and the Geuvadis consortium [24], respectively.

On the one hand, among ~ 31 millions of high-confidence DMs provided by 1000 Genomes Project, about 578,000 of them could be examined in the paired 403 RNA-seq datasets with expression (≥ 3 hits per billion mapped bases [14], HPB), mutation frequency (≥ 0.05), and mutation read (≥ 2) cutoffs. Similar distributions of all 12 DM types were observed between samples, exemplified by 2 of them from NA12890 and NA19213 (Additional file 1: Fig. S1a, Methods). On the other hand, high-confidence REs in each of 403 RNA-seq datasets were individually detected by the previously published RADAR pipeline [25], by which expressed DMs were filtered by the paired WGS dataset to obtain REs only (Additional file 1: Fig. S1b, Methods). Of note, the same expression

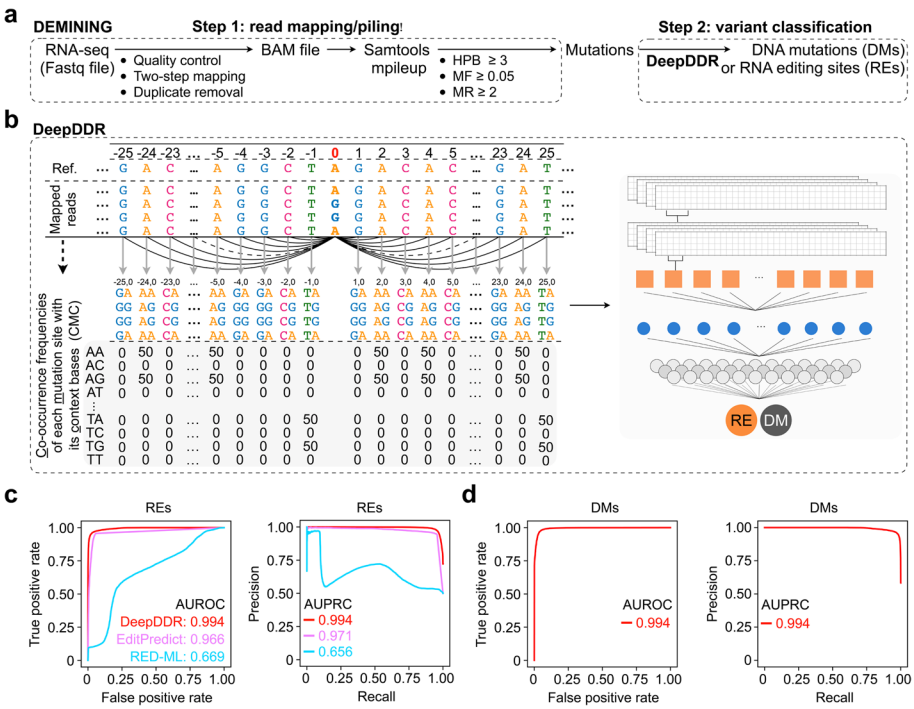

Fig. 1 Developing DEMINING embedded DeepDDR model for DNA mutations (DMs) and RNA editing sites (REs) classification. **a** Construction of a stepwise DEMINING computational framework for direct DNA mutation (DM) and RNA editing (RE) classification. HPB hits per billion mapped bases, MF mutation frequency, MR mutation read. See the "Methods" section for details. **b** Schematic diagram of an embedded DeepDDR model for DM and RE classification. Left, features extract strategy by the *co*-occurrence frequencies of each *m*utation site with its *c*ontext bases (CMC). Right, DeepDDR model architecture. See the "Methods" section for details. **c** Evaluation of different models on RE identification. Receiver operating characteristic (ROC, left) curves and precision recall curves (PRC, right) of DeepDDR (red), EditPredict (purple), and RED-ML (blue) were shown to indicate their performance on RE identification with the test set. Area under ROC (AUROC) and area under PRC (AUPRC) values of DeepDDR (red), EditPredict (purple), and RED-ML (blue) were included in the figure. **d** Evaluation of DeepDDR on DM identification. ROC (left) and PRC (right) of DeepDDR were shown to indicate its performance on DM identification with the test set. AUROC and AUPRC values of DeepDDR were included in the figure

(≥ 3 HPB) and mutation frequency (≥ 0.05) cutoffs, together with mutation read numbers (≥ 2), were used for the selection of REs. As previously reported [11–13], over 98% of identified REs in each of 403 RNA-seq datasets were A-to-G and T-to-C mismatches, mainly in *Alu* regions, indicating high-confidence A-to-I REs [11] called by RADAR [25] (Additional file 1: Fig. S1b). These high-confidence A-to-I REs identified by the RADAR pipeline in each of 403 RNA-seq datasets were then combined together to generate the collection of 122,872 unique REs (Additional file 3: Table S2). In parallel, 122,872 unique DMs were randomly selected from 578,000 of expressed DMs (Additional file 4: Table S3). The distribution of mutation frequencies of DMs and REs was shown in Additional file 1: Fig. S1c. These DMs and REs were then individually split into training, validation, and test sets with an 8:1:1 ratio for the model development and evaluation (Additional file 1: Fig. S1d).

To obtain features of these mutations, 51-nucleotide sequences containing a given DM (or RE) in the middle and its 50 context bases (25 bases in the upstream region and 25 bases in the downstream region) were extracted from aligned RNA-seq reads as described in "Methods (Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline, Step 1: RNA-seq read mapping/piling)" and then piled up for subsequent analysis (Fig. 1b, Methods). Next, a matrix of the *c*o-occurrence frequencies of each *m*utation site with its *c*ontext bases (CMC) was calculated, and CMC matrices of all DMs and REs in the training set were used as the input to train the DeepDDR model that contains two layers of convolutional neural network (CNN) (Fig. 1b, Methods). DeepDDR outputs a prediction probability from 0 to 1 for a given DM or RE, followed by the classification into the DM group (with prediction probabilities ≥ 0.5) or the RE group (with prediction probabilities < 0.5). When evaluated with the test set, the majority (> 98%) of DMs were shown with prediction probabilities between 0.5 and 1 and most (> 92%) REs with prediction probabilities between 0 and 0.5 (Additional file 1: Fig. S2a), suggesting the efficient separation of DMs and REs by DeepDDR. In addition, compared to previously published EditPredict [26] and RED-ML [22] methods, DeepDDR could find more true REs as well (Additional file 1: Fig. S2a-c) with much higher AUROC (area under the receiver operating characteristic curve) values (Fig. 1c and Additional file 1: Fig. S2d). Out of our expectation, DeepDDR achieved similar high (over 0.99) AUROC values on DM (Fig. 1d and Additional file 1: Fig. S2e) prediction, suggesting its potential application in detecting expressed DMs from massive transcriptomic datasets.

To further evaluate DeepDDR, an independent test set (SampleID: HG00145 from the 1000 Genomes Project and the Geuvadis consortium) containing paired WGS and RNA-seq was downloaded. From this independent test set, 17,230 true DMs and 5395 true REs were individually extracted from paired WGS and RNA-seq files with the same strategies (Fig. 2a, Additional file 1: Fig. S3a and S3b). Meanwhile, we applied DEMINING to the RNA-seq file of HG00145 only to obtain predicted DMs and REs by DeepDDR (Fig. 2a, Additional file 5: Table S4). By comparing true DMs/REs with predicted ones from the HG00145 independent test set, DeepDDR outperformed EditPredict and RED-ML, with the highest AUROC and AUPRC values on the prediction of REs (Fig. 2b and c, Additional file 1: Fig. S3c-e). Of special note, DeepDDR achieved much higher recall rate compared to EditPredict and RED-ML did when using the independent test set for evaluation (Fig. 2b), aligned well with the finding

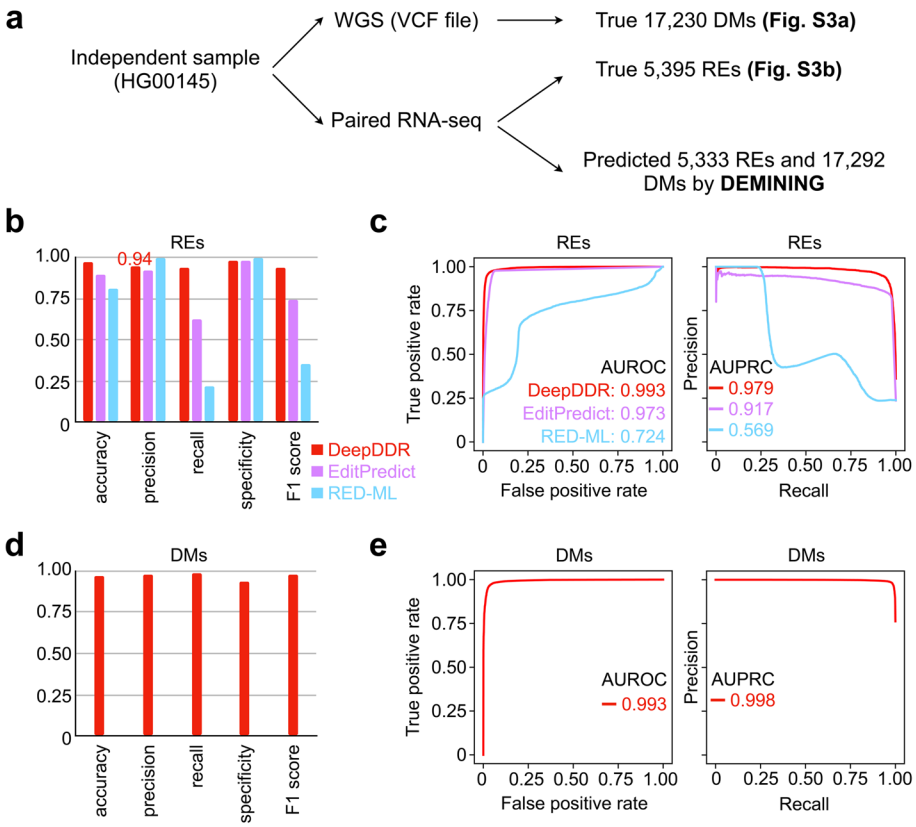

Fig. 2 Model performance on human independent test set. **a** Prediction of DMs and REs on independent test set with DEMINING/DeepDDR model. True DMs and REs were classified from paired WGS and RNA-seq data. DEMINING/DeepDDR was used to predict DMs and REs from training independent RNA-seq data of sample with SampleID HG00145. See the "Methods" section for details. **b** Evaluation metrics for RE identification on HG00145, including accuracy, precision, recall, specificity, and F1 score, comparing DeepDDR (red bar), EditPredict (purple bar) and RED-ML (blue bar). **c** Evaluation of different models on RE identification on HG00145. Receiver operating characteristic (ROC, left) curves and precision recall curves (PRC, right) of DeepDDR (red), EditPredict (purple), and RED-ML (blue) were shown to indicate their performance on RE identification. Area under ROC (AUROC) and area under PRC (AUPRC) values of DeepDDR (red), EditPredict (purple), and RED-ML (blue) were included in the figure. **d** Evaluation metrics for DM identification on HG00145, including accuracy, precision, recall, specificity, and F1 score, for DeepDDR. **e** Evaluation of DeepDDR on DM identification on HG00145. ROC (left) and PRC (right) of DeepDDR were shown to indicate its performance on DM identification. AUROC and AUPRC values of DeepDDR were included in the figure

