Decoding the stochastic profile of human N6-methyladenosine (m6A) over the entire transcriptome

Jiaying Wang[1], Zhen Wei[1, 2, *]

[1]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China

[2]Institute of Ageing & Chronic Disease, University of Liverpool, L7 8TX Liverpool, UK

*Corresponding author:

Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China; E-mail: zhen.wei01@xjtlu.edu.cn (ZW).

(P.S. This is a draft, and the complete article is planned to be submitted to the journal in one month)

1. ABSTRACT

N6-methyladenosine (m^6A), an abundant eukaryotic mRNA modification, is a crucial marker dynamically regulated by demethylase (Erasers), methyltransferase (Writers), and binding proteins (Readers). Hence, decoding the stochastic profile of m6A over transcriptome is invaluable to our understanding of the biological functions of RNA. The m^6A frequencies over 1624625 DRACH motifs on human exons were summarized from 40 single-based m6A experiments. Four machine learning algorithms, generalized linear model (GLM), multi-layer perceptron (MLP), extreme gradient boosting (XGBoost), and random forest (RF), were implemented to build the Poisson regression models. Compared with the classification models used in previous studies, our Poisson regression approaches provide a new framework for integrating multiple single-based RNA modification datasets. We demonstrate that the Poisson regressors can better predict the biological and technical variation between experiments than the corresponding classifiers trained using the same feature set. In addition, we for the first time introduced the binding sites of 17 m6A regulators as the predictive features. Compared to only using the sequence-derived and genome-derived features (MSE 1.020 / CE 0.579; AUC 0.854 / MCC 0.410), predictive performances can be significantly improved after adding the regulator features (MSE 0.855 / CE 0.503; AUC 0.883 / MCC 0.469). These results suggest the importance of the information of protein regulators when building high accuracy epi-transcriptomic predictors. Finally, we

provide a predicted stochastic m^6A map on the entire human exonic region, and an in-depth analysis is performed on the feature importance of both the linear and non-linear models.

2. KEY WORDS

RNA modification, N6-methyladenosine (m^6A), m^6A regulators, Poisson Regression, Machine Learning

3. INTRODUCTION

N6-methyladenosine (m^6A) was firstly discovered on messenger RNA as an abundant nucleotide modification (Schwartz et al., 1974). However, it was not until a few years ago that the interest on the fundamental mRNA marker regained its momentum, mainly due to the recognition of its' ubiquity and the functional significant roles on gene expression regulation (Roundtree et al., 2017), DNA repair (Xiang et al., 2017), cell differentiation (Frye et al., 2018) and regulation of viral infection (Kennedy et al., 2016). Therefore, it is an essential task of RNA biochemistry to identify the specific position of m^6A residues and to reveal its function mechanisms in cell biology. Several independent shreds of evidence integrated to reveal that N6-methyladenosine (m6A) occurs mainly in DRACH consensus motif (D for A/G/U, R for A/G, H for A/C/U) (Linder et al., 2015). While the frequency of DRACH motifs appearing in the transcriptome is around 10 folds higher than the observed m^6A methylation sites, criterion other than motif are required to account for the selectivity.

MeRIP-Seq is the first technique developed to determine the transcriptome-wide mapping of mA sites (Meyer et al., 2012), in which the Poly(A)+-selected RNA were randomly fragmented into 100 nt, followed by immunoprecipitation and purification. High-throughput sequencing (like Illumina GAIIx) was used to sequence the Immunoprecipitated and input control samples (non-IP sample) (Meyer et al., 2012). m6A peaks were further identified by MACS peak-calling algorithms or other peak callers such as exomePeak (Meng et al., 2014), MeTPeak (Cui et al., 2016), MeTDiff (Cui et al., 2018). The location of the m^6A site can be narrowed down by finding the DRACH motif within the m^6A containing peaks. The method can lead to many potential false positives, since all DRACH motifs located within the peak are all selected as the m^6A site. (Chen et al., 2019).

To solve the low-resolution defect of the m6A-seq, base-resolution HTP techniques, such as miCLIP (Linder et al., 2015) and m6A-CLIP (Ke et al., 2015), were developed. However, since those two technologies are antibody-dependent, it is hard for them to define the stoichiometry of individual m^6A sites. Moreover, the library preparation requires RNA fragmentation, which can lead to the underrepresented coverage around RNA 3' ends (Parker et al., 2020). Therefore, two antibody-independent techniques MAZTER-seq and m6A-REF-Seq have been developed by two independent groups in 2019 (Garcia-Campos et al., 2019; Zhang et al., 2019). Another technique, DART-Seq, applies standard RNA-Seq to detect C-to-U deamination at sites adjacent to m^6A

residues induced by fusing cytidine deaminase APOBEC1 to the m^{6}A-binding YTH domain (Meyer, 2019). All of the antibody-independent techniques are well reviewed and can provide high-resolution detection of m6A epitranscriptome (Huang et al., 2020).

Although various high throughput techniques can provide high-resolution mapping of m^{6}A, the considerable cost and complicated experiments made it hard to be widely applied. As an extension of the wet lab efforts, many web-based machine learning site predictors have subsequently been developed based on the published data obtained from the base resolution experiments. SPAMP was devised for mammalian m^{6}A predication, combining three RF classifiers on sequence-derived features including positional binary encoding of nucleotide sequence, the K-nearest neighbor (KNN) encoding, and nucleotide pair spectrum encoding (Zhou et al., 2016). Recently, a prediction framework WHISTLE has obtained significantly higher performances by adding the genome-derived features onto the sequence-derived features under the classifiers of RF and SVM (Chen et al., 2019).

It has been demonstrated that m^{6}A is a dynamic and reversible modification that can be installed by m^{6}A methyltransferases and erased by demethylases (Liu et al., 2014). However, the existing m^{6}A predictors are mostly using the classification algorithms such as SVM or RF, which treated the m^{6}A status as inherently binary. By taking the union of sites from different experiments as the ground truth positives, the classification framework often ignores the existence of biological variation and technical sensitivities

over different samples. Furthermore, the union approach will consider a site to be positive when it was detected at least once in multiple samples, which easily introduces all the technically false positives and biologically weak signals without any differentiation. As an improvement, we represent the times a specific site to be detected among multiple experiments as a count data, and we use Poisson regression model to better fit the count information from multiple datasets. Moreover, feature importance metrics from Poisson regression are calculated to reveal the factors that can generally increase the rate of detection among different experiments. Last but not least, majority of the previous predictors are constructed using the sequence-derived information and genome-derived features, none of them have incorporated the binding sites of all m^6A regulators, which can be the deterministic genomic factors for m^6A methylation, into their analysis.

