Machine learning algorithm for precise prediction of 2'-O-methylation (Nm) sites from experimental RiboMethSeq datasets

Florian Pichot, Virginie Marchand, Mark Helm, Yuri Motorin

▶ **To cite this version:**

Florian Pichot, Virginie Marchand, Mark Helm, Yuri Motorin. Machine learning algorithm for precise prediction of 2'-O-methylation (Nm) sites from experimental RiboMethSeq datasets. Methods, 2022, 203, pp.311-321. 10.1016/j.ymeth.2022.03.007 . hal-03616347

HAL Id: hal-03616347

https://hal.univ-lorraine.fr/hal-03616347

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S1046202322000743
Manuscript_9c703aa326ed7c00efeb3061c9dc93df

prepared for METHODS special issue Bioinformatics, revised

Machine learning algorithm for precise prediction of 2'-O-methylation (Nm) sites from experimental RiboMethSeq datasets

Florian Pichot[1,2], Virginie Marchand[2], Mark Helm[1] and Yuri Motorin[2,3]

[1] Institute of Pharmacy of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Mainz, Germany.

[2] Université de Lorraine, CNRS, INSERM, UAR2008/US40 IBSLor, EpiRNA-Seq Core facility, F-54000 Nancy, France.

[3] Université de Lorraine, CNRS, UMR7365 IMoPA, F-54000 Nancy, France.

* corresponding author: Yuri MOTORIN

Email: Yuri.Motorin@univ-lorraine.fr

Keywords: deep sequencing, RNA modification, 2'-O-methylation, machine learning, Random Forest

© 2022 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

Abstract

Analysis of epitranscriptomic RNA modifications by deep sequencing-based approaches brings an essential contribution to the general knowledge on their precise locations and relative stoichiometry in cellular RNAs. To reveal RNA modifications, several analytical approaches have been proposed, including antibody-driven enrichment, analysis of RT-signatures and specific chemical treatments. However, analysis and interpretation of these massive datasets, especially for low abundant cellular RNAs (e.g. mRNA and lncRNA) is not easy nor straightforward, since the insufficient specificity and selectivity are leading to massive false-positive and false-negative identifications. The main issue in the application of these methods relies on a subjective classification of potentially modified positions, mostly based on arbitrarily defined threshold values for different scores. Such approach using pre-defined scores' values was revealed to be appropriate for limited complexity datasets (for tRNA and/or rRNA analysis), but application to longer reference sequences requires much better classification algorithms. In this work we applied a machine learning algorithm (Random Forest, RF) to create a predictive model for analysis of 2'-O-methylated sites in RNA using RiboMethSeq datasets. Model's training was performed on a large collection of human rRNA datasets with well-known modification profiles and the performance of the prediction was assessed using experimentally defined profiles for other eukaryotic rRNAs (*S. cerevisiae* and *A. thaliana*). Application of this Random Forest prediction model for detection of other RNA modifications and to more complex datasets is discussed.

Introduction

Mapping and precise quantification of epitranscriptomic RNA modifications are essential to better understand their functions in different steps of RNA metabolism. RNA modifications were shown to regulate RNA stability, RNA processing and translation [1–11]. Amongst approaches, capable of detecting RNA modifications, the deep sequencing-based protocols now play a major role [2,12–16]. Only these techniques do not need purification of totally homogeneous RNA prior to analysis and, with optimized protocols of library preparation, require only a tiny amount of input RNA. Due to these essential features, deep sequencing is now widely used as a gold standard method for detection, mapping and quantification of RNA modified residues [15,16].

However, such high-throughput analysis of complex reference sequences precludes site-by-site validation of the potential modification candidates, and thus such approaches should have extraordinary specificity and precision, to provide reliable RNA modification profiles in an automatic bioinformatic pipeline [17–21]. Even a relatively low False Discovery Rate (FDR) will generate hundreds or thousands of false positive identifications, even when applied to transcriptomes of 10^6-10^7 nucleotides. Considering this, application of simple thresholds for detection of modifications in highly abundant and extensively modified RNA (like rRNAs) is still possible but becomes of limited value when analysis should be performed for sparsely (and, in addition, only partially) modified RNA species.

Alkaline hydrolysis-based RiboMethSeq protocol is now extensively used for quantification of RNA modifications (mostly in rRNA and tRNAs) [14,22–26], and, in some instances, was also applied to establish *de novo* modification profiles. Such *de novo* mapping was recently performed for *A. thaliana* and *X. laevis* rRNA [27,28], and for FTSJ3-dependent Nm modifications in HIV-1 RNA [24]. In all these cases, multiple false positive hits were detected, and additional validation by complementary approaches was necessary to establish a reliable Nm modification pattern.

In the current model used for prediction of Nm modification sites in unknown RNA, only two major RiboMethSeq scores [29–31] define classification of a potential candidate site. ScoreMean provides sensitivity, but has somehow a limited specificity, while ScoreA displays the opposite behavior (**Figure 1**). Thus, in practice, the best compromise in detection of modified positions is defined as a combination of two thresholds (e.g. ScoreMean > 0.92 and ScoreA > 0.5), this combination was determined through the previous analysis of ROC curves [29]. In most cases, this default combination of scores ensures efficient detection of highly modified Nm sites (observed MethScore=ScoreC > 0.85) but shows only a limited performance in prediction of partially modified positions (with MethScore < 0.75). Additional RiboMethSeq scores [32,33] (ScoreB and MethScore) are rarely used for *de novo* mapping since they show only limited discrimination of modified and

unmodified positions. Thus, to improve both specificity and sensitivity of Nm mapping from experimental RiboMethSeq datasets [34], more advanced approaches are required, ideally including combination of different scores and other features of the analyzed RNA sequence. In this perspective, the use of machine learning algorithms may fulfill criteria for more accurate approaches.