when using the test set for evaluation (Additional file 1: Fig. S2d). Other than this, DeepDDR also performed well in terms of AUROC and AUPRC values for predicting DMs, while other pipelines were not applicable for DM identification from RNA-seq datasets (Fig. 2d and e).

To assess the specific features from the input CMC matrix that may contribute to the classification accuracy of DeepDDR, we conducted a gradient analysis [27, 28] to identify key sequences that may contribute to distinguish REs from DMs. Briefly, for a given input CMC matrix and the corresponding output of DeepDDR, a gradient analysis calculates a rectified derivative for each input neuron in a deep neural network by backpropagating the output. This derivative value then represents the importance of each neuron in determining the output. The average gradient values for all 5333 A-to-I REs and 3006 A-to-G (out of all 17,292) DMs were calculated. With gradient

heatmaps as shown in Additional file 1: Fig. S4a, it indicated that bases immediately upstream and downstream of REs or DMs were crucial for DeepDDR to differentiate REs from DMs. We then extracted 50 nt sequences surrounding all 5333 A-to-I REs and 3006 A-to-G DMs to explore possible sequence motifs crucial for RE and DM differentiation. As shown in Additional file 1: Fig. S4b, the 5' upstream nucleotide adjacent to REs suggested significant depletion of G, while the 3' downstream nucleotide suggested significant enrichment of G (top panel), which is consistent with previous reports of sequence features surrounding REs (bottom panel) [29–31]. In contrast, the 5' upstream nucleotide adjacent to DMs indicated a significant enrichment of C, while the 3' downstream nucleotide shows a significant enrichment of T (top panel of Additional file 1: Fig. S4c), mirroring the pattern of annotated A-to-G DMs (retrieved from dbSNP v151) located in mature mRNA regions (bottom panel of Additional file 1: Fig. S4c).

We also examined several other features that might contribute to distinguish REs and DMs, including repetitive element component, gene body location, clustering, and RNA secondary structure. As expected, REs were primarily located in *Alu* elements as previously reported [11–13], while DMs were predominantly found in non-repetitive regions (Additional file 1: Fig. S4d), consistent with previous studies [2]. In addition, REs were mainly located in the 3' UTR [32], whereas expressed DMs were found in both the 3' UTR and coding sequences (CDS) (Additional file 1: Fig. S4e) [2]. Moreover, detected REs tended to be closely clustered (mean distance between two REs is 9, Additional file 1: Fig. S4f), consistent with previous reports [14, 33], while expressed DMs are scattered (mean distance between two DMs is 576, Additional file 1: Fig. S4f). Finally, the minimum free energy (MFE) analyzed by RNAfold indicated that sequences surrounding REs exhibited lower MFE than those around expressed DMs, suggesting that RE-surrounding sequences tended to form stable dsRNA structures (Additional file 1: Fig. S4g) [30, 34]. Although these features could not be directly highlighted by the gradient heatmap (Additional file 1: Fig. S4a and S4b), they likely contributed to the DeepDDR model's ability to distinguish REs and DMs.

Taken together, we developed a useful computational tool, DEMINING, to directly predict and distinguish DMs and REs from mapped RNA-seq reads.

Extended application of DEMINING in non-primate RNA-seq datasets after transfer learning

Other than the successful application in human samples, we set to apply DEMINING trained with human data to identify DMs and REs from other non-primate datasets, which has been previously proven to be challenging [20] because characteristics and patterns of A-to-I REs in primates, such as human, and non-primates, such as mouse, are very different. Nevertheless, we first tried the original DEMINING to differentiate DMs and REs from a publicly available mouse bone marrow RNA-seq dataset (Methods, Fig. 3a, top). Because no corresponding WGS was available, true DMs and REs in this mouse dataset were obtained by comparing paired RNA-seq samples of wild-type (WT) and *Adar* knockout conditions (Additional file 1: Fig. S5a, Additional file 6: Table S5, and Methods) and then used to evaluate the performance of the original DeepDDR model in the mouse sample. Since the features used by RED-ML only designed to be extracted

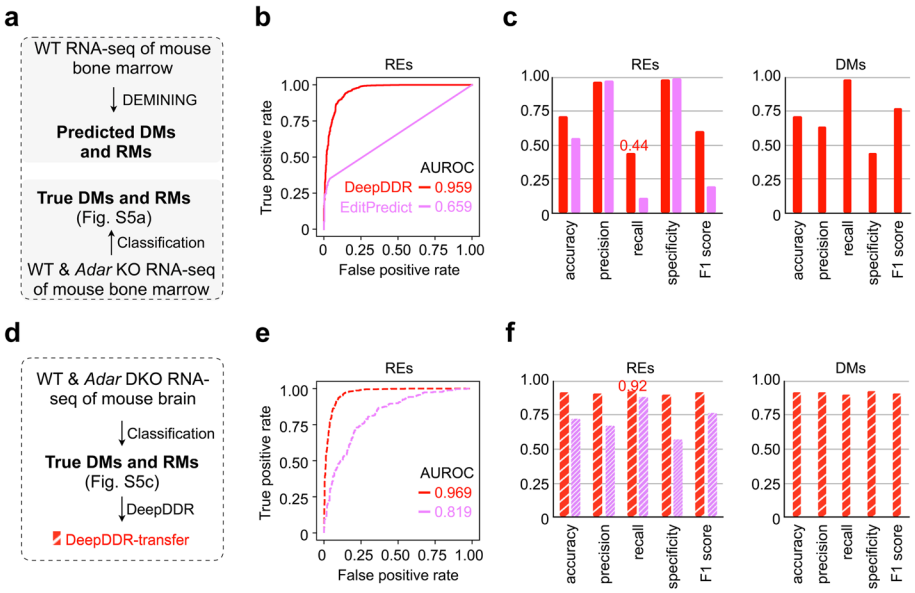

Fig. 3 Model trained on human data accurately classify DMs and REs on mouse data using transfer learning. **a** Prediction of DMs and REs in mouse datasets with original DeepDDR model. DeepDDR was used to predict DMs and REs from WT RNA-seq data of mouse bone marrow, and true DMs and REs were classified by comparing WT and *Adar* KO RNA-seq data. See the "Methods" section for details. **b** Evaluation of different models on RE identification. ROC of DeepDDR (red) and EditPredict (purple) were shown to indicate their performance on RE identification with the mouse bone marrow dataset. AUROC values of DeepDDR (red) and EditPredict (purple) were included in the figure. Of note, since the features used by RED-ML only designed to be extracted from the human genome, RED-ML was failed to be included in this and below comparisons. **c** Evaluation metrics for RE identification (left) and DM identification (right), including accuracy, precision, recall, specificity, and F1 score, by DeepDDR (red) and EditPredict (purple). **d** Schematic of constructing DeepDDR-transfer model. An additional mouse brain RNA-seq datasets containing WT and *Adars* (*Adar* and *Adarb1*) double knockout (DKO) samples were used for transfer learning. See the "Methods" section for details. **e** Evaluation of different models after transfer learning on RE identification. ROC of DeepDDR-transfer (red dashed line) and EditPredict-transfer (purple dashed line) were shown to indicate their performance on RE identification with the same mouse bone marrow dataset. AUROC values of DeepDDR-transfer (red) and EditPredict-transfer (purple) were included in the figure. **f** Evaluation metrics for RE identification (left) and DM identification (right), including accuracy, precision, recall, specificity, and F1 score, by DeepDDR-transfer (red shaded bar) and EditPredict-transfer (purple shaded bar)

from the human genome, RED-ML was failed to be included in this comparison, and only EditPredict could be applied here for the comparison with DeepDDR. As shown in Fig. 3b, although DeepDDR achieved higher AUROC value in the prediction of REs than EditPredict did, the RE recall rate by DeepDDR was only 0.44, when using the same cutoffs as in human (prediction probabilities of DMs ≥ 0.5 and prediction probabilities of REs < 0.5) (Fig. 3c). Further examination showed that the prediction probabilities of most mouse REs by DeepDDR were ranged between 0 and 0.8 (Additional file 1: Fig. S5b) compared to human ones between 0 and 0.5 (Additional file 1: Fig. S2a and Fig. S3c). In this case, the original DeepDDR model trained with human datasets might not be suitable for direct DM and RE classification in mouse datasets.