In this study, we characterized the rate of methylation on all exonic DRACH motifs using Poisson regression model over 40 independent m6A experiments. The binding sites of m^6A regulators was combined with the sequence-derived features and genome-derived features in WHISTLE to achieve higher predictive accuracy. Various evaluation metrics including cross entropy were calculated to assess the performance of models. Multiple machine learning algorithms and the combination of features sets are implemented. Moreover, comprehensive model interpretations were conducted to investigate the relationship between genome features and the m^6A modification rate.

4. Material and method

4.1 Datasets preparation

The datasets used for model building and performance evaluation were manually collected from 40 single-base resolution m6A experiments (Supplementary Table S1). More specifically, all the m6A sites in the datasets are on the human (Homo Hsapiens) genome, and the experimental technologies used include MAZTER-Seq, m6A-REF-Seq, and DART-Seq together with the CLIP-based including miCLIP, m6A-PA-CLIP, and m6A-CLIP. Those data were manually collected from Gene Expression Omnibus (GEO).

4.2 Features used for machine learning predication

4.2.1 *Feature iRNA*

The *iRNA* (Chen et al., 2017) feature is a sequence-derived feature which encodes the nucleotide sequences in the 41 nt flanking windows with DRACH motif at the center. Feature iRNA is the combination of four sub-features. First three features focus on the chemical property of different nucleotides and chemical structures and chemical binding. For chemical property, adenine(A) and guanine(G) contain a fused ring structure derived from purine, while cytosine(C) and uracil(U) contains single ring skeletal structure derived of pyrimidine (1 encodes for A or G, 0 encodes for C or U). For chemical functionality, adenine (A) and cytosine (C) have the amino group on the ring, while guanin (G) and uracil (U) have the keto group on the ring (1 encodes for A

or C, 0 encodes for G or U). For chemical binding, Adenine(A) and uracil(U) will form strong hydrogen bonds, guanine(G) and cytosine(C) will form weak hydrogen bonds (1 encodes for A or C, 0 encodes for G or U). The final feature is the nucleotide frequency, this feature shows the nucleotide and nucleotide distribution in 41 bp franking window centered on DRACH motif, and it is calculated according to the formula below:

$$d_i = \frac{1}{|N_i|}\sum_{j=1}^{l} f(n_j), \quad f(n_j) = \begin{cases} 1 & if\ n_j = q \\ 0 & other\ cases \end{cases}$$

Where d_i denotes the density of any nucleotide n_j at position i in RNA sequence, l denotes the sequence length, $|N_i|$ is the length of the i-th prefix string $\{n_1, n_2, n_3 \ldots, n_j\}$ in the sequence, and $q \in \{A, C, G, U\}$. For better understanding, take the sequence "AAACU" as an example, it will be encoded as a vector $[1(1/1), 1(2/2), 1(3/3), 0.25(1/4), 0.2(1/5)]$.

4.2.2 *Genome-derived feature*

The prediction framework WHISTLE reported that the model's prediction performace had been greatly improved after joining the genome-derived features (Chen et al., 2019). In our models, what we want to find is the probability of the site to be stably methylated according to their feature state combination. Therefore, it is necessary to add these biologically meaningful features. There are 35 genome-derived features, Feature 1-13 are dummy variable features that show whether the site is overlapped to a certain topological region, and they were generated by the Genomic Features R/Bioconductor

166 package using the transcript annotations hg19 TxDb package (Lawrence et al., 2013).
167 To reduce the influence of isoform ambiguity, the transcript sub-regions on the primary
168 transcripts of each gene were extracted. Feature 14-16 (relative position on 5'UTR,
169 3'UTR, exon) are about the relative position of the transcript regions in the form of
170 proportion. Feature 17-19 (length of 5'UTR, 3'UTR, exon) are about the length of the
171 transcript region containing DRACH sites which were shown in bp length, for those
172 sites that do not belong to the region, the value was set to zero. Feature 20-21 are about
173 the nucleotide distances toward the 5' splicing junctions and 3' splicing junctions.
174 Feature 22 is about the distance from the sites to the thier closest neighbor DRACH site.
175 Feature 23-26 are about the conservation degree of the sites and their flanking regions,
176 which are calculated using Phast-Cons score (Siepel et al., 2005) and the fitness
177 consequence (Gulko et al., 2015). Feature 27-28 are about the RNA secondary structure
178 (predicted RNA hybridized region and loop region), the structures are predicted by
179 RNAfold from the Vienna RNA package (Gruber et al., 2015). Feature 29-31 are related
180 to m6A biology, like eCLIP data of HNRNPC RNA binding sites (Consortium, 2012),
181 the miRNA targeted sites (Betel et al., 2010;Agarwal et al., 2015), feature 32-35 are
182 about the attributes of the gene or transcripts containing the m6A site (here indicate
183 DRACH site), more detailed information about genomic features can be found in
184 Supplementary Table S2.
185
186 **4.2.3** *Regulators feature*
187 Here a total of 17 regulators were included in feature "*Regulators*" for each site

information was downloaded from GEO, the relative information can be found in Supplementary Table S2. Recall that the m6A regulation pathway is composed of three classes of protein factors their roles as “writers”, “erasers”, and “readers”. METTL3, METTL14, METTL16, and WTAP are components of the methyltransferase complex that catalyzes m6A methylation, serve as “writers” (Yao et al., 2021). For FTO and ALKBH5, the demethylases catalyze oxidative demethylation of m6A, known as “erasers” (Yang and Nam, 2020). The m6A-modified RNA reader proteins include YTH-domain containing proteins YTHDF1 (promotes mRNA translation), YTHDF2 (reduces mRNA stability), YTHDF3 (mediates the translation or degradation), YTHDC1 (promotes RNA splicing and translocation), and YTHDF2 (reduces mRNA stability) (Jiang et al., 2021;Lou et al., 2021). IGF2BP1/2/3 (enhances mRNA stability), protein HNRNPC (mediates the mRNA splicing), and elF3, those effectors can exert various functions in all stages of the RNA life cycle (Yao et al., 2021). Moreover, Histone H3 Trimethylation at Lysine 36 (H3K36me3) is related to Histone modification (Huang et al., 2019). By mapping that information together with the DRACH site on the genome, we can get whether those regulator binding sites are overlapped on the DRACH site or not, this is how regulator features were implemented.