Figure 1 Scatter plot for scores mean and A2 for randomly selected RiboMethSeq sample of human 28S rRNA. Unmodified residues (A, C, G and U) are shown as dots with pastel color, known modified sites are in solid color. Commonly used combination of thresholds for score A2 (> 0.5) and score mean (> 0.92) is shown as dashed red lines. Positions holding exceeding values for both scores are considered to be 2'-O-methylated candidates.

Machine learning algorithms' applications for data treatment are recurrent, especially in biomedical/clinic research fields, due to very high global volume of data to be analyzed. Depending on the goal behind data analysis, multiple machine learning algorithms offering different approaches can be used. Unsupervised learning algorithms have proved to be able to discover new mechanisms of pathology through studious analyses of genomic and transcriptomic data [35–37]. Supervised learning algorithms, like PSI-Sigma or DeepM6ASeq [38,39], are used in epitranscriptomics for modification detection, on the basis of the sequence context and secondary structure predictions. Such algorithms are also used in Nanopore sequencing for base calling and modification predictions [40,41]. Such tools could also be applied to chemically treated or pre-enriched RNAs, like in the case of Bisulfite sequencing, MeRIP-seq or in this work, to RiboMethSeq.

Kernel machines such as support-vector machines (SVM), decision trees with Random Forest (RF) and artificial neural networks seem to be suitable for prediction of modified RNA sites from experimental datasets. Artificial neural networks is the main model used in Nanopore sequencing technology [42–44] due to its capacity to mimic brain behavior for decision making. However, complexity of such models makes model's prediction traceback difficult. SVM allow a clear view on how the predictions are done, and tuning the prediction performances can be done easily, but the kernel trick approach shows weaknesses for noisy data and overlapping classes. RF, based on predictions by multiple decision trees, makes the prediction procedure clear enough to understand how it is achieved, shows better accuracy than most of the other classification algorithms while providing probabilistic estimates through trees classification votes' rate. This tool has already been used for modification predictions by analysis of RT signatures [45,46]. However, RF is also known to be highly biased if incorrectly configured. Out of these three approaches, RF seems to be the best algorithm for analysis of RiboMethSeq data and prediction of the potentially modified Nm sites.

In this manuscript we describe application of RF prediction model for a comprehensive analysis of RiboMethSeq datasets.

Materials and Methods

RiboMethSeq datasets used in the study

RF models have been created and tested on 62 complete wild-type human rRNA samples (446 214 nucleotide positions) from different cell cultures (Table S1, [34], ENA accession n° PRJEB46620). To reduce biases of sample selection, 80% of the data (356 970 of randomly selected sites) has been used as a training dataset, while the last 20% (89 243 sites) served as an internal testing dataset for initial assessment of models' performance.

RiboMethSeq datasets for yeast *S. cerevisiae* (6 datasets) and plant *A. thaliana* (3 datasets) have been described previously [22,27] and available in public databases (ENA accession n° PRJEB35565). Data storage and analyses have been handled by a Unix server Dell R740 under Ubuntu LTS 20.4. Raw data have been trimmed by Trimmomatic version 0.32/0.39 and aligned by Bowtie2 version 2.2.2/2.2.4 as described [22,27,34].

All calculations for RF application and model analysis has been done on RStudio Server Version "Juliet Rose" 1.4.1717 with R version “Bird Hippie“ 4.1.2. Specific data transformation required R package “reshape2” version 1.4.4, RF algorithm has been provided by the R package “randomForest” version 4.6-14 and graphics have been produced by R packages “ggplot2” version 3.3.5 and “RColorBrewer” version version 1.1-2.

Results

Brief outline of RiboMethSeq analysis

Analysis of Nm sites in RNA by RiboMethSeq is essentially based on a relative protection against alkaline hydrolysis, which can be provided by different factors, including adjacent 2’-O-methylation. The replacement of 2’-OH by 2’-O-CH_3 very substantially reduces the nucleophilicity of the 2'-oxygen atom in ribose, thus reducing the cleavage of the 3’-adjacent phosphodiester bond in the RNA sequence. This protection is not fully quantitative, but 95-99% reduction of the cleavage rate is generally observed. Partial methylation provides only partial protection, therefore the modification rate can be rather precisely defined [47]. In a randomly cleaved RNA, fragments starting at N+1 position and ending exactly at an Nm residue are substantially under-represented. This under-representation is detected and quantified by deep sequencing, after conversion of the RNA fragments to sequencing library and sequencing in single read (generally SR50) or paired-end mode. Both count profiles for 5’-reads’ extremities at N+1 and 3’-ends at N positions show a characteristic coverage gap at the position of modification. Initial versions of RiboMethSeq analysis mainly used 5’-end count showing typical N+1 nt gap [22,47], but cumulated 5’/3’-ends’ count provides a better discrimination of modified positions and more precise quantification of the modification rate [29]. During merging of 5’-end and 3’-end counts, the -1 backshift for 5’-ends is introduced, to align the gap exactly to the position of Nm residue in RNA (**Figure 2**).