To solve this problem, we then leveraged the original DeepDDR model trained with human datasets as a pre-trained model for transfer learning with another mouse dataset containing mouse brain *Adars* (*Adar* and *Adarb1*) double knockout (DKO) and WT samples (Methods). True DMs and REs in these mouse brain datasets were similarly

retrieved by comparing ratio changes between WT and *Adar* DKO brain samples (Additional file 1: Fig. S5c, Additional file 7: Table S6, and Methods) and further used as inputs for the transfer learning to obtain a fine-tuned DeepDDR-transfer model (Fig. 3d and Additional file 1: Fig. S5d). Similarly, transfer learning was applied to EditPredict using its default requirement of 201-nucleotide sequences extracted around the same true DMs and REs from mouse brain datasets, resulting in the generation of EditPredict-transfer (Additional file 1: Fig. S5d). DeepDDR-transfer and EditPredict-transfer were thus used to reanalyze mouse bone marrow WT RNA-seq samples that were not used for transfer learning. As expected, DeepDDR-transfer also exhibited higher AUROC value compared to EditPredict-transfer (Fig. 3e). More importantly, the recall ratio of mouse REs was increased from 0.44 by the original DeepDDR model to 0.92 by the fine-tuned DeepDDR-transfer model (compared Fig. 3c with Fig. 3f), indicating that the transfer learning approach significantly improved the model performance. Closer examination showed that prediction probabilities of 90.78% mouse bone marrow A-to-I REs ranged between 0 and 0.05 by the fine-tuned DeepDDR-transfer model (Additional file 1: Fig. S5e), compared to those between 0 and 0.8 by the original DeepDDR model prior to transfer learning (Additional file 1: Fig. S5b).

Similar improvement was also achieved in a downloaded nematode dataset that contains *adr* (*adr-1* and *adr-2*) DKO and WT samples (Methods, Additional file 8: Table S7), in which the recall ratio was increased from 0.18 by DeepDDR to 0.85 by DeepDDR-transfer (Additional file 1: Fig. S6). These results together suggested the expanded application of DeepDDR-transfer in the prediction of RNA A-to-I REs from non-primate samples. Interestingly, when applied the fine-tuned DeepDDR-transfer model back to the human samples, the precision of RE prediction was significantly dropped from 0.94 by the original DeepDDR models (Fig. 2b) to 0.56 by DeepDDR-transfer (Additional file 1: Fig. S7a). Correspondingly, many true DMs from this human dataset were wrongly classified as REs by DeepDDR-transfer (Additional file 1: Fig. S7b). In this scenario, the DeepDDR and DeepDDR-transfer models should be individually used for the corresponding analyses in primate or non-primate samples.

During the study, we also developed other machine learning models, such as Light Gradient Boosting Machine (LightGBM), logistic regression (LR), and random forest (RF), and deep learning models, such as recurrent neural network (RNN) and a hybrid of CNN and RNN (CNN + RNN), for the analyses. Among all these models, DeepDDR with two layers of CNN and the CNN + RNN hybrid model demonstrated comparable performance in distinguishing differentiating DMs and REs in both human (Additional file 1: Fig. S8) and mouse datasets, before and after transfer learning (Additional file 1: Fig. S9-10), superior to those of all the other models.

Application of DEMINING to detect disease-associated DMs/REs from patient samples

Next, we set to apply the DeepDDR-embedded DEMINING framework to identify disease-related mutations from publicly available RNA-seq datasets of individuals, such as acute myeloid leukemia (AML) patient samples. From a published collection of 19 AML patient RNA-seq datasets (from peripheral blood or bone marrow tissue) [35], DEMINING totally identified 195,256 DMs (Fig. 4a, Additional file 9: Table S8) and 137,682 REs

(Additional file 10: Table S9). Since the heterogeneous nature of AML has been reported to be associated with numerous pathogenesis-related mutations [36], we then set to further investigate these DEMINING-identified mutations from AML samples.

Among 195,256 DMs, we first filtered out those could be also called in at least one of 17 normal control samples (medullary thymic or myelocytic precursor cells) in the same collection [35] or had a high minor allele frequency (MAF > 0.05) in 1000 Genomes Project (Fig. 4a). As a result, about 57,150 DMs were remained by this filtering step, indicating as AML-specific ones (Fig. 4a, Additional file 11: Table S10), with a similar distribution pattern of mutation types (Additional file 1: Fig. S11a) as those retrieved from 1000 Genomes Project (Additional file 1: Fig. S1a). Although the majority (~ 70.91%) of them were found to be overlapped with annotated SNPs in dbSNP (Fig. 4b), only a small portion (~ 13.32%) of these 57,150 AML-specific DMs were previously reported to be associated with diseases by ClinVar and/or COSMIC databases. These results thus implied that many previously underappreciated DMs might be associated with AML pathogenesis. Given the fact that their mutation frequencies were generally lower than those reported in the 1000 Genomes Project (compared Fig. 4c with the left panel of Additional file 1: Fig. S1c), it suggested that AML-specific DMs identified by DEMINING were likely acquired mutations during AML pathogenesis.

Correlation of DEMINING-identified DMs and mis-regulated gene expression

Then, we performed additional analyses to address links of these previously underappreciated DMs with AML. Among 57,150 AML-specific DMs identified by

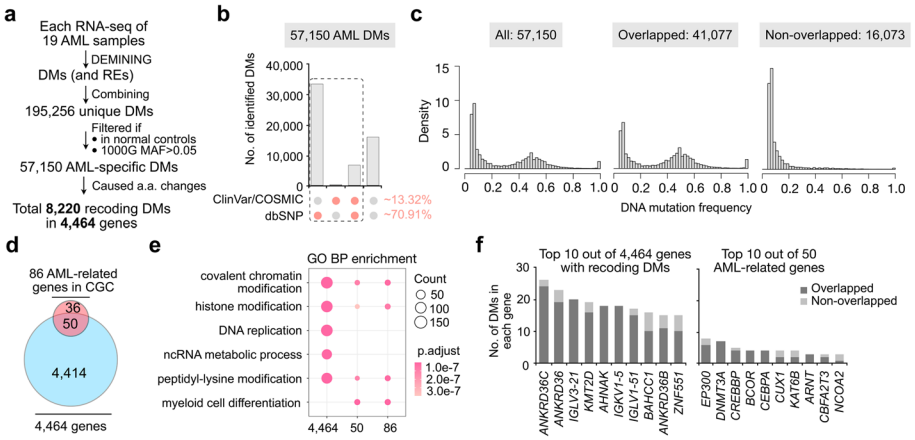

Fig. 4 Applying DEMINING framework to identify disease-related mutations in acute myeloid leukemia (AML). **a** Identification of AML-associated DMs from corresponding RNA-seq datasets by DEMINING. **b** Overlapping of AML-specific DMs and reported SNVs in public databases, including ClinVar 2023.04 (https://www.ncbi.nlm.nih.gov/clinvar/), COSMIC (version 97, https://cancer.sanger.ac.uk/cosmic/) and dbSNP (version 156, https://www.ncbi.nlm.nih.gov/snp/). **c** Mutation frequency distribution of all AML-specific DMs (left), overlapped AML-specific DMs (middle), and non-overlapped AML-specific DMs (right). **d** Overlapping of 4464 mutated genes carrying AML-specific recoding DMs with 50 AML-associated genes listed in COSMIC Cancer Gene Consensus (CGC). **e** Gene Ontology (GO) enrichment analysis in biological process (BP) terms for three gene sets including all mutated 4464 genes, 86 AML-associated genes listed in CGC, and their overlapping 50 genes. Top GO terms ordered by adjusted *P* value in at least one gene set were kept and compared. **f** The number of overlapped (dark gray) and non-overlapped (light gray) DMs in the top 10 genes. Left, top 10 genes out of 4464 genes with recoding DMs; right, top 10 genes out of 50 AML-associated genes listed in COSMIC CGC

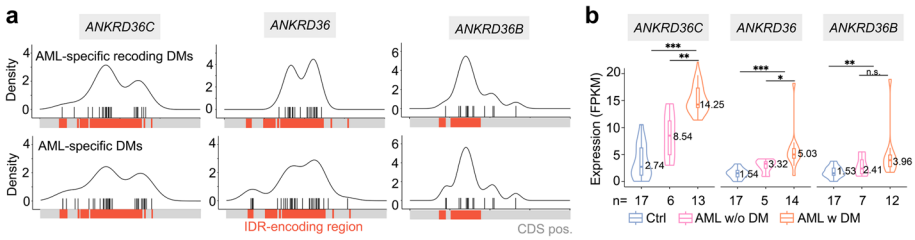

Fig. 5 AML-specific DMs identified in three *ANKRD* genes are enriched in IDR coding regions and correlated with expression. **a** Distribution of identified AML-specific recoding DMs (top) and AML-specific DMs (bottom) along coding sequences (CDS) of the *ANKRD36C* (left), *ANKRD36* (middle), and *ANKRD36B* (right). Distribution (black plot) of predicted DMs (black vertical lines) by DEMINING were shown in genes' CDS regions (gray and red rectangles). Red rectangles represent CDS regions that encoding intrinsically disordered regions (IDRs) predicted by MobiDB-lite integrated in the InterPro database (https://www.ebi.ac.uk/interpro/). **b** Comparison of gene expression of *ANKRD36C*,*ANKRD36*, and *ANKRD36B* in 17 normal control samples (Ctrl) and 19 AML patients with or without recoding DMs (AML w DM: AML patients with recoding DM; AML w/o DM: AML patients without recoding DM). The boxplot summarizes results for all samples with the number of samples *n* shown below. Center line: median. Box bottom and top edges: 25th and 75th percentiles. Whiskers extend to extreme points excluding outliers (1.5 times above or below the interquartile range). Outliers omitted for clarity. Violin-shaped areas: Kernel density estimate of data distribution. Statistical significance was assessed with two-tailed Wilcoxon rank-sum test. **P* < 0.05, ***P* < 0.01, ****P* < 0.001

DEMINING, 8220 were predicted to cause amino acid changes, hereafter called recoding DMs. These recoding DMs were found in 4464 genes (Fig. 4d), including 50 out of 86 AML-associated genes listed in the COSMIC Cancer Gene Consensus (CGC) that have evidence-based functions in the development and progression of AML [37]. Importantly, similar biological pathways (BPs) were enriched by Gene Ontology (GO) analyses of 4464 genes with AML-specific recoding DMs, 86 AML-associated genes listed in CGC, and their overlapping 50 genes (Fig. 4e). These enriched GO pathways included covalent chromatin modification, histone modification and peptidyl-lysine modification (Fig. 4e), which were also found to be mis-regulated during tumorigenesis in previous studies [38].