4.3 Machine learning approach

4.3.1 machine learning algorithms

In this study, we decided to try four algorithms representing different levels of complexity: generalized linear modeling (GLM), Distributed Random Forest (DRF),

Deep Learning (Neural Networks), and extreme gradient boosting (XGboost). More specifically, GLM estimate regression models for response variables following exponential family-like Gaussian, Poisson, and gamma distributions (Click et al., 2016). For RF, it is a widely used algorithm in computational biology, SRAMP as a computational predictor of mammalian m6A sites employs the Random Forest classifier (Zhou et al., 2016). Another m6A site predictor also applied this algorithm for model construction (Chen et al., 2019). Deep learning is a powerful method that has recently been shown impressively in the data-rich domain like biology due to its' ability to integrate colossal data sets and learn arbitrarily complex relationships (Ching et al., 2018;Wainberg et al., 2018). For XGboost, it is an algorithm that implements a scalable end-to-end tree boosting which gives state-of-the-art result for various problems, and the innovation on algorithmic optimization made it scalable in different problem solving, which essentially speed up the computation (Chen and Guestrin, 2016;Mitchell and Frank, 2017), but this algorithm has not been used in the post-transcriptional site prediction. In this project, we will apply the above four algorithms for model building.

4.3.2 Poisson approximation to PBD

To handle this kind of response variable which comes from 40 experiments, one reasonable way is to return a probability. We firstly denote the probability distribution over possible labels $p(y|x, D)$, x is the input vector of the different feature combinations, D is the training set.

Let dummy variable y_{ij} denote whether the adenosine i on the center of the DRACH was detected to be methylated (1 for methylated, 0 for unmethylated) under experiment j. p_{ij} is the unknown actual probability that the y_{ij} is 1 (methylated). There are a total of 40 CLIP-based experiments (seen as number S of events), we may empirically assume that each experiment j is an independent and repeatable experiment with the same p_{ij} for each specific site i, so it is precisely the binomial distribution $B(n, p_i)$. Since n is large enough and p_i is quite small for majority sites, each binomial distribution can be further approximated by Poisson distribution with expectation $n * p_i$. However, due to different cell lines together with technical artificial variance, it is not reasonable to regard each experiment j for a specific site as i.i.d (or equally likely to succeed), then S_i has the distribution sometimes called Poisson binomial distribution (PBD) where p_{ij} are not necessarily identical for each site i. Fortunately, Hodges and Le Cam provided (Hodges and Lecam, 1960) an approximation theorem that helps to explain how this situation can be well approximated by Poisson distribution (Detailed prove were presented in the Supplementary File S1):

We focus on random site i, let x_{ij} indicate the random variables that have Poisson distribution with $E(x_{ij}) = p_i$, following are the joint distribution of x_{ij} and y_{ij}.

For the additive property of Poisson variables, $T_i = sum(x_{ij})$ has the Poisson distribution. We aim to show that $S_i = sum(y_{ij})$ has a very similar distribution. We let

$$D_i = \sup_u |P(S_i \leq u) - P(T_i \leq u)| \quad (1)$$

D denotes the maximum absolute difference between the cumulates of S and T, and what we want is to find the condition under which D is small.

It can be further narrowed down to:

$$D_i \leq 2\sum_{j=1}^{n} {p_{ij}}^2 \tag{2}$$

Then we denote the random variable $Z_i = X_i - Y_i$, $E(Z_i) = 0$, While

$$Var(Z_i) = E({Z_i}^2) = p_{ij}(1 - e^{-p_i}) + \sum_{k=2}^{\infty} k^2({p_{ij}}^k e^{-p_{ij}})/k! \leq 3{p_{ij}}^2$$

let $\sum Z_i = U_i$, then$E(U) = 0$, $Var(U) \leq 3\mu$

Here we introduce a to be any positive number a. Let $T_i = S_i + U_i \leq \nu - a$, we can further get:

$$D_i = \sup_{\nu} |P(S_i \leq \nu) - P(T_i \leq \nu)| \leq \sup_{\nu} P(\nu \leq T_i \leq \nu + a) + P(|U_i| \geq a) \tag{3}$$

By using Chebycheff inequality together with the upper bound of the T_i which is "$(1 + 1/12\lambda)/(2\pi\lambda)^{1/2}$ "(will not go into detail here), the flowing equation can prove:

$$D_i \leq (3\mu/a^2) + (a+1)(1 + 1/12\lambda_i)/(2\pi\lambda_i)^{1/2} \tag{4}$$

Using (2) and (4) we can get the result：

$$D_i \leq 3\sqrt[3]{a_i} \tag{5}$$

More specifically, in our situation, the above approximation theorem implies that maximum absolute difference D_i between the cumulative distributions of S_i and Poisson distribution $\mathcal{p}(\sum p_{ij})$ tends to 0, as $\alpha_i = max\{p_{i1}, \dots, p_{in}\} \to 0$. Moreover, the approximation theorems also suggest that the condition that $\alpha_i \to 0$ is sufficient but not necessary for $D_i \to 0$, in another words, S_i will have approximately a Poisson distribution even if few of p_{ij} are quite large, provided these values contribute only a

small part of the total $\sum pi$ which make this model more robust.

4.3.3 dataset processing & model building

Since we specify the response variables follow a conditional Poisson distribution in the regression model, it is more suitable to use cross entropy for model evaluation. createFolds function in caret package was used to split the dataset into five equally sized groups. In each loop, four groups will be combined as the training data, the performance of the model will be tested on the rest data group using predict function in h2o package to get the prediction. After five loops, we can get the prediction for the whole dataset and various metrics will be calculated to evaluate the difference between the prediction and the actual label. The features' relative importance from each tree-based model were retrieved for later analysis, and for GLM models the standard coefficient for each feature were extracted.

To compare the classic binary classification models with the Poisson regression models that model the distribution of the target variable, we extracted the p1 value returned from the classification models for each site, here p1 is the predicted probability of the site to be the positive site, it could seem equally as the outcome of the regression model, for classification model the prediction is made by applying a threshold on the p1 value for getting the max F1. Since there are 40 independent experiments, metrics used the same as the regression model were calculated on the total number of experiments multiplied by the p1 for each site which can be seen as a continuous response (see Table

3).

To obtain what with what kind of characteristics, the site will be stably detected positive in different independent experiments, the DRACH motif sites that were detected as positive once or more in 40 experiments were extracted for model building separately from the original models. We focus on the relative importance of features extracted from the models, mainly pursuing scientific meanings and explanatory characters.

4.4 Performance evaluation of the models

For the classification model, the prediction accuracy was measured by the receiver operating characteristic (ROC) curve, showing the model's performance at all classification thresholds. AUROC calculated the two-dimensional area underneath the entire ROC curve, it is equal to the value of the Wilcoxon-Mann-Whitney test static and measures the discrimination. AUROC was our classification model's main performance evaluation metrics and widely used in previous m6A predictors' classification models like WHISTLE. In addition, Matthew correlation coefficient (MCC) or phi coefficient (φ) was calculated to measure the quality of binary classifications predictors:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (1)$$

Where *TP* stands for true positive, *FP* for false positive, *FN* for false negative, *TN* for true negative. From the equation, MCC takes the whole confusion matrix into account and is perfectly symmetric, valued between 1 and -1.