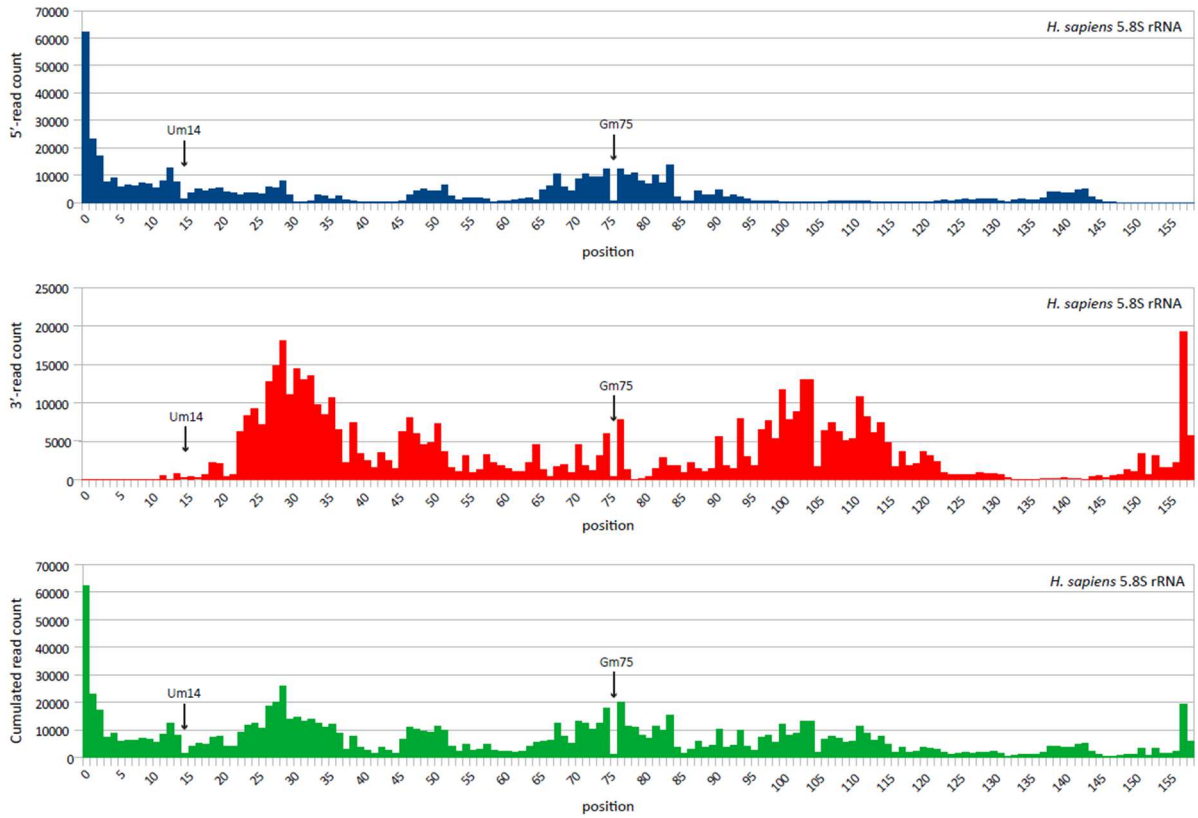

Figure 2 Typical RiboMethSeq profiles for human 5.8S rRNA. Top panel - 5'-end count (blue), middle panel - 3'-end count (red) and merged profile (bottom, green). Positions of two known Nm residues (Um14 and Gm75) in 5.8S rRNA are shown. Numbering is in 0->n-1 format (from *.bed format), so Um14 and Gm75 signals appear at the nucleotide positions N and not at N+1.

Parameters and scores extracted from RiboMethSeq dataset

The depth and other characteristic features of the Nm-related gap in cumulated profile is measured using several RiboMethSeq specific scores. Construction of this scoring system was previously described in details [22,29,32] and outlined in **Supp Figure S1**. Since the random alkaline cleavage profile for real RNA is not perfectly regular, some nucleotide positions may show partial protection, related to multiple factors, such as residual RNA structure under alkaline hydrolysis, unequal cleavage at different nucleotides or nucleotide's environments, as well as to biases related to preferential adapter ligation to RNA fragments and PCR amplification during library preparation. At the clustering and sequencing steps, some sequence contexts may be under- or over-represented, and exclusion of ambiguously mapped reads also alters the observed ends' coverage. All these factors collectively contribute to irregular cleavage profiles obtained for different RNA species. The easiest and the most straightforward parameter indicating a potential Nm residue in RNA sequence might be global protection from cleavage, compared to other neighboring positions.

However, in practice, this global (raw) protection is not a sufficiently robust criteria for precise discrimination between irregular cleavage and the presence of Nm residue.
Considering various factors contributing to irregularities of RNA cleavage, RiboMethSeq scores are designed to include the cleavage profile at the neighboring positions in the value calculated for a given nucleotide. Initial protocol arbitrarily suggested inclusion of positions from -6 to +6, albeit with reduced weight coefficients, varying from 1.0 to 0.5, depending on their relative distance to the Nm site [32]. Later, thoughtful analysis of different calculation intervals demonstrated that the optimal distance, providing the best discrimination of Nm positions and their quantification is +/-2 nt (instead of +/-6 nt used previously) [29]. Thus, all scores used in this work were calculated for +/-2 nt interval with weight coefficients 1 (pos -1/+1) and 0.9 (pos -2/+2). In addition to the classical RiboMethSeq scoring system [29,32], an additional ScoreAngle was suggested [26]. This score represents only the depth of the gap, without considering other neighboring positions in the profile. Thus, ScoreAngle provides very high sensitivity, but relatively low specificity, since any accidental drop in the coverage profile may be considered as a valid signal. Other RiboMethSeq scores (ScoreA, ScoreB and ScoreC=MethScore) are less sensitive to position-to-position variations, since they consider global variations of the coverage in a given region.