After analyzing the distribution of these recoding DMs in their host genes, we observed that over 15 recoding DMs were enriched in each of top 10 of aforementioned 4464 genes (left panel of Fig. 4f, Additional file 1: Fig. S11b), compared to less than 10 recoding DMs in AML-associated genes reported by CGC (right panel of Fig. 4f). Top 10 genes with most enriched recoding DMs included *ANKRD36C* (26 recoding DMs), *ANKRD36* (23 recoding DMs), and *ANKRD36B* (15 recoding DMs), which all belong to the ankyrin repeat domain (*ANKRD*) gene family.

Interestingly, these recoding DMs and all AML-specific DMs identified by DEMINING in the three *ANKRD* gene loci were mainly clustered in predicted sequences encoding intrinsically disordered regions (IDRs) (Fig. 5a). Previous studies showed that some mutations mapped to sequences encoding predicted IDRs were related to diseases [39]. Since IDRs could drive liquid-liquid phase separation (LLPS) [40] and the dysregulation of LLPS was a key event in the initiation and/or evolution of cancer [41], the identification of recoding DMs in regions encoding predicted IDRs of *ANKRD* gene loci suggested the interplay of dysregulated LLPS by mutated *ANKRD* genes along AML pathogenesis. Together with a recent finding of *ANKRD36* as a

biomarker of chronic myeloid leukemia (CML) [42], these results strongly suggested the possible association of enriched recoding DMs in *ANKRD* genes with AML.

What was the consequence of these recoding DMs on their host gene expression? We observed all three *ANKRD* genes were significantly upregulated in AML samples, especially in those with DEMINING-identified recoding DMs (Fig. 5b). These results indicated that DEMINING-identified DMs might be functional as *cis*-expression quantitative trait loci (*cis*-eQTLs), while we could not exclude the possibility of indirect effects of DMs on their host gene expression.

The cluster of 8220 AML-specific recoding DMs was selected due to their capacity of causing amino acid changes (Fig. 4a). Although lack of the whole-proteome sequencing datasets, corresponding liquid chromatography-coupled mass spectrometry (LC–MS/MS) spectra of major histocompatibility complex (MHC)-associated peptides from 19 AML patients were available in the same study [35]. Thus, it was attractive to examine whether 8220 DEMINING-identified AML-specific recoding DMs (and recoding 149 REs, Additional file 1: Fig. S12, Additional file 11: Table S10 and Additional file 12: Table S11) could cause the production of neoantigen(s) that were displayed by MHC. To achieve this goal, we created an in silico library containing 3,581,675 of predicted peptides (8-30 amino acids (a.a.) long) from 4464 genes with 8220 recoding DMs (Fig. 4a) and 121 genes with 149 recoding REs (Additional file 1: Fig. S12) as queries (Additional file 1: Fig. S13a, Methods). Decoyed peptides from the predicted peptides and proteins from human UniProt proteome were also added to this library to be used as negative controls (Additional file 1: Fig. S13a, Methods). This in silico library containing both predicted peptides with recoding DMs/REs and negative control peptides was used as bait to screen LC-MS/MS spectra of MHC-associated peptides from AML patients. This analysis led to the identification of three neoantigen candidates, two caused by distinct DMs and one caused by an RE (Additional file 1: Fig. S13b). Of note, none of these three neoantigens were listed in the published analysis of LC-MS/MS spectra [35], indicating the importance of DM/RE-guided strategy for neoantigen identification.

Lastly, it has been reported that DMs or REs can alter RNA splicing or RNA stability, which is correlated with human diseases [43, 44]. To evaluate whether DMs and REs identified by DEMINING could influence splicing or generate stop codons that trigger nonsense-mediated decay (NMD), we adopted the plugin MaxEntScan of Ensembl Variant Effect Predictor (VEP) for analyses [45]. As shown in Additional file 1: Fig. S14a and S14b, some DEMINING-identified DMs and REs were shown to affect splicing or introduce stop codon. Specifically, among 57,150 AML-specific DMs found in 19 AML samples, 22 were predicted to generate new splicing sites and 41 to abolish splicing sites. Of note, an obvious splicing pattern change could be determined with an AML-specific DM in the 08H053 sample, when compared to the 07H122 sample without the AML-specific DM (Additional file 1: Fig. S14c). Interestingly, among 58,045 AML-specific REs, only 13 were shown to abolish splice sites. As for whether AML-specific DMs or REs could generate stop codons for possible NMD, we found that 327 DMs caused stop codon gaining, while none with REs. These results suggested that DEMINING-identified DMs might dysregulate gene expression by altering RNA splicing and stability, which requires further investigation. Notably, since only high-confidence DMs and REs were selected with stringent cutoffs (such as mutation reads ≥ 2 and mutation frequencies ≥ 0.05, Methods),

some low abundant ones that could affect RNA splicing and stability might be missed in this analysis.

Discussion

Here, we reported a deep learning model embedded computational framework, DEMINING, to achieve efficient and accurate prediction of REs and DMs from RNA-seq datasets only (Figs. 1 and 2). After transfer learning, DEMINING-transfer has exhibited remarkable adaptability in datasets from non-human species, including mouse and nematode (Fig. 3), indicating its broad application in DM and RE profiling across species. The successful identification of DMs from the independent dataset (Fig. 2) and AML patient samples (Figs. 4 and 5) suggested its application in the disease-related DM profiling.

Identifying disease-related DMs is essential to our understanding of genetic disorders contributing to diseases. Despite that a large number of WGS and WES datasets have been accumulated to contribute to the profiling of genetic DMs [2, 46], retrieving high-confidence mutations from WGS and WES datasets still suffered with high cost, low efficiency, and time-consuming. Instead, whole transcriptomic RNA-seq, but not corresponding WGS, datasets are prevalently available from individuals, including patients [47]. Thus, it is luring to use the developed DEMINING framework to directly differentiate expressed DMs and REs from RNA-seq datasets.

In this proof-of-concept study, we applied DEMINING to RNA-seq datasets from AML patient samples [35] and uncovered previously unreported DMs and REs in some uncharacterized AML-associated gene loci, such as *ANKRD36C*, *ANKRD36*, and *ANKRD36B* (Fig. 4f). On the one hand, many of recoding DMs are enriched in these *ANKRD* gene regions encoding IDRs (Fig. 5a), thus suggesting a possible link between dysregulated LLPS and AML pathogenesis. On the other hand, a positive correlation between DEMINING-identified recoding DMs and upregulated expression of these *ANKRD* mRNAs has been observed, providing potential targets for AML diagnosis and/or therapeutics (Fig. 5b). In addition, some AML-specific recoding mutations, including two DMs and one RE, were found to be responsible for the production of neoantigens validated by corresponding mass spectrometry data (Additional file 1: Fig. S13). Finally, some AML-specific DMs and REs were also found to cause splicing changes or stop codon gaining (Additional file 1: Fig. S14), while their effects on RNA stability and function require further investigation. These findings together demonstrated the successful application of DEMINING in identification of disease-associated mutations, providing additional insights in pathogenesis and therapeutics.

Conclusions

Here, we developed a computational framework, DEMINING, empowered by an embedded DeepDDR model to differentiate DMs and REs directly from RNA-seq datasets. After transfer learning, the DeepDDR model can be extended for DM and RE differentiation from other non-primate samples. Given the fact that DEMINING is characteristic to efficiently and specifically detect DMs from RNA-seq datasets,

its application in AML samples successfully identified previously underappreciated mutations in unreported AML-associated genes. Since whole transcriptomic RNA-seq datasets are prevalently available from individuals, including patients, we expect that, when applied in a broader spectrum of human disease RNA-seq samples, DEMINING will uncover more disease-related mutations and genes for potential diagnosis and therapeutics.

Of note, since both expression (≥ 3 HPB) and mutation frequency (≥ 0.05 cutoffs), together with mutation reads (≥ 2), were used for the filtering, lowly expressed DM/RE sites (< 3 HPB) or sites with low mutation frequency (< 0.05) or with low mutation read number (< 2) were excluded by DEMINING. Thus, caution should be exercised when using samples with different sequencing strategies/depths for similar comparative analyses by DEMINING.

Methods

Collection of published WGS and/or RNA-seq datasets for analyses

A series of WGS and RNA-seq datasets from independent resources were collected and used in this study (Additional file 2: Table S1).

The first data collection consisted of data from 403 donors, which represents the intersection of individuals from the 1000 Genomes Project [2] phase 1, phase 3, and the Geuvadis consortium [24]. For each donor, data from whole-genome sequencing (WGS) and/or whole-exome sequencing (WES) were obtained in the form of variant call format (VCF) files. These VCF files were obtained from both phase 1 (1000 Genomes Project phase 1: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521) and phase 3 (1000 Genomes Project phase 3: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502) of the 1000 Genomes Project and were merged into a single file if they were from one donor. Additionally, RNA-seq data in FASTQ file format were obtained from the Geuvadis consortium for the same set of 403 donors. These data were used to construct DNA mutation (DM) and RNA editing site (RE) sets for deep learning model training and evaluation.

The second data collection consisted of an additional donor with SampleID HG00145, who was exclusively included in the 1000 Genomes Project phase 3 and the Geuvadis consortium. The genetic mutations obtained from both WGS and WES for this donor were merged into a single VCF file from the 1000 Genomes Project phase 3. Furthermore, RNA-seq data in FASTQ file format for donor HG00145 were obtained from the Geuvadis consortium. These data were used to construct DM and RE sets, which were used as an independent test set for deep learning model evaluation.