For the Poisson regression models, the prediction ability was evaluated using the mean squared error (MSE) and cross entropy (CE) mentioned previously. Kullback-Leibler Divergence (KL divergence) is frequently used in machine learning to quantify the difference between the two probability distributions (2) given the measure η, recall that the probability distributions of an exponential family have the general form (3). Combining (2) and (3), we can easily get the KL divergence of the exponential family (4), in our project, η_1 stand for the real distribution, and η_2 stands for the predicted distribution since the η_1 is fixed, we introduce the cross entropy as our regression models' metrics, which can be expressed in the form of (5), where $\eta_1 = E_{\eta_1}[T(X)]$ is the mean parameter. For Poisson distribution, the cross entropy is (6), where true λ was estimated using the Law of large number (LLN), λ_{pre} is the predicted value from the Poisson regression models. The details of the derivation of the process formula can be viewed in the subsequent documents (Supplementary File S1):

$$D(p(x|\eta_1) \parallel p(x|\eta_2)) = E_{\theta_1} \log \frac{p(X|\eta_1)}{p(X|\eta_2)} \quad (2)$$

$$p(x|\eta_1) = h(x)exp\eta^T T(x) - A(\eta) \quad (3)$$

$$D(\eta_1 \| \eta_2) = (\eta_1 - \eta_2)^T \mu_1 - A(\eta_1) + A(\eta_2) \quad (4)$$

$$CE = -E_{\eta_1} logh(x) - {\eta_2}^T \mu_1 + A(\eta_2) \quad (5)$$

$$CE = log(\lambda!) - \lambda * log\lambda_{pre} + \lambda_{pre} \quad (6)$$

5. Results and discussion

5.1. Performance evaluation & comparison between the classification model and the Poisson regression model

5.1.1 Performance evaluation: classical classification

The classic classification predictors of the m6A site were firstly generated on the whole exonic DRACH motif sites. AUROC and MCC were used as metrics to measure the accuracy of the classification models quantitatively, and the performance was evaluated using 5-fold cross-validation. The results are presented in Table 1. The classification model using Xgboost algorithm has 0.645 AUROC on the sequence-based feature, and AUROC is 0.854 on sequenced-based feature and genomic features, the AUROC improves again when adding the regulator features (AUROC of 0.883), so as the same trend on other algorithms, which derived the same conclusion as the WHISTLE got that the genomic feature will vastly improve the performance of the predictor. Since the performance has once again improved by a notch after combining the 17 m6A regulator features suggesting that the regulator may have a potential effect on m6A site methylation prediction. Furthermore, it is clear that the Xgboost (Highest AUROC 0.883) slightly outperformed than Random Forest (Highest AUROC 0.880), which was widely used in many classification predictors including WHISTLE (Chen et al., 2019) and SAMP (Zhou et al., 2016), this may be because that Xgboost works well on imbalanced data than Random Forest, when the model fails to predict abnormally for

the first time, it will be given more preference in the following iterations to adjust the prediction on low participation rate categories. (Both of them are ensemble methods that proved to generally outperform basic algorithms due to the robust accuracy of ensemble tree models.)

5.1.2 Performance evaluation: Poisson regression

Compared with the classical classification models, our Poisson regression models are the exact times the m6A site was detected to be methylated among 40 independent experiments. To quantitatively measure the accuracy of our Poisson regression models, we calculated Cross entropy (CE) and mean square error (MSE) to represent the consistency between the estimated distribution and the ground truth distribution (Table 2), while the ground truth was approximated using the Law of large number. The overall result shows that the Poisson regression model fit the dataset well, the RF slightly overperformed among those four algorithms with the minimum cross entropy of 0.503 (on the combination of whole features). Meanwhile, the results on feature improvement (genome feature improves the performance largely based on the sequence-based features and the regulator feature) were consistent with the classification models, again demonstrate the regulator feature is a strong predictor of m6A site.

5.2 Comparison between the classification model and the Poisson regression model

As each classification model has the corresponding regression model for which it is

established with the same features and machine learning algorithm, their performance can be compared together with the "decisions" made by our models. Comparing the Poisson regression model with the classification model using the same metrics (Table 3), shows that the regression model substantially outperformed the classification model on predicting the probability of the m6A site methylation. Besides, based on the metric cross entropy, the classification model improves slowly when adding the genome and regulator features. The value of MSE even becomes larger when adding those features, it can be concluded that while handling more features, the bias will also increase. Figure 2AB give an overall visualization of the prediction of both kind of models, it is evident that the Poisson regression model gives a better prediction. In contrast, the classification model generally deviates from the actual m6A site methylation preference, and the distribution of the log prediction is approximately uniform which is different from the real situation. Based on the result of the Poisson regression models with the best performance algorithms (Random Forest), the maps of RNA modification probability on the whole exonic DRACH motif are summarized in Supplementary information (Supplementary File S2) together with the genome information.

To further demonstrate the power of the Regression model in handling the multi-experiment datasets, we compared the predictability of two groups of models on the potential m6A site that at least methylated once. The predictive results, namely, the cross entropy and mean square error shown in Supplementary Table S3A (for Poisson regression models) and Supplementary Table S3B (for classification models), from

which we found that the predictive accuracies obtained by the best regression model (CE 1.868, MSE 4.352) outperformed the best classical classification models (CE 7.651, MSE 135.761). Interestingly, when more features were added, the performance of the models dropped instead (Supplementary Table S3B), the possible explanation is that when adding those features classification models, the bias of the prediction on at least one methylated site becomes larger. By centering the predicted probability to the original mean, the bias gets smaller (CE 2.706, MSE 17.369), which is shown in Supplementary Table S3C, but the overall result still cannot exceed the performance of the regression model (CE 1.868, MSE 4.352).

6 Model interpretation. Regression model (whole dataset, subdataset) comparison.

***For the sake of simplicity, we use "subdataset" to indicate the sites used for model building and testing are the sites that are detected to be methylated at least in one experiment, "whole dataset" indicates the sites used were the whole dataset.**

Our main goal is to discover potentially causal features associations that may guide methylation of the m6A site. Here we prefer models which are interpretable over models which might give relatively better performance. Both Poisson regression models (using subdataset and whole dataset) were taken for model interpretation. The Poisson models built on the whole dataset are mainly for predicting all the exonic DRACH motifs, including the potential m6A sites. The Poisson regression models built

on the subdataset gave insight into what features state combination will increase the probability that a site can be methylated to be detected under different parallel experiments or stably methylated under various biological conditions. Although it is still possible that the methylated sites have not been detected in 40 experiments, and the unmethylated sites may be false positives, these interferences can be ignored for large samples.