Discrimination of Nm from other modified nucleotides

An additional layer of complexity in the analysis of RiboMethSeq data, especially for eukaryotic rRNA, is related to the false positive (FP) signals raised by other modified nucleotides, mostly pseudouridines (ψ), which are also rather frequent in these RNA species. The exact origin of such non-Nm signals observed in RiboMethSeq remains unclear, it was suggested that they may appear due to adapter ligation biases [22,32]. Similar behavior was also noticed for m^7G (modification common in tRNAs) [14] and, in part, for $m^6{}_2A$ (a highly conserved double methylation found at the 3'-end of 16S/18S rRNA). The signal for $m^6{}_2A$ may be related to both ligation bias and to strong RT-arrest signal during primer extension. However, other RT-arresting RNA modifications (like m^1A, m^1G, m^3U, $m^1acp^3\psi$, …) were not shown to generate sharp RiboMethSeq signals, but rather progressively reduce global coverage in the region (~20 nt) upstream of such modification. RT-silent modified residues (m^5C, m^2G, ac^4C, …) are not visible in RiboMethSeq protection profiles. Considering these points, the discrimination model for prediction of Nm sites should include also other potential RNA modifications (at least pseudouridines), especially if prediction is done for eukaryotic rRNA, rich in such modified residues.

Construction of the predictive Random Forest model

In order to include in the Random Forest (RF) model the maximum of information about the RNA sequence as well as on other features of the cleavage profile, the parameters of different nature were retained for model construction. We considered sequence features (nature of the base included as "base" parameter and identity of -6/+6 surrounding bases), calculated RiboMethSeq scores (in -2/+2 neighbourhood) and score-related values for a given position (ratioL/R, aroundL/R, protection, scores, etc, see **Supp Figure S1** for more details on all these scores and related values), scores calculated for neighboring positions (called prevScores) and auxiliary information to annotate low coverage regions (see **Table 1** for a full list). Three main RF models were created, and their performance evaluated: "Full model" including all possible features, "no_Surrounding_base_info" model retaining all scores but omitting sequence information (except the identity of the modified residue), and "no_Surrounding_info" model where both sequence and scores for neighboring positions were omitted.

Importance of each parameter was evaluated using different RF model trained on a large collection of experimentally obtained RiboMethSeq datasets for human rRNA, and considering possibility to predict Nm and pseudouridines only, as well as other modified nucleotides found in human rRNA (separated in two groups: RT-arresting and RT-silent RNA modifications).

Overall inspection of models' performance in prediction of Nm and ψ residues, as well as RT-arresting and RT-silent residues (see **Supp Figures S2-S6**) pointed out the importance of both RiboMethSeq scores and sequence information for prediction of modifications. However, for prediction of Nm and ψ residues the nature of the modified nucleotide (base) plays a major role in discrimination (**Supp Figure S2**), followed by values of RiboMethSeq scores, while the sequence context brings only a minor contribution. In contrast, the prediction of RT-arresting and RT-silent nucleotides is mostly based on surrounding sequence features, since these modifications are rare in rRNA and found at conserved sequences in all samples in the dataset.

The performance of different models was evaluated by "precision" and "sensitivity" parameters (**Supp Figure S3**). Precision was calculated as ratio of true positives to total predicted positives, while sensitivity represents the ratio of true positives to total (actual) positives in the data. "Full model" and "no_Surrounding_base_info" models show very similar precision, while sensitivity for low and variably modified positions decreases without surrounding sequence information, since these modifications with low RiboMethSeq scores are mostly predicted in a sequence context, like RT-silent and RT-arresting residues.

The same conclusion can be drawn from inspection of prediction frequency (**Supp Figure S4**); the "Full model" model reliably predicts both highly modified and invariable residues

(Am/Cm/Gm/Um/ψm) and also variably modified and low modified/almost unmodified positions (vAm/vCm/vGm/vUm/_Cm/_Um/_Gm). However, when surrounding sequence information is not included (model "no_Surrounding_base_info"), prediction of highly modified positions is only minorly affected, while for others classification into unmodified group substantially increases. This clearly indicates that those variable and low-modified residues (as well as RT-silent and most of RT-arresting modifications) are classified mostly relying on the sequence context, and not according to the RiboMethSeq signals. Such prediction using the conserved cleavage profile around the modified site as prediction criteria also explains very good results in prediction of Ψm residues in the internal human dataset.

Taking into account these observations, the model excluding sequence related information for the neighboring positions ("no_Surrounding_base_info") shows the best results in prediction of Nm and Ψ residues and thus was investigated in more details. The model excluding all sequence and score-related information on the neighbors of the modified nucleotide ("no_Surrounding_info") showed considerably degraded performance compared the two others.