The third data collection consisted of mouse bone marrow RNA-seq of three *Adar1* knockout and three wild-type (WT) samples [48], mouse brain RNA-seq of three *Adar1* and *Adar2* knockout and three WT samples [49], and nematode RNA-seq of four *adr-1* and *adr-2* knockout and four WT samples [50]. These data were used to construct DeepDDR-transfer model and identify DMs and REs from other non-primate by DEMINING framework.

The fourth data collection consisted of RNA-seq datasets of 19 acute myeloid leukemia (AML) patients (peripheral blood or bone marrow tissue) and 17 normal control samples (medullary thymic or myelocytic precursor cells). These RNA-seq datasets were

available from the Gene Expression Omnibus (GEO) with accession number GSE147524 and GSE98310 [35], respectively. These data were used to identify disease-related mutations in AML by DeepDDR-embedded DEMINING framework.

Detailed information of all these WGS and/or RNA-seq datasets was listed in Additional file 2: Table S1.

Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline

The previously published RADAR pipeline [25] was applied to identify high-confidence REs from each of 403 RNA-seq datasets, including two major steps.

Step 1: RNA-seq read mapping/piling

After performing a quality check using FastQC (version 0.11.4) with default parameters, RNA-seq reads were trimmed by Trimmomatic (version 0.36) with the following parameters: TruSeq3-PE-2.fa:2:30:10 TRAILING:25 MINLEN:30. Reads mapped to ribosomal DNA (rDNA) by BWA-MEM algorithm (version 0.7.9a) with default parameters were then removed. To enhance the detection of mutations from reads with more mismatches, a two-round unique mapping strategy [14] was implemented to align high-quality RNA-seq reads to the human hg38 reference genome with the GENCODE gene annotation (version 41) by HISAT2 (version 2.1.0) with parameters: –rna-strandness RF –no-mixed –secondary –no-temp-splicesite –known-splicesite-infile –no-soft-clip –score-min L,-16,0 –mp 7,7 –rfg 0,7 –rdg 0,7 –max-seeds 20 –k 10 –dta and BWA-MEM (version 0.7.9a) with default parameters. Uniquely mapped RNA-seq reads by HISAT2 and BWA-MEM with up to six mismatches were selected and combined for subsequent analysis. The combined reads were analyzed to identify read pairs that were likely derived from duplicates of the same original DNA fragments due to artifactual processes.

As these duplicates were considered non-independent observations, we employed Picard MarkDuplicates (version 2.7.1) with the following parameters: CREATE_INDEX = true and VALIDATION_STRINGENCY = SILENT. This tool tagged all but a single read pair within each set of duplicates, causing the marked pairs to be ignored during subsequent processing steps.

To address issues arising from RNA-seq reads containing Ns in their cigar string, we utilized the GATK (version 4.1.2.0) command SplitNCigarReads with default parameters. This command split reads that contain Ns in their cigar string, typically associated with spanning splicing events in RNA-seq data. The process involved identifying all N cigar elements and generating $k+1$ new reads, where k represents the number of N cigar elements detected. The first read of generated $k+1$ new reads included the bases that located to the left of the first N element, while the part of the read that located to the right of the N (including the Ns) was hard clipped and so on for the rest of the new reads.

Base quality scores were crucial for assessing the reliability of base calls and were used to weigh the evidence for or against potential mutations during the mutation discovery process. Systematic biases that affected base quality scores could arise from various factors during the generation of RNA-seq data such as library preparation, sequencing, chip manufacturing defects, or sequencer instrumentation defects. To address systematic biases, we applied a base quality score recalibration procedure using two GATK commands: BaseRecalibrator and ApplyBQSR, with default parameters. During the

recalibration process, covariate measurements from all uniquely mapped reads were collected, including read group, base quality score, machine cycle producing the base (*N*th cycle = *N*th base from the start of the read), and the current base in conjunction with the previous base (dinucleotide). These covariate measurements were used to build a model that captured the systematic biases present in the data. Based on this model, base quality adjustments were applied to the dataset to correct for the observed biases.

Mutation sites were determined from the BAM file containing uniquely mapped reads by using the Samtools (version 1.9) command mpileup with parameters: -Q 0. High-confidence mutations with base quality ≥ 20, mutation reads ≥ 2, hits per billion mapped bases (HPB) ≥ 3, and mutation frequencies ≥ 0.05 were selected for subsequent analyses.

Step 2: Identification of true REs

To identify high-confidence REs, expressed mutation sites (with base quality ≥ 20, mutation reads ≥ 2, HPB ≥ 3, and mutation frequencies ≥ 0.05) which were overlapped with DMs listed by 1000 Genomes Project were first removed, and non-overlapped ones were then classified into *Alu* or non-*Alu* regions according to the genomic location as previously described [12].

Next, mutations in non-*Alu* regions were subjected to additional stringent filtering processes to remove false positives, including:

1) Those overlapped with SNPs from dbSNP version 151 (https://www.ncbi.nlm.nih.gov/SNP/), the 1000 Genomes Project (https://www.internationalgenome.org/), or the University of Washington Exome Sequencing Project (https://evs.gs.washington.edu/EVS/).
2) Those in simple repeats, homopolymer runs of ≥ 5 base pairs.
3) Intronic candidates if they were located within 4 base pairs of a known splice junction.
4) Those mapped to highly similar regions by BLAST-like alignment tools (BLAT [51], version 364, parameters: -repMatch = 2253 -stepSize = 5).

REs in non-*Alu* regions after this stringent selection were combined with those in *Alu* regions to obtain high-confidence REs in each RNA-seq dataset. Of note, over 98% of identified REs in each of 403 RNA-seq datasets were A-to-G and T-to-C mismatches, mainly in *Alu* regions, indicating high-confidence A-to-I [11] called by RADAR [25] (Additional file 1: Fig. S1b). A total of 122,872 unique A-to-I sites were finally determined from the 403 RNA-seq datasets as the RE set, which is divided into training, validation, and test sets using an 8:1:1 ratio for the development and evaluation of DeepDDR (Additional file 1: Fig. S1d).

Detailed information of these REs was listed in Additional file 3: Table S2.

Identification of true DMs from VCF files of 403 paired human WGS datasets

True DMs were directly obtained from about 31 million SNPs documented in VCF files from the 1000 Genomes Project [2] (see the "Collection of published WGS and/or RNA-seq datasets for analyses" section for detail). Expressed DMs were then determined with base quality ≥ 20, mutation reads ≥ 2, HPB ≥ 3, and mutation frequencies ≥ 0.05 from

their corresponding human RNA-seq datasets, as described in "Step 1: RNA-seq read mapping/piling" of the "Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline" section. In total, from approximately 578,000 expressed DMs, a subset of 122,872 expressed DMs was randomly selected and further divided into training, validation, and test sets using an 8:1:1 ratio for the development and evaluation of DeepDDR (Additional file 1: Fig. S1d).

Detailed information of these 122,872 expressed DMs was listed in Additional file 4: Table S3.

Construction of a deep learning model for the differentiation of expressed DMs from REs directly from aligned RNA-seq reads

To differentiate expressed DMs from REs directly from aligned RNA-seq reads, a deep learning neural network, called DeepDDR, was trained with true DM and RE sets identified from 403 paired human WGS and RNA-seq datasets as described above in the "Identification of true DMs from VCF files of 403 paired human WGS datasets" and "Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline" sections, respectively.

To determine the most suitable sequence length, we tested a range of lengths, including 12, 24, 50, 100, and 200 nts. The sequence length of 50 nts provided the best performance in terms of area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and F1 score (data not shown). Thus, we chose 50 context bases surrounding the mutation sites for the subsequent model training.

To train DeepDDR, the *c*o-occurrence frequencies of each *m*utation site with its *c*ontext bases (CMC) were determined by analyzing 51-nucleotide sequences containing the mutation in the middle and its 50 context bases (25 bases in the upstream region and 25 bases in the downstream region). Briefly, 51-nucleotide sequences with the mutation in the middle were extracted from aligned RNA-seq reads as described in "Step 1: RNA-seq read mapping/piling" of the "Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline" section. A 16×50 CMC matrix, indicating frequencies of all 16 types of dinucleotides by 50 combinations of the given DM/RE individually with each of 25 upstream and 25 downstream bases, was obtained from piled 51-bp sequencing, shown in Fig. 1b. CMC matrices of all DMs and REs in training set were then used as the input sets for the DeepDDR model training. Of note, one DM or RE can be identified in multiple samples within the 403 human RNA-seq dataset; if so, only one sample is randomly selected for constructing the CMC matrix to mitigate the risk of overfitting.

The DeepDDR model consists of two layers of convolutional neural networks (CNNs) implemented with the TensorFlow v2.0.0 backend in Python v3.7.8. Using CMC matrices as input, the CNN module captures the specific co-occurrence frequencies around DMs or REs. The model architecture consists of two convolutional layers with rectified linear unit (ReLU) activation function, followed by a max pooling layer. The hyperparameters of the model were optimized through a search process. Specifically, we explored different values for the number of filters in the first convolutional layer, choosing from options such as 32, 64, and 128. Additionally, we varied the number of units in the dense layer, selecting values such as 256, 512, and 1024.

During the optimization process, we evaluated the models based on their AUROC (area under the receiver operating characteristic curve) performance on the validation set, as described in "Step 1: RNA-seq read mapping/piling" of the "Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline" section. The best-performing model had the following configuration: the first convolutional layer had 128 filters with a width of 12 bases and 4 channels. The second convolutional layer had 64 filters, covering the 128 output channels from the first layer, with filters of width 6 bases. The max pooling layer subsampled the signal with a stride of 4. The resulting signal was then passed to a dense layer with 1024 hidden units, incorporating a dropout ratio of 0.3 to prevent overfitting. The dense layer was connected to two output nodes using the softmax activation function, which allowed for the prediction of probabilities for the two classes: DM or RE. To make a prediction, DeepDDR selected the class with the highest prediction probability as the predicted class. Mutations with prediction probabilities ≥ 0.5 were considered as DMs, mutations with prediction probabilities < 0.5 were considered as REs.