6.2.1 Directed interpretation

GLM models always seen as an interpretable model that can help uncover causal structure in observational data (Athey and Imbens, 2016). Figure 2CD shows the ranked top 7 most standard coefficient in three feature groups respectively from two GLM regression models (subdataset and whole dataset), where a positive coefficient indicates a positive relationship between the feature and the response (shown in red), while a negative coefficient indicates an increase in the feature corresponds with a decrease in the response (shown in blue). From Figure 2C, we can get that, genomic feature "FitCons_101bp" (standardized coefficients 0.528) has the most substantial relationship with the methylation on the m6A sites, followed by "3'UTR" (standardized coefficients 0.394) and "Long_exon" (standardized coefficients 0.322), those results show that the average *fitness consequence* (fitCons) scores within the flanking 50 bp region are the most profound feature related to the preference of the methylation of the potential m6A site, where fitCons was generated to measure the potential genomic function based on an evolutionary perspective (Gulko et al., 2015), here the strong

relation with the methylation may indicate the role of m6A methylation at the level of gene function. Moreover, if the candidate site is near the 3'UTR and within the long exons will increase the probability of methylation, the previous studies have also found that a large proportion of m6A residues are in the last exons allowing the 3'UTR regulation (Ke et al., 2015), validating its model interpretation capability. It is worth noting that in the ranking of the iRNA feature group, the significant position is centered on the m6A site, we can guarantee that the chemical property and functionality of the nucleotide will influence the methylation probabilities. Specifically, negative "FG_23" (standardized coefficients -0.463) and the positive "RS_23" (standardized coefficients 0.437) indicate that having Cytosine on the final position of the DRACH motif site is not favored by methylation. In addition, Figure 3EF give us a clear visualization of the how near positions relate or influence the methylation of the m6A site, it can be concluded that the methylation of the m6A will have a preference of the specific kind of ribonucleic acid kind, some kind of RNA will also have negative relation on m6A methylation, like the Cytosine on the final position of DRACH, and the results draw from the whole dataset and subdataset are quite similar. In addition, we also multiply each DRACH site with the frequency to be detected as methylated and draw the sequence logo on them (shown in Figure 2GH). When it comes to the regulator feature groups, METTL14 (standardized coefficients 0.256) has a significant positive relationship with methylation since METTLE14 is a component of the methyltransferase complex. Figure 3D shows the standard coefficient magnitudes for GLM regression model that builds on the site that has at least one methylated in 40

474 samples, the conclusions of iRNA features are consistent with the previous model, and
475 “FitCons_101bp” (standardized coefficients 0.109) and “Long_exon” (standardized
476 coefficients 0.102) are still significant in genomic feature group. The difference is that
477 the relative impact of the regulator on the methylation improved a lot compared with
478 the model build on the whole dataset, the METTL14 (standardized coefficients 0.163)
479 showing significant importance among all the features followed by regulator YTHDF1
480 (standardized coefficients 0.114) and YTHDF2 (standardized coefficients 0.107),
481 where YTHDF1 and YTHDF2 are m6A-related RBPs (RNA-binding proteins) that
482 selectively recognize m6A-modified mRNAs, YTHDF1 promotes protein translation
483 (Wang et al., 2015) and YTHDF2 reduces mRNA stability (Shi et al., 2017). The
484 Supplementary information (Supplementary Table S4; Supplementary Figure S1) gives
485 the standard coefficient magnitude value for each feature, and nearly all the regulators
486 show the positive effect on methylation. Although the GLM’s standardized coefficient
487 magnitudes maybe seem intuitive, they can be fragile concerning feature selection and
488 processing.
489
490 **6.2.2 Undirected interpretation**
491 The model interpretation on tree-based models RF and Xgboost was based on the
492 relative feature importance. Here the feature importance represents the statistical
493 significance of each variable in terms of its influence on the model based on how much
494 the squared error changed between that node and its children nodes (scaled to 100%).
495 The regression model builds on whole dataset and subdataset were compared (given

different lenses on interpretability). The bar plots in Figure 2I-L depict the feature importance of the Random Forest models, computed using the 5-fold cross-validation. For the Poisson regression model built on the whole dataset, Feature “HNRNPC” and “FitCons_101bp” are in the model (Figure 2I), while “Long_exon” and “Dist_3’sj” have the most contribution to the model prediction in the model built on the subdataset (Figure 2J). When combing with the complete m6A regulators, feature “YTHDF1” has the most contribution to the model prediction, followed by “YTHDF2”, “METTL14” in both models (Figure KL). Moreover, the standard deviation from the 5-fold shows the percentage of importance for each feature are quite “confident”. For another kind of tree-based model, Xgboost is shown in Supplementary information (Supplementary Figure S2).

7. Conclusion

Benefitting from the burst of high throughput epitranscriptomic data deposited on public resources, multiple machine learning site predictors were built up to predict the selectivity of m^6A methylation over DRACH motifs. However, almost all of the predictors developed are under the classification framework, which failed to represent the stoichiometry of the biological replicates. Furthermore, the sequence-derived features are widely used in RNA modification predictors and some bioinformatics databases, genomic-derived features including the binding sites of m^6A regulators are often ignored despite their substantial potential in the performance improvement.

Since the m^6A methylation is essentially a dynamic process, the building of classifiers over independent experiment or seeing at least methylated m6A site among several experiments as the true m6A site as previous studies did may not be the best solution. Therefore, by integrating and analyzing the high-throughput sequencing data, in the present work, we proposed Poisson regression models, using four algorithms (GLM, Deep learning, Xgboost, Random Forest) representing different levels of complexity to predict the possibility of methylation on whole exon DRACH motif. By combing the new 17 regulator features with the genome-derived features and sequence-derived features, the overall performance vastly improved compared with the previous work. Comparing the Poisson regression model with the corresponding classic classification on the same metrics, we can conclude that the Poisson regression model outperformed the classification model on overall performance. Moreover, the Poisson regressors gives a more comprehensive way to model the data from the multiple independent high-throughput sequencing experiments. On the perspective of model interpretation, three key features were identified that largely influence the probability that the site will be predicted as methylated. Therefore, it is anticipated that this approach will become a meaningful way to integrate the current base-resolution m6A data and give a feature interpretation on m6A which can be further applied in other post-transcriptional modifications or research fields.

8. Appendices

9. Acknowledgements

10. Reference

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

Athey, S., and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113**,** 7353-7360.

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 11**,** R90.

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., Su, J., De Magalhaes, J.P., Rigden, D.J., and Meng, J. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res* 47**,** e41.