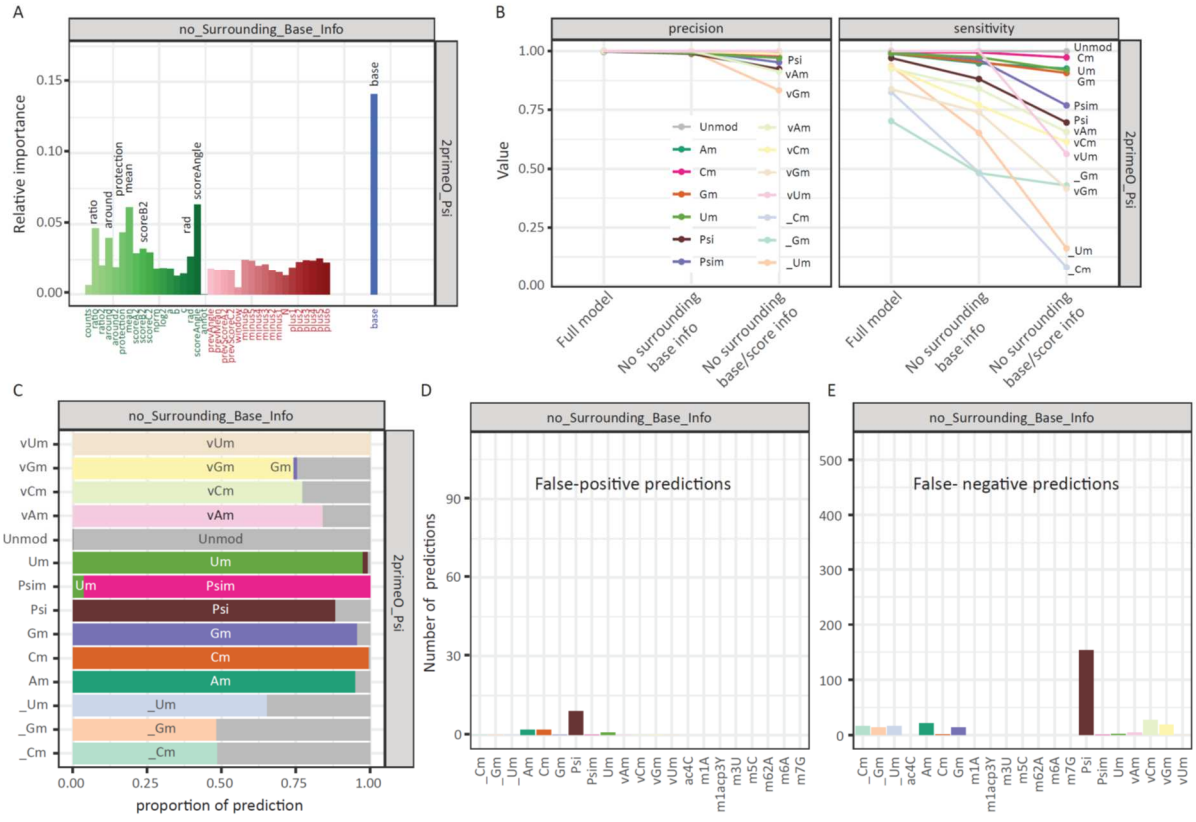

Figure 3 Importance of RF model's parameters and performance of "no_Surrounding_Base_info" model in detection of Nm and pseudouridine (Psi/ψ) residues (other models and their performance are illustrated in **Supp Figures S2-S6**). **Panel A** – Relative importance for each feature has been calculated by normalizing Gini importance (or mean decrease impurity) across all features. **Panel B**

shows the influence of the RF model on the precision (left) and sensitivity (right) of Nm and ψ detection. Identity of the models is shown on the bottom (Full model, no surrounding base info, no surrounding info). **Panel C** shows the results of predictions for confirmed and invariable Nm residues (Am, Cm, Gm, Um), variable positions Nm (vAm, vCm, vGm, vUm) and some low modified positions (_Um, _Gm, _Cm). No low modified _Am residues were present in the training dataset. Unmod stands for unmodified nucleotide in the sequence. **Panels D** and **E** show distribution of False-positive (FP) and False-negative (FN) hits for predictions using internal training dataset (20% of human RiboMethSeq datasets kept aside for validation, see Materials&Methods).

The behavior of the "no_Surrounding_base_info" model is illustrated in **Figure 3**. As mentioned above, the nature of the base, the ratioL/aroundL values as well as protection, ScoreMean and ScoreAngle play major roles in discrimination (**Figure 3A**). Precision of the model is very high, the value is very close to 1.0 and does not differ much from the "Full model" for Nm prediction, while sensitivity remained high for constitutively modified sites (**Figure 3B**). Proportion of correctly predicted Nm and Ψ positions was high for constitutive rRNA modifications and moderate for variably and low modified sites (annotated as vNm and _Nm, respectively) (**Figure 3C**). Analysis of FP hits showed a very low number of incorrectly predicted locations (~10 unmodified positions were predicted as Ψ residues for the whole dataset) **(Figure 3D**). Moreover, FN predictions were also moderate, with <10 Nm which were not predicted, however this proportion is much higher for Ψ residues (**Figure 3E**). Overall, considering the dataset used for validation (20% of total dataset, so ~ 13 samples), the "no_Surrounding_base_info" model showed almost exceptional performance in the Nm detection and discrimination from Ψ residues.

Validation of the predictive RF model using *S. cerevisiae* and *A. thaliana* RiboMethSeq datasets

Since the use of identical sequences (here *H. sapiens* rRNA datasets) for both training and performance measurements is highly biased by predictions based on "known sequence" and "known profile", the performance of the RF models have been also assessed using additional RiboMethSeq datasets, for relatively distant rRNAs from *S. cerevisiae* and from *A. thaliana*.