The DeepDDR model was trained by the CMC matrices of DMs and REs with a batch size of 5000 and 100 epochs. Early stopping was employed with a patience of ten rounds, terminating the training process if no improvement in performance was observed. The model with best performance during training was kept at last.

Construction of the DEMINING framework

DEMINING is a stepwise computational framework to classify DMs and REs directly from RNA-seq data, including two major steps, RNA-seq read mapping/piling to call high confidence mutations, which are further classified as DMs and REs by DeepDDR. Briefly, each RNA-seq dataset was first aligned and piled up as described in "Step 1: RNA-seq read mapping/piling" of the "Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline" section. After this step, high-confidence mutations were extracted with base quality ≥ 20, mutation reads ≥ 2, HPB ≥ 3, and mutation frequencies ≥ 0.05. Next, these identified high-confidence mutations were then fed into the DeepDDR model for DM and RE classification as described in the "Construction of a deep learning model for the differentiation of expressed DMs from REs directly from aligned RNA-seq reads" section. Basically, DeepDDR analyzed the CMC of each input mutation and assigned it a predicted probability score. Mutations with prediction probabilities ≥ 0.5 were considered as DMs, and mutations with prediction probabilities < 0.5 were considered as REs.

True DMs and REs obtained from HG00145 as an independent test set

The same RADAR pipeline [25] was applied to identify 5395 REs from an independent donor with SampleID HG00145. With its paired WGS (see the "Collection of published WGS and/or RNA-seq datasets for analyses" section for detail), 17,230 expressed DMs were also retrieved from 1000 Genomes Project and detected in the RNA-seq dataset with base quality ≥ 20, mutation reads ≥ 2, HPB ≥ 3, and mutation

frequencies ≥ 0.05. These DMs and REs were used as independent sets for deep learning model validation and comparison.

Detailed information of these 5395 REs and 17,230 DMs were listed in Additional file 5: Table S4.

Prediction of REs by EditPredict and RED-ML

For EditPredict, the scripts "get_seq.py" and "editPredict.py" were obtained from the EditPredict GitHub repository (https://github.com/wjd198605/EditPredict). The "get_seq.py" script was used to extract the flanking 200 bp sequence of each mutation from the hg38 reference genome. This was done by running the script with the following parameter: "python get_seq.py -f hg38.fa -m b -l 200". The extracted sequences were then fed to the "editPredict.py" script, which was used to predict RNA editing sites in the input sequences. The EditPredict weights and construction files were obtained from the same repository.

For RED-ML, the RED-ML executable commands "MismatchStat", "MutDetML", and the script "red_ML.pl" were obtained from the RED-ML GitHub repository (https://github.com/BGIRED/RED-ML). The features used by RED-ML for each mutation were extracted using the "MutDetML" command. This was done with the following parameters: "MismatchStat -i $RNA_BAM -o stat.txt -u -q 20" and "MutDetML -i $RNA_BAM -r hg38.fa -v stat.txt -u -q 20 -Q 0 -d 0 -a 0 -t -o Mut.txt.gz". The extracted features were then used as input to the "red_ML.pl" script to predict A-to-I editing. Of note, RED-ML was only designed to extract features from the human genome for human A-to-I prediction.

Gene expression analysis

Gene expression of AML RNA-seq samples was determined by StringTie (version 2.1.4) with default parameters on BAM files after RNA-seq read mapping, as previously described [52].

Gene ontology analysis

After annotating mutations using the ANNOVAR [53] (version 7 June 2020), GO analyses of three gene sets, including all 4464 genes containing 8220 recoding DMs, 86 AML-associated genes listed in CGC, and their overlapped 50 genes, were performed by using clusterProfiler enrichGO function (ont = "BP", pAdjustMethod = "BH", pvalueCutoff = 0.2, qvalueCutoff = 0.2) and reduced redundancy of GO terms by simplify function (cutoff = 0.7, by = "p.adjust", select_fun = min, measure = "Wang"). Top GO terms ordered by adjusted *P* value in each gene set were kept and compared (Fig. 4e).

Mass spectrometric data analysis

Raw MS data of major histocompatibility complex (MHC) associated peptides from 19 AML patients were available in a public repository from the ProteomeXchange Consortium via the PRIDE partner repository by the dataset identifier PXD018542 [35]. To analyze these raw MS data, we constructed an in silico library containing predicted peptides ranging from 8 to 30 amino acids in length from 4464 genes with

recoding DMs and 121 genes with recoding REs. Decoyed peptides from the predicted peptides and proteins from human reference proteome (UniProt 2023-04-04) were also added to this library to be used as negative controls. Subsequently, the in silico library served as a reference database for searching LC-MS/MS spectra. Raw MS data of each sample was transformed into mzML format using MSConvert [54] (version 3.0.23090). The mzML format of MS data was then searched against the customized database using Comet [55] (version 201,901 rev. 1). The search was performed with 10 ppm precursor mass tolerance and 0.02 fragment ions mass tolerance. The search results were analyzed by PeptideProphet [56] (version 5.2.1). To ensure a reliable peptide identification, the search results were filtered to achieve an estimated peptide FDR of 5%. Only mutation-containing peptides that did not occur in the reference proteome were included in the final list. The search and filtering processes were implemented by the Philosopher (version 4.8.1) framework [57]. The detailed information of the search results was listed in Additional file 1: Fig. S13b.

Identification of true DMs and REs from paired adenosine deaminase(s) knockout and wild-type samples

To identify true DMs and REs from adenosine deaminase(s) knockout (KO) and wild-type (WT) data, the following steps were performed:

> Step 1: RNA-seq data from the adenosine deaminase(s) KO and WT samples were mapped to the corresponding reference genomes (mouse: mm10, nematode: ce11) using the same method described in "Step 1: RNA-seq read mapping/piling" of the "Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline" section.
>
> Step 2: Mutations were determined from the BAM file containing uniquely mapped reads by using the Samtools (version 1.9) command mpileup with parameters: -Q 0. High-confidence mutations with base quality ≥ 20, mutation reads ≥ 2, HPB ≥ 3, and mutation frequencies ≥ 0.05 were selected for further analyses.

Step 3: Calculating mutation frequency ratio

Firstly, calculate the average mutation frequency in the wild-type (WT) samples by summing up the mutation frequencies of all the WT samples and dividing by the total number of WT samples. Secondly, calculate the average mutation frequency in the adenosine deaminase(s) knockout (KO) samples by summing up the mutation frequencies of all the KO samples and dividing by the total number of KO samples. Thirdly, subtract the average mutation frequency of the WT samples from the average mutation frequency of the KO samples.

Mutation frequency difference = Average mutation frequency in KO samples − Average mutation frequency in WT samples.

Lastly, divide the difference obtained in the last step by the average mutation frequency in the WT samples.

Mutation frequency ratio = Mutation frequency difference / Average mutation frequency in WT samples.

Step 4: Classification of true DMs and REs

True DMs were selected by criteria: (1) the absolute value of mutation frequency ratio < 0.2, (2) the mutation frequencies in WT samples ≥ 0.05, and (3) not overlapped with sites in the REDIportal database [58].

Of note, as the REDIportal did not contain nematodes, true DMs from nematode dataset [50] were only selected by the absolute value of mutation frequency ratio < 0.2 and the mutation frequencies in WT samples ≥ 0.05.

True REs were selected by criteria: (1) the mutation frequency ratio was lower than -0.2; (2) the mutation frequencies in WT samples were greater than 0.05 and lower than 0.5; (3) the mutation frequencies in adenosine deaminase(s) KO samples were lower than 0.05; (4) the mutation types were A-to-G or T-to-C, indicating an A-to-I editing event; and (5) the mutations did not overlap with SNPs in the dbSNP database (dbSNP version for mouse: 150, dbSNP version for nematode: 138).

Detailed information of these REs and DMs were listed in Additional file 6: Table S5 (mouse bone marrow dataset), Additional file 7: Table S6 (mouse brain dataset), and Additional file 8: Table S7 (nematode dataset).

Fine-tuning of DeepDDR by transfer learning

To perform fine-tuning and validation of DeepDDR on mouse bone marrow datasets using transfer learning, the following steps were taken:

Step 1: Obtain true DMs and REs from the mouse bone marrow datasets using the method described in the "Identification of true DMs and REs from paired adenosine deaminase(s) knockout and wild-type samples" section.
Step 2: Split the DMs and REs individually into transfer training, transfer validation, and transfer test sets with a ratio of 70:15:15.
Step 3: Sequentially select transfer training sets of different sizes from the original 70% split transfer training set. The aim is to find the smallest transfer training set size that maximizes the recall ratio. Evaluate the recall ratio and other evaluation metrics on the transfer validation sets to determine the optimal size.
Step 4: Based on the evaluation results, select the subset of 500 DMs and 500 REs as the final transfer training set.
Step 5: Construct the DeepDDR-transfer model using a transfer learning strategy. Start with the DeepDDR model trained on human datasets as described in the "Construction of a deep learning model for the differentiation of expressed DMs from REs directly from aligned RNA-seq reads" section. Fine-tune this model using the final transfer training set from the mouse bone marrow datasets. This process was implemented with the TensorFlow v2.0.0 backend in Python v3.7.8.
Step 6: Test the performance of the DeepDDR-transfer model on the transfer test set.

Construction of other deep learning models

To train other deep learning-based models for comparison, a RNN and a hybrid model of CNN and RNN (CNN + RNN) were constructed using the TensorFlow (version 2.0.0) backend in Python (version 3.7.8).

For the RNN model, it consisted of one bidirectional LSTM layer and one dense layer. The hyperparameters that were varied during training include the number of hidden units in the bidirectional LSTM layer, such as 16, 32, or 64, and the number of units in the dense layer, such as 256, 512, or 1024.

For the CNN + RNN model, it combined two convolutional layers, one bidirectional LSTM layer and one dense layer. The hyperparameters that were varied include the number of filters in the first convolutional layer, such as 32, 64, or 128, and the number of hidden units in the bidirectional LSTM layer, such as 16, 32, or 64.