Chen, T., and Guestrin, C. (Year). "Xgboost: A scalable tree boosting system", in: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*), 785-794.

Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: a web server for identification of N(6)-methyladenosine sites. *J Biomol Struct Dyn* 35**,** 683-687.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., Decaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L.K., Segler, M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., and Greene, C.S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 15.

Click, C., Malohlava, M., Candel, A., Roark, H., and Parmar, V. (2016). Gradient boosting machine with h2o. *H2O. ai* 11**,** 12.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489**,** 57-74.

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

Cui, X., Zhang, L., Meng, J., Rao, M.K., Chen, Y., and Huang, Y. (2018). MeTDiff: A Novel Differential RNA Methylation Analysis for MeRIP-Seq Data. *IEEE/ACM Trans Comput Biol Bioinform* 15**,** 526-534.

Frye, M., Harada, B.T., Behm, M., and He, C. (2018). RNA modifications modulate gene expression during development. *Science* 361**,** 1346-1349.

Garcia-Campos, M.A., Edelheit, S., Toth, U., Safra, M., Shachar, R., Viukov, S., Winkler, R., Nir, R., Lasman, L., Brandis, A., Hanna, J.H., Rossmanith, W., and Schwartz, S. (2019). Deciphering the "m(6)A Code" via Antibody-Independent Quantitative Profiling. *Cell* 178**,** 731-747 e716.

Gruber, A.R., Bernhart, S.H., and Lorenz, R. (2015). The ViennaRNA web services. *Methods Mol Biol* 1269**,** 307-326.

Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* 47**,** 276-283.

Hodges, J.L., and Lecam, L. (1960). The Poisson Approximation to the Poisson Binomial-Distribution. *Annals of Mathematical Statistics* 31**,** 737-740.

Huang, H., Weng, H., and Chen, J. (2020). The Biogenesis and Precise Control of RNA m(6)A Methylation. *Trends Genet* 36**,** 44-52.

Huang, H., Weng, H., Zhou, K., Wu, T., Zhao, B.S., Sun, M., Chen, Z., Deng, X., Xiao, G., Auer, F., Klemm, L., Wu, H., Zuo, Z., Qin, X., Dong, Y., Zhou, Y., Qin, H., Tao, S., Du, J., Liu, J., Lu, Z., Yin, H., Mesquita, A., Yuan, C.L., Hu, Y.C., Sun, W., Su, R., Dong, L., Shen, C., Li, C., Qing, Y., Jiang, X., Wu, X., Sun, M., Guan, J.L., Qu, L., Wei, M., Muschen, M., Huang, G., He, C., Yang, J., and Chen, J. (2019). Histone H3 trimethylation at lysine 36 guides m(6)A RNA modification co-transcriptionally. *Nature* 567**,** 414-419.

Jiang, X., Liu, B., Nie, Z., Duan, L., Xiong, Q., Jin, Z., Yang, C., and Chen, Y. (2021). The role of m6A modification in the biological functions and diseases. *Signal Transduct Target Ther* 6**,** 74.

Ke, S., Alemu, E.A., Mertens, C., Gantman, E.C., Fak, J.J., Mele, A., Haripal, B., Zucker-Scharff, I., Moore, M.J., Park, C.Y., Vagbo, C.B., Kussnierczyk, A., Klungland, A., Darnell, J.E., Jr., and Darnell, R.B. (2015). A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev* 29**,** 2037-2053.

Kennedy, E.M., Bogerd, H.P., Kornepati, A.V., Kang, D., Ghoshal, D., Marshall, J.B., Poling, B.C., Tsai, K., Gokhale, N.S., Horner, S.M., and Cullen, B.R. (2016). Posttranscriptional m(6)A Editing of HIV-1 mRNAs Enhances Viral Gene Expression. *Cell Host Microbe* 19**,** 675-685.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol* 9**,** e1003118.

Linder, B., Grozhik, A.V., Olarerin-George, A.O., Meydan, C., Mason, C.E., and Jaffrey, S.R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 12**,** 767-772.

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., Dai, Q., Chen, W., and He, C. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat Chem Biol* 10**,** 93-95.

Lou, X., Wang, J.J., Wei, Y.Q., and Sun, J.J. (2021). Emerging role of RNA modification N6-methyladenosine in immune evasion. *Cell Death Dis* 12**,** 300.

624 Meng, J., Lu, Z., Liu, H., Zhang, L., Zhang, S., Chen, Y., Rao, M.K., and Huang, Y.
625 (2014). A protocol for RNA methylation differential analysis with MeRIP-Seq
626 data and exomePeak R/Bioconductor package. *Methods* 69**,** 274-281.
627 Meyer, K.D. (2019). DART-seq: an antibody-free method for global m(6)A detection.
628 *Nat Methods* 16**,** 1275-1280.
629 Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C.E., and Jaffrey, S.R.
630 (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3'
631 UTRs and near stop codons. *Cell* 149**,** 1635-1646.
632 Mitchell, R., and Frank, E. (2017). Accelerating the XGBoost algorithm using GPU
633 computing. *PeerJ Computer Science* 3**,** e127.
634 Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D.,
635 Hall, A.J., Barton, G.J., and Simpson, G.G. (2020). Nanopore direct RNA
636 sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A
637 modification. *Elife* 9.
638 Roundtree, I.A., Evans, M.E., Pan, T., and He, C. (2017). Dynamic RNA Modifications
639 in Gene Expression Regulation. *Cell* 169**,** 1187-1200.
640 Schwartz, N.B., Galligani, L., Ho, P.L., and Dorfman, A. (1974). Stimulation of
641 synthesis of free chondroitin sulfate chains by beta-D-xylosides in cultured cells.
642 *Proc Natl Acad Sci U S A* 71**,** 4047-4051.
643 Shi, H., Wang, X., Lu, Z., Zhao, B.S., Ma, H., Hsu, P.J., Liu, C., and He, C. (2017).
644 YTHDF3 facilitates translation and decay of N(6)-methyladenosine-modified
645 RNA. *Cell Res* 27**,** 315-328.
646 Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K.,
647 Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson,
648 R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. (2005).
649 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
650 genomes. *Genome Res* 15**,** 1034-1050.
651 Wainberg, M., Merico, D., Delong, A., and Frey, B.J. (2018). Deep learning in
652 biomedicine. *Nat Biotechnol* 36**,** 829-838.
653 Wang, X., Zhao, B.S., Roundtree, I.A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K.,
654 Shi, H., and He, C. (2015). N(6)-methyladenosine Modulates Messenger RNA
655 Translation Efficiency. *Cell* 161**,** 1388-1399.
656 Xiang, Y., Laurent, B., Hsu, C.H., Nachtergaele, S., Lu, Z., Sheng, W., Xu, C., Chen,
657 H., Ouyang, J., Wang, S., Ling, D., Hsu, P.H., Zou, L., Jambhekar, A., He, C.,
658 and Shi, Y. (2017). RNA m(6)A methylation regulates the ultraviolet-induced
659 DNA damage response. *Nature* 543**,** 573-576.
660 Yang, H.D., and Nam, S.W. (2020). Pathogenic diversity of RNA variants and RNA
661 variation-associated factors in cancer development. *Exp Mol Med* 52**,** 582-593.
662 Yao, L., Yin, H., Hong, M., Wang, Y., Yu, T., Teng, Y., Li, T., and Wu, Q. (2021). RNA
663 methylation in hematological malignancies and its interactions with other
664 epigenetic modifications. *Leukemia* 35**,** 1243-1257.
665 Zhang, Z., Chen, L.Q., Zhao, Y.L., Yang, C.G., Roundtree, I.A., Zhang, Z., Ren, J., Xie,
666 W., He, C., and Luo, G.Z. (2019). Single-base mapping of m(6)A by an
667 antibody-independent method. *Sci Adv* 5**,** eaax0250.