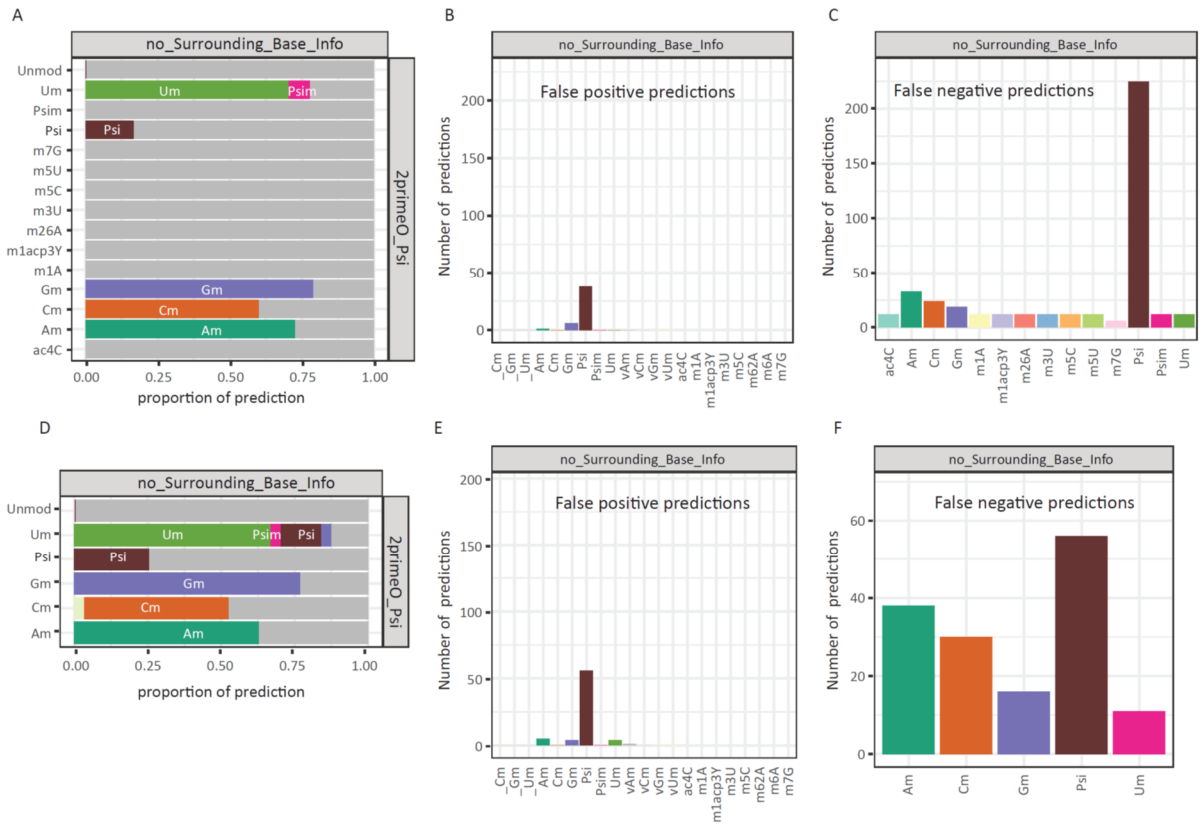

Figure 4 Validation of RF models using RiboMethSeq datasets for *S. cerevisiae* and *A. thaliana* rRNAs. Only Nm and Ψ residues were reliably mapped in *A. thaliana* rRNAs and this limited modification profile was used as a reference. Results of prediction for Nm, ψ and other RNA modifications are shown for "no_Surrounding_Base_info" model applied for *S. cerevisiae* (**A**) and *A. thaliana* (**D**) RiboMethSeq datasets. Identity of prediction is shown by color. **Panels B** and **F** show FP hits for *S. cerevisiae* (**B**) and *A. thaliana* (**E**) datasets, while **panels C** and **F** list FN predictions for both.

Figure 4 demonstrates that analysis of *S. cerevisiae* and *A. thaliana* RiboMethSeq datasets gives very similar results for "no_Surrounding_Base_info" RF model, even with relatively different rRNA sequences. Substantial proportion of Nm residues was correctly predicted (**Figure 4A-D**), moreover, the FP/FN predictions remain very limited (**Figure 4B-F**), again as in the case of internal human dataset, most FP/FN hits are ψ, but not Nm residues.

Overall, Nm sites in both datasets were predicted with a high precision albeit with somehow reduced sensitivity, which is also related to incomplete modification at some *S. cerevisiae* and *A. thaliana* rRNA sites.

More detailed analysis of performance and predictions obtained using other RF models (**Supp Figures S2-S6**) demonstrated that the results obtained with "Full model" are rather similar to "no_Surrounding_Base_info" model, except detection of almost universal modification 18S-m$^6{}_2$A,

which is present in highly conserved sequence context and thus efficiently detected by “Full model” in yeast rRNA. Other RT-arresting or RT-silent modifications were not detected, since no RiboMethSeq signal was generated for those residues. Overall performance of the selected RF model is much better than combination of scores’ thresholds (**Supp Figure S7**).

Performance of reduced RF model for Nm and Ψ prediction

Since the importance of different RF models’ features is not equal, to reduce computational efforts and complexity of the scripts, we decided to test the possibility to use the reduced RF prediction model including only 8 most important parameters, namely nature of nucleotide (base), and different RiboMethSeq scores (ratioL, ratioR, around L, aroundR, protection, ScoreMean and ScoreAngle, see **Table 1** and **Figure 5A**). Performance of this reduced model was evaluated as three others, using first internal human dataset and validation datasets from *S. cerevisiae* and *A. thaliana*. As shown in **Figure 5B**, the proportion of correctly predicted human rRNA modifications remained correct, even if overall prediction was less precise compared to “no_Surrounding_Base_info” model. However, the validation with *S. cerevisiae* and *A. thaliana* RiboMethSeq data pointed out a very poor performance of the reduced model for these different eukaryotic species. Only Um and Am residues were correctly predicted (**Figure 5C**) and even if FP identifications remained limited (**Figure 5D**), the number of FN predictions becomes very high. Precision of the reduced model was still acceptable, but the sensitivity was clearly degraded (**Supp Figure S3-S4**). This model has degraded performance and is not suitable for sensitive prediction of Nm and ψ residues.