During training, the models were fed with the CMC matrix as input. The performance of each model was evaluated based on the AUROC (area under the receiver operating characteristic curve) on the validation set. The model that achieved the highest AUROC on the validation set was chosen for further comparison.

Construction of machine learning models

To train the machine learning models, including LightGBM, LR, and RF, four features were extracted from the mutations in the training set, as described in "Step 1: RNA-seq read mapping/piling" of the "Identification of true REs from 403 human RNA-seq datasets by the RADAR pipeline" section. These features include mutation frequency, base substitution type, whether the mutation is located in the *Alu* region, and the count of base pairs from the closest mutation in the genomic sequence. These features were selected based on insights obtained from previous bioinformatic tools and the current understanding of RNA editing mechanisms.

For the LightGBM model, the LightGBM Python package (version 3.1.1.99) was used. For LR and RF models, the Python scikit-learn package (version 0.20.3) was employed. The extracted four features were used as inputs for these models.

To optimize the hyperparameters of each model, 150 combinations were selected for LightGBM, logistic regression (LR), and random forest (RF) models, individually. For LightGBM, we searched 150 combinations chosen from following hyperparameter configurations: learning rate (chosen from [0.1, 0.05, 0.02, 0.01]), the number of estimators (chosen from 24 points that were evenly spaced between 100 and 2400), the maximum depth of the individual estimators (chosen from [2–5, 10, 20, 40, 51]), the minimum number of data in one leaf (chosen from 22 points that were evenly spaced between 1 and 44), the fraction of subset on each estimator (chosen from [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]), the frequency of bagging (chosen from [0, 1, 2]), the penalty of L1 regularization (chosen from [0, 0.001, 0.005, 0.01, 0.1]), and the penalty of L2 regularization (chosen from [0, 0.001, 0.005, 0.01, 0.1]). For LR, the penalty of L2 regularization was randomly chosen from integer between 50 and 1000 to optimize hyperparameters. For RF, we searched 150 combinations through the following hyperparameter configurations: the number of trees in the forest (randomly chosen from integer between 1 and 250) and the maximum depth of the tree (randomly chosen from integer between 5 and 255).

To select the optimal hyperparameters for each model, a fivefold cross-validation based on RandomizedSearchCV was performed. The F1 score was used as the evaluation metric, and the model that achieved the highest mean F1 score during cross-validation was chosen for further comparison and analysis.

Statistical analyses

Statistically significant differences were assessed as described in correspondent figure legends. All statistic tests were performed with the R platform (version 4.1.1) (http://www.R-project.org/).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03397-2.

Additional file 1: Fig. S1-S14 with figure legends.

Additional file 2: Table S1. Summary of downloaded datasets used in this study. Include 403 donors' datasets for DeepDDR model establish, two independent validation datasets, RNA-seq data from 19 AML patients and 17 healthy donors and datasets for DeepDDR-transfer training.

Additional file 3: Table S2. Information of RNA editing sites (REs) for DeepDDR development and evaluation. In total, 122,872 RNA editing sites (REs) were used to develop and evaluate DeepDDR. Genomic coordinates (hg38), SampleID, Mutation type, Mutation frequency, Hits per billion mapped bases (HPB), Host gene, Repetitive element, Gene body, Dataset type and Label are included.

Additional file 4: Table S3. Information of DNA mutations (DMs) for DeepDDR development and evaluation. In total, 122,872 DNA mutations (DMs) were used to develop and evaluate DeepDDR. Genomic coordinates (hg38), SampleID, Mutation type, Mutation frequency, Hits per billion mapped bases (HPB), Host gene, Repetitive element, Gene body, Dataset type and Label are included.

Additiona file 5: Table S4. Information of true REs and DMs from independent HG00145. In total, 5,395 REs and 17,230 DMs were identified from HG00145 for deep learning model evaluation. Genomic coordinates (hg38), SampleID, Mutation type, Mutation frequency, Hits per billion mapped bases (HPB), Host gene, Repetitive element, Gene body and Label are included.

Additional file 6: Table S5. Information of true REs and DMs from mouse bone marrow dataset including three *Adar* knockout samples and three wild-type samples. In total, 345 REs and 345 DMs were identified from mouse bone marrow dataset used to deep learning model evaluation. Genomic coordinates (mm10) and Label are included.

Additional file 7: Table S6. Information of true REs and DMs from mouse brain dataset including three *Adar* and *Adarb1* knockout samples and three wild-type samples. In total, 5,586 REs and 5,586 DMs identified from mouse brain dataset were used to transfer learning. Genomic coordinates (mm10) and Label are included.

Additional file 8: Table S7. Information of true REs and DMs from nematode dataset including four *adrs* double knockout and four wild-type samples. In total, 403 REs and 403 DMs identified from nematode dataset used to deep learning model evaluation. Genomic coordinates (ce11) and Label are included.

Additional file 9: Table S8. Information of DMs identified in AML patients by DEMINING. In total, 195,256 DMs were identified by DEMINING in RNA-seq from 19 AML patients. Genomic coordinates (hg38), SampleID, Mutation type, Mutation frequency, Hits per billion mapped bases (HPB) and Condition of AML-specific are included.

Additional file 10: Table S9. Information of REs identified in AML patients by DEMINING. In total, 137,682 REs identified by DEMINING in RNA-seq from 19 AML patients. Genomic coordinates (hg38), SampleID, Mutation type, Mutation frequency, Hits per billion mapped bases (HPB) and Condition of AML-specific are included.

Additional file 11: Table S10. Information of AML-specific DMs. In total, 57,150 AML-specific DMs identified by filtering those also detected in 17 normal control samples or had a high minor allele frequency (MAF > 0.05) in 1000 Genomes Project. Genomic coordinates (hg38), SampleID, Mutation type, Mutation frequency, Hits per billion mapped bases (HPB), Repetitive element, Host gene, Gene body, Condition of recoding, Amino acid change type, Condition of overlapped with ClinVar, COSMIC or dbSNP database are included.

Additional file 12: Table S11. Information of AML-specific REs. In total, 58,045 AML-specific REs were identified in AML patients by filtering those also detected in 17 normal control samples. Genomic coordinates (hg38), SampleID, Mutation type, Mutation frequency, Hits per billion mapped bases (HPB), Repetitive element, Host gene, Gene body, Condition of recoding, Amino acid change type, Condition of overlapped with RADAR database and Condition of overlapped with REDIportal database are included.

Additional file 13: Review History.

Acknowledgements

We thank all members of the Yang laboratory for discussion and support.

Review history

The review history is available as Additional File 13.

Peer review information

Kevin Pang and Veronique van den Berghe were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
L.Y. conceived and supervised the project. Z.-C.F. and B.-Q.G. performed most analyses with the help of F.N. and X.-K.M., supervised by L.Y. L.Y. wrote the paper with input from all authors. All authors read and approved the final manuscript.

Funding
This work was supported by National Natural Science Foundation of China (NSFC, 31925011), the Ministry of Science and Technology of China (MoST, 2019YFA0802804, 2021YFA1300503), the Shanghai Municipal Science and Technology Commission (STCSM, 23JS1400300 and 23DX1900102), and the Chinese Academy of Sciences (CAS, XDB38040300) to L.Y., by STCSM (23YF1407400) to X.K.M., and by China National Postdoctoral Program for Innovative Talents (BX20220077) and Shanghai Post-doctoral Excellence Program, China (2022728) to F.N.

Availability of data and materials
The statements of data availability and their associated accession codes and references are available in the "Methods" section and Additional file 2: Table S1. Multiple datasets from independent resources were used in this study. The first collection consisted of WGS/WES and RNA-seq datasets of 403 donors was available from 1000 Genomes Project phase 1 [59], phase 3 [60], and the Geuvadis consortium [61]. The second collection consisted of WGS/WES and RNA-seq datasets of an additional donor was available from 1000 Genomes Project phase 3 [60] and the Geuvadis consortium [61]. The third data collection consisted of RNA-seq datasets of 19 acute myeloid leukemia (AML) patients (peripheral blood or bone marrow tissue) and 17 normal control samples (medullary thymic or myelocytic precursor cells) was available from the Gene Expression Omnibus (GEO) with accession number GSE147524 [62] and GSE98310 [63]. The raw mass spectrometric dataset of major histocompatibility complex (MHC) associated peptides from 19 AML patient was available from ProteomeXchange Consortium with accession number PXD018542 [64]. The final collection consisted of RNA-seq of mouse bone marrow, mouse brain RNA-seq, and nematode which was available from the Sequence Read Archive with project ID PRJEB31568 [65], PRJNA546532 [66], and PRJNA215361 [67], respectively.
All scripts used in this study, including DEMINING framework, DeepDDR model, and related codes, are currently available at https://github.com/YangLab/DEMINING [68] with GPLv3 license for open source use and in Zenodo with DOI: https://doi.org/10.5281/zenodo.12903872 [69].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
F.N. and L.Y. have filed a patent application (202310642373.8) relating to this work through Children's Hospital of Fudan University. However, the patent does not restrict the educational, research, and not-for-profit purposes. The remaining authors declare no competing interests.

Received: 15 December 2023 Accepted: 20 September 2024
Published online: 08 October 2024

References

1. Chen LL, Yang L. ALUternative regulation for gene expression. Trends Cell Biol. 2017;27:480–90.
2. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
3. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 2014;42:D109-113.
4. Wang C, Davila JI, Baheti S, Bhagwate AV, Wang X, Kocher JP, Slager SL, Feldman AL, Novak AJ, Cerhan JR, et al. RVboost: RNA-seq variants prioritization using a boosting method. Bioinformatics. 2014;30:3414–6.
5. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93:641–51.
6. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.
7. Higuchi M, Single FN, Kohler M, Sommer B, Sprengel R, Seeburg PH. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. Cell. 1993;75:1361–70.
8. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004;2: e391.
9. Moller-Krull M, Zemann A, Roos C, Brosius J, Schmitz J. Beyond DNA: RNA editing and steps toward Alu exonization in primates. J Mol Biol. 2008;382:601–9.
10. Bass B, Hundley H, Li JB, Peng Z, Pickrell J, Xiao XG, Yang L. The difficult calls in RNA editing. Interviewed by H Craig Mak. Nat Biotechnol. 2012;30:1207–9.
11. Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. Nat Methods. 2012;9:579–81.
12. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. Identifying RNA editing sites using RNA sequencing data alone. Nat Methods. 2013;10:128–32.
13. Zhang Q, Xiao X. Genome sequence-independent identification of RNA editing sites. Nat Methods. 2015;12:347–50.