668 Zhou, Y., Zeng, P., Li, Y.H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of
669 mammalian N6-methyladenosine (m6A) sites based on sequence-derived
670 features. *Nucleic Acids Res* 44**,** e91.

671

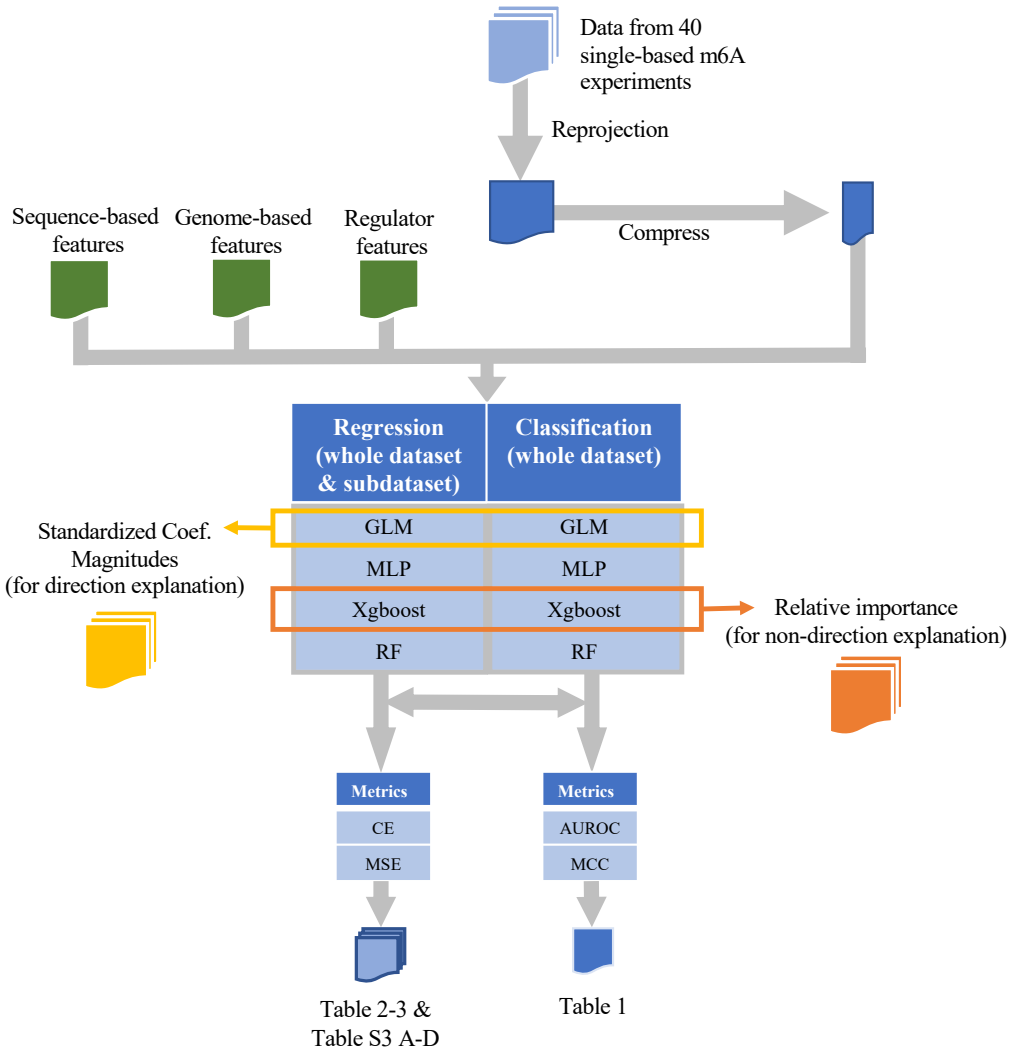

Figure 1 . The straightforward computational framework of this project.