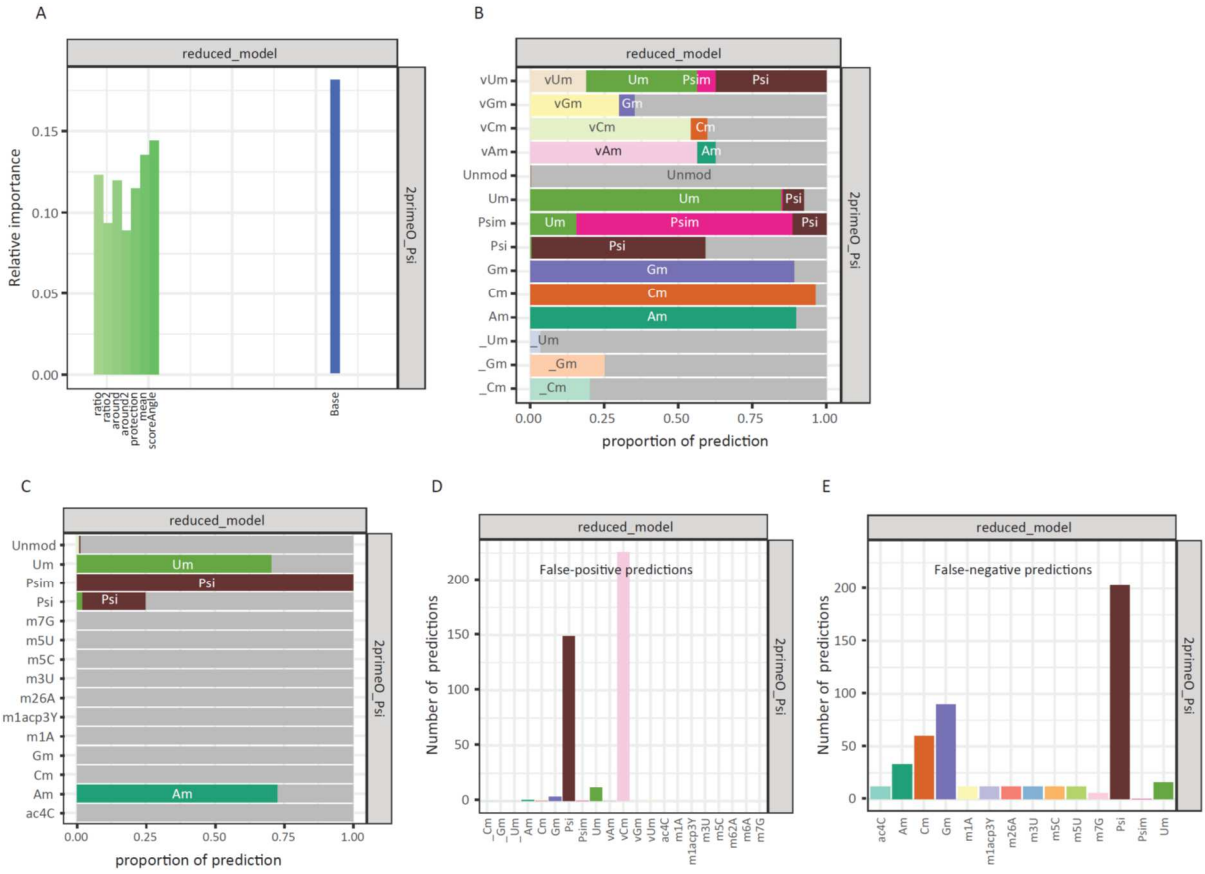

Figure 5 Performance of the reduced RF model for prediction of Nm and ψ residues. Relative importance of different features included in the reduced model (**A**). Prediction of human rRNA modifications using the internal dataset (**B**). Prediction of yeast S. cerevisiae rRNA modifications (**C**). Number and distribution of FP (**D**) and FN (**E**) hits. Notice that Cm and Gm residues are not at all predicted to be modified by the reduced RF model.

Discussion

The choice of the predictive model

In this study (overall design is outlined in **Figure 6**) we compared behavior and performance of four RF models ("Full model", "no_Surrounding_base_info", "no_Surrounding_info" and "Reduced model"), differing in the number and nature of parameters taken for the training and prediction. Performance of these models was evaluated for prediction of all possible modifications, for the target group (Nm and ψ residues), as well as for detection of RT-arresting/RT-silent RNA modifications. As anticipated, the "Full model" demonstrates a better performance for internal dataset of the same nature and for RNA modifications appearing in a highly conserved sequence. However, in case of external validation datasets (*S. cerevisiae* and *A. thaliana*), the best performance is obtained for "no_Surrounding_base_info" model, which does not use sequence consensus for prediction, but essentially based on the values of RiboMethSeq scores in the considered region. This model is well adapted for prediction of Nm and ψ residues, but, as anticipated, is not suitable for other RNA modifications, since the latter do not generate any distinct RiboMethSeq signal. Even if some non-Nm residues (e.g. $m^6{}_2A$) generate RT-stop and sudden drop in 5'-coverage, other approaches and additional features (like mutational RT signatures) have to be used in such cases of RT-arresting RNA modifications [45,46,48]. The attempt to reduce the number of RF model's parameters to the minimum ("Reduced model") clearly demonstrated that this approach does not provide a sufficient sensitivity in detection of true modified sites. This is likely related to the cumulative character of the prediction, every parameter brings incremental, but still important contribution into final decision, and thus reduction of the model leads to a substantial decrease of the performance.