14. Zhu S, Xiang JF, Chen T, Chen LL, Yang L. Prediction of constitutive A-to-I editing sites from human transcriptomes in the absence of genomic sequences. BMC Genomics. 2013;14: 206.
15. Park E, Williams B, Wold BJ, Mortazavi A. RNA editing in the human ENCODE RNA-seq data. Genome Res. 2012;22:1626–33.
16. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011;333:53–8.
17. Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." Science. 2012;335:1302 ; author reply 1302.
18. Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." Science. 2012;335:1302 ; author reply 1302.
19. Lin W, Piskol R, Tan MH, Li JB. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." Science. 2012;335:1302 ; author reply 1302.
20. Zhang P, Zhu Y, Guo Q, Li J, Zhan X, Yu H, Xie N, Tan H, Lundholm N, Garcia-Cuetos L, et al. On the origin and evolution of RNA editing in metazoans. Cell Rep. 2023;42:112112.
21. Zhou ZY, Hu Y, Li A, Li YJ, Zhao H, Wang SQ, Otecko NO, Zhang D, Wang JH, Liu Y, et al. Genome wide analyses uncover allele-specific RNA editing in human and mouse. Nucleic Acids Res. 2018;46:8888–97.
22. Xiong H, Liu D, Li Q, Lei M, Xu L, Wu L, Wang Z, Ren S, Li W, Xia M, et al. RED-ML: a novel, effective RNA editing detection method based on machine learning. Gigascience. 2017;6:1–8.
23. Kim MS, Hur B, Kim S. RDDpred: a condition-specific RNA-editing prediction model from RNA-seq data. BMC Genomics. 2016;17(Suppl 1):5.
24. Lappalainen T, Sammeth M, Friedländer MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506–11.
25. Wang X, Ding C, Yu W, Wang Y, He S, Yang B, Xiong YC, Wei J, Li J, Liang J, et al. Cas12a base editors induce efficient and specific editing with low DNA damage response. Cell Rep. 2020;31: 107723.
26. Wang J, Ness S, Brown R, Yu H, Oyebamiji O, Jiang L, Sheng Q, Samuels DC, Zhao YY, Tang J, Guo Y. EditPredict: prediction of RNA editable sites with convolutional neural network. Genomics. 2021;113:3864–71.
27. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 22–29 Oct. 2017. 2017. p. 618–26.
28. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929: IEEE Computer Society. 2016. p. 2921–9.
29. Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res. 2012;22:142–50.
30. Brümmer A, Yang Y, Chan TW, Xiao X. Structure-mediated modulation of mRNA abundance by A-to-I editing. Nat Commun. 2017;8:1255.
31. Ouyang Z, Ren C, Liu F, An G, Bo X, Shu W. The landscape of the A-to-I RNA editome from 462 human genomes. Sci Rep. 2018;8:12069.
32. Jain M, Jantsch MF, Licht K. The editor's I on disease development. Trends Genet. 2019;35:903–13.
33. Porath HT, Carmi S, Levanon EY. A genome-wide map of hyper-edited RNA reveals numerous new sites. Nat Commun. 2014;5:4726.
34. Liu X, Sun T, Shcherbina A, Li Q, Jarmoskaite I, Kappel K, Ramaswami G, Das R, Kundaje A, Li JB. Learning cis-regulatory principles of ADAR-based RNA editing from CRISPR-mediated mutagenesis. Nat Commun. 2021;12:2165.
35. Ehx G, Larouche JD, Durette C, Laverdure JP, Hesnard L, Vincent K, Hardy MP, Theriault C, Rulleau C, Lanoix J, et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Immunity. 2021;54(737–752): e710.
36. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, Long N, Schultz AR, Traer E, Abel M, et al. Functional genomic landscape of acute myeloid leukaemia. Nature. 2018;562:526–31.
37. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18:696–705.
38. Zhao S, Allis CD, Wang GG. The language of chromatin modification in human cancers. Nat Rev Cancer. 2021;21:413–30.
39. Meszaros B, Hajdu-Soltesz B, Zeke A, Dosztanyi Z. Mutations of Intrinsically disordered protein regions can drive cancer but lack therapeutic strategies. Biomolecules. 2021;11:381.
40. Borcherds W, Bremer A, Borgia MB, Mittag T. How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? Curr Opin Struct Biol. 2021;67:41–50.
41. Tong X, Tang R, Xu J, Wang W, Zhao Y, Yu X, Shi S. Liquid-liquid phase separation in tumor biology. Signal Transduct Target Ther. 2022;7:221.
42. Iqbal Z, Absar M, Akhtar T, Aleem A, Jameel A, Basit S, Ullah A, Afzal S, Ramzan K, Rasool M, et al. Integrated genomic analysis identifies ANKRD36 gene as a novel and common biomarker of disease progression in chronic myeloid leukemia. Biology (Basel). 2021;10:1182.
43. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17:19–32.
44. Lindeboom RGH, Vermeulen M, Lehner B, Supek F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. Nat Genet. 2019;51:1645–51.
45. Shamsani J, Kazakoff SH, Armean IM, McLaren W, Parsons MT, Thompson BA, O'Mara TA, Hunt SE, Waddell N, Spurdle AB. A plugin for the Ensembl Variant Effect Predictor that uses MaxEntScan to predict variant spliceogenicity. Bioinformatics. 2019;35:2315–7.
46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

47. Yepez VA, Gusic M, Kopajtich R, Mertes C, Smith NH, Alston CL, Ban R, Beblo S, Berutti R, Blessing H, et al. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. 2022;14:38.
48. Kapoor U, Licht K, Amman F, Jakobi T, Martin D, Dieterich C, Jantsch MF. ADAR-deficiency perturbs the global splicing landscape in mouse tissues. Genome Res. 2020;30:1107–18.
49. Chalk AM, Taylor S, Heraud-Farlow JE, Walkley CR. The majority of A-to-I RNA editing is not required for mammalian homeostasis. Genome Biol. 2019;20:268.
50. Zhao HQ, Zhang P, Gao H, He X, Dou Y, Huang AY, Liu XM, Ye AY, Dong MQ, Wei L. Profiling the RNA editomes of wild-type C. elegans and ADAR mutants. Genome Res. 2015;25:66–75.
51. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
52. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11:1650–67.
53. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38: e164.
54. Adusumilli R, Mallick P. Data conversion with ProteoWizard msConvert. Methods Mol Biol. 2017;1550:339–68.
55. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013;13:22–4.
56. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002;74:5383–92.
57. da Veiga LF, Haynes SE, Avtonomov DM, Chang HY, Shanmugam AK, Mellacheruvu D, Kong AT, Nesvizhskii AI. Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods. 2020;17:869–70.
58. Mansi L, Tangaro MA, Lo Giudice C, Flati T, Kopel E, Schaffer AA, Castrignano T, Chillemi G, Pesole G, Picardi E. REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. Nucleic Acids Res. 2021;49:D1012–9.
59. Fairley S, Lowy-Gallego E, Perry E, Flicek P: The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Datasets. The International Genome Sample Resource. 2011. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521.
60. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Datasets. The International Genome Sample Resource. 2013. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502.
61. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al: Transcriptome and genome sequencing uncovers functional variation in humans. Datasets. European Nucleotide Archive. 2013. https://identifiers.org/ena.embl:ERP001942.
62. Ehx G, Larouche JD, Durette C, Laverdure JP, Hesnard L, Vincent K, Hardy MP, Theriault C, Rulleau C, Lanoix J, et al: Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Datasets. Gene Expression Omnibus. 2021. https://identifiers.org/geo:GSE147524.
63. Maiga A, Lemieux S, Pabst C, Lavallee VP, Bouvier M, Sauvageau G, Hebert J. Transcriptome analysis of G protein-coupled receptors in distinct genetic subgroups of acute myeloid leukemia: identification of potential disease-specific targets. Datasets. Gene Expression Omnibus. 2017. https://identifiers.org/geo:GSE98310.
64. Ehx G, Larouche JD, Durette C, Laverdure JP, Hesnard L, Vincent K, Hardy MP, Theriault C, Rulleau C, Lanoix J, et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Datasets. ProteomeXchange. 2021. https://identifiers.org/px:PXD018542.
65. Kapoor U, Licht K, Amman F, Jakobi T, Martin D, Dieterich C, Jantsch MF. ADAR-deficiency perturbs the global splicing landscape in mouse tissues. Datasets. Sequence read archive. 2020. https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31568.
66. Chalk AM, Taylor S, Heraud-Farlow JE, Walkley CR. The majority of A-to-I RNA editing is not required for mammalian homeostasis. Datasets. Sequence read archive. 2019. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA546532.
67. Zhao HQ, Zhang P, Gao H, He X, Dou Y, Huang AY, Liu XM, Ye AY, Dong MQ, Wei L: Profiling the RNA editomes of wild-type C. elegans and ADAR mutants. Datasets. Sequence Read Archive. 2015. https://www.ncbi.nlm.nih.gov/bioproject/215361.
68. Fu ZC, Gao BQ, Nan F, Ma XK, Yang L. DEMINING: a deep learning model embedded framework to distinguish RNA editing from DNA mutations in RNA sequencing data. GitHub. 2024. https://github.com/YangLab/DEMINING.
69. Fu ZC, Gao BQ, Nan F, Ma XK, Yang L. DEMINING: a deep learning model embedded framework to distinguish RNA editing from DNA mutations in RNA sequencing data. 2024. Zenodo. https://doi.org/10.5281/zenodo.12903872.

Publisher's Note