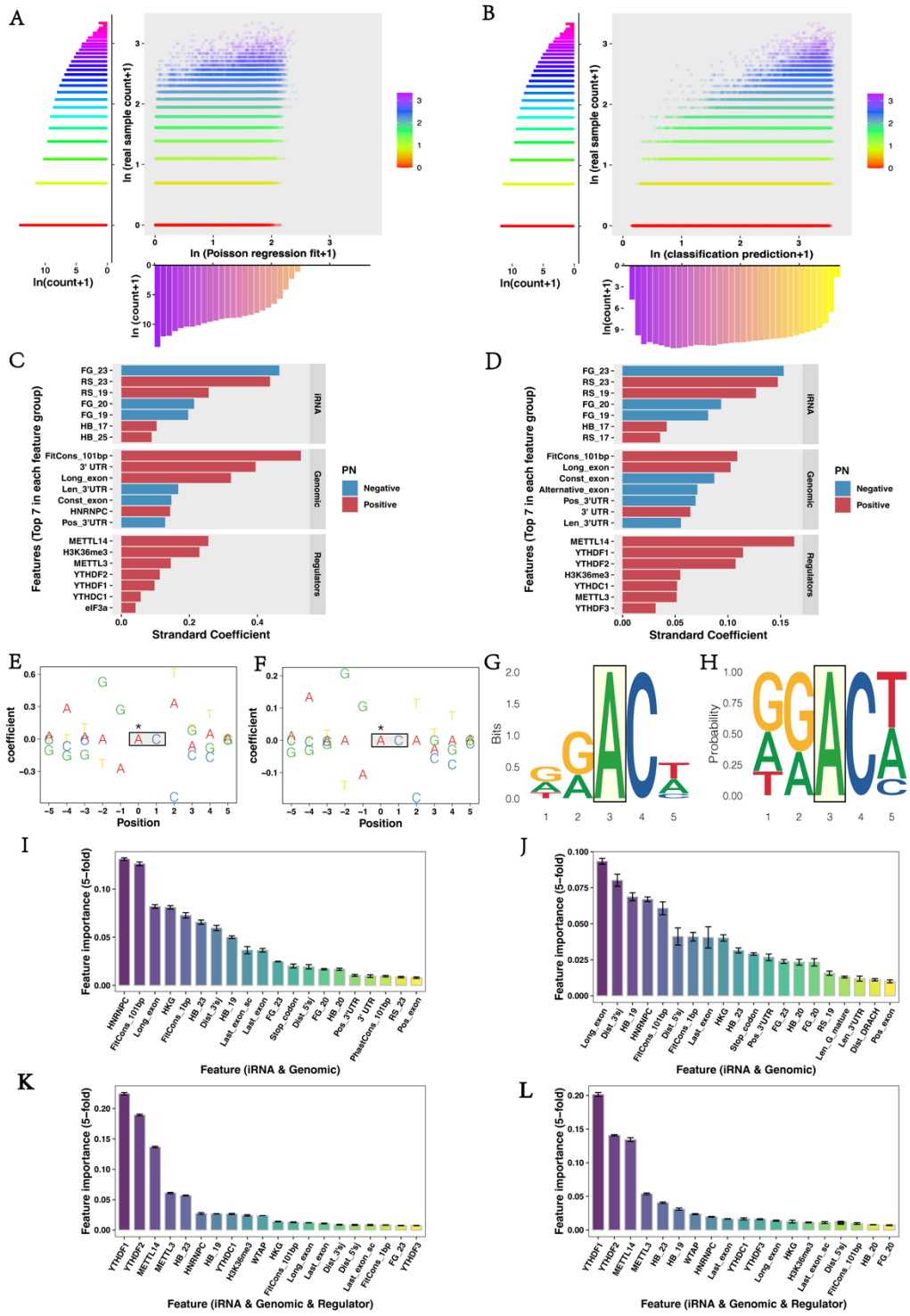

Figure 2.

(A)(B) Two figures show the relationship between the real sample count and the model prediction (Figure 2A for RF Poisson regression model, Figure 2B for RF classification

model). Upright is the main scatter plot, the color of the dot shows the value of the ln (real sample count + 1) corresponding to the left figure, the length of those colorful bar (ln(count +1)) in the left figure indicate the distribution of the ln(real sample count + 1). For the downright figure length of the bars shows the distribution of the ln (prediction value + 1), the meaning of two auxiliary figures is to Visualize the frequency of points at overly dense locations in a scatter plot. (C) The plot about Standard Coefficient Magnitudes for GLM regression model builds on the whole dataset. (D) The plot about Standard Coefficient Magnitudes for GLM regression model builds on the subdataset. (E) (F)Two figures depict the relative contribution of different nucleotides around the m6A site to the methylation. X-axis shows the relative position concerning the m6A site (x=0). Figure 2E is results of the Poisson regression model built on the whole dataset, while Figure 2F is from the regression model built on the subdataset, both models only used one hot-encoding feature. (G) (H) Sequence logos for sites identified from 40 single-based m6A experiments (According to the frequency). X-axis shows the relative position with respect to the m6A site (x=3). (I) (J) (K) (L)Feature importance for Random Forest Poisson Regression models, x-axis indicates the features used for model building, Figure IK are Poisson regression models built on the whole dataset, Figure JL are Poisson regression models on the subdateset.

* For the detailed meaning of features used in Figure G-L are shown below. For “iRNA” features, “RS here indicates “two ring structure”, here indicate Adenin and Guanineu . “HB” indicate “hydrogen bond”, here indecates Cytosine and Guanine. “FG” indicate “Functional group”, here indicates Cytonine and Adenine. For genomic feature, 3’UTR

is a dummy variable indicating whether the site overlaps with the 3'UTR, Alternative_exon and Const_exon indicated whether the site is on the alternative exon region or constitutive exon region. Long_exon indicates whether the site is on the exon that is longer than 400bp. Pos_3'UTR indicate the relative position on 3'UTR. Dist_3'sj indicates the distance between the site to the 5'splicing junction. Len_3'UTR is the length of the 3'UTR. FitCons_1bp indicates the FitCons scores of the nucleotide and FitCons_101bp indicate the average Fitcons scores within the flanking 50bp region.

Table 1. Performance evaluation of the four classical classification model on whole DRACH sits

Feature	Metrics	Four classificaiton models: Generalized Linear Model	Multi-layer Perceptron	Xgboost	Random Forest
iRNA	AUROC	0.644	0.641	0.645	0.647
	MCC	0.133	0.130	0.135	0.136
iRNA&Genomic features*	AUROC	0.838	0.820	0.854	0.847
	MCC	0.388	0.376	0.410	0.399
iRNA&Regulators*	AUROC	0.839	0.831	0.839	0.839
	MCC	0.431	0.424	0.421	0.421
iRNA&Genomic features&Regulators	AUROC	0.881	0.861	0.883	0.880
	MCC	0.473	0.463	0.469	0.459

Table 2. Performance evaluation of the four Poisson regression models on the whole DRACH motif sits

Feature	Metrics	Four Poisson Regression Models			
		Generalized Linear Model	Multi-layer Perceptron	Xgboost	Random Forest
iRNA	MSE	1.319	1.322	1.325	1.318
	CE	0.819	0.823	0.830	0.817
iRNA+Genomic features	MSE	1.106	1.153	1.054	1.020
	CE	0.609	0.659	0.616	0.579
iRNA+Regulators	MSE	1.091	0.993	0.932	0.905
	CE	0.568	0.582	0.580	0.543
iRNA+Genomic features+Regulators	MSE	1.050	1.038	0.856	0.855
	CE	0.517	0.571	0.526	0.503

Table 3. Performance evaluation of the four classical classification models with the regression metrics

		Four Classification Models			
Feature	Metrics	Generalized Linear Model	Multi-layer Perceptron	Xgboost	Random Forest
iRNA	MSE	23.179	22.653	32.459	21.182
	CE	4.302	4.262	5.476	4.313
iRNA+Genomic features	MSE	47.617	48.025	41.345	30.587
	CE	4.099	4.126	5.329	4.138
iRNA+Regulators	MSE	53.281	52.825	44.361	35.747
	CE	4.048	4.012	5.282	4.089
iRNA+Genomic features+Regulators	MSE	60.939	60.731	49.606	38.834
	CE	4.002	4.034	5.246	4.049