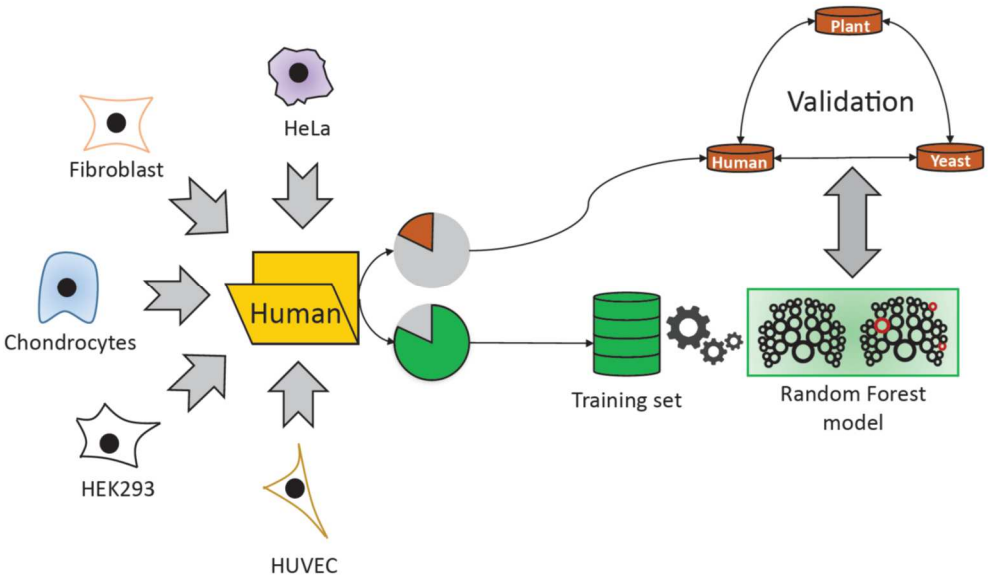

Figure 6. Outline of the RF model construction and validation. RiboMethSeq datasets from various human cell lines have been used as training and internal validation. To reduce sample selection biases, 80% of the data (356 970 of sites randomly selected from 62 samples) has been used as a training dataset, while the last 20% (89 243 sites) served as an internal testing dataset for initial assessment of models' performance. Further validation of the model and evaluation of its performance was done using yeast *S. cerevisiae* and plant *A. thaliana* RiboMethSeq datasets.

In order to reduce the number of FP and FN hits, the models used in this work have been trained on a validated collection of constitutively modified human 2'-O-methylated sites. We introduced the specific annotations for variably modified (vNm) or low modified (_Nm) positions. This annotation scheme showed to be efficient, since the models' precision is relatively high compared to preliminary models (not shown), where simple annotation even variable or low modified sites as Nm was used. However, RF is known to have multiple biases [49] on variable importance and/or classification and such biases can be anticipated to influence the results of sequencing detection methods. These biases may be related to variations between sequencing/analysis runs (for example, due to different behavior of the Reverse Transcriptase, used library preparation kit and parameters for reads alignment/trimming). However, these sources of variability were not explored in this study, since all samples were prepared using the same library preparation kit with the same RT enzyme.
In conclusion, considering degraded performance of other tested RF models, we recommend using "no_Surrounding_base_info" model for training and prediction of Nm and ψ residues.
In contrast to purely computational machine learning methods essentially based on primary sequencing data(like NmSEER V2.0 [50], NmRF [51], or DeepOMe [52]), the approach proposed in

this work is intended to be applied to deep analysis experimental RiboMethSeq datasets, and thus demonstrates superior performance in the absence of directly entered sequencing information.

Precision and sensitivity of RF model

Type I (FP hits) and type II (FN hits) errors are the main parameters assessed for predictive model performance. Although model optimization tends to minimize both error types, applications of such predictors are not really suitable for biological projects, since even relatively low False Discovery Rate (FDR) can lead to tremendous numbers of FP hits, since certain modifications are rare or even very rare in RNA. This is particularly true for mRNA analysis where overlap in modification mapping between different working groups varies from 30% to 60% only, even when the same detection method was used [18]. This is the reason to prioritize minimization of type I errors over type II in mapping of potential modified sites in unknown RNA.

Analysis of the essential performance parameters (mainly precision and sensitivity) demonstrates that all tested models show very low FP rate, albeit at the expense of some loss in sensitivity, and thus an increasing number of FN hits. Since the main challenge in the analysis of epitranscriptomics data, especially for large transcriptomics datasets of millions of nucleotides, is the highest precision/low FP rate (ratio of true positives to total predicted positives) some limited loss of sensitivity is quite acceptable. The observed precision close to 1 (0.99695282 for prediction of Nm and ψ) is largely sufficient for rRNA analysis (~10,000 nt) and similarly sized RNA species but is still inappropriate for very large and sparsely modified transcriptomics data, since the proportion of FP hits may be still very high. Combination of different approaches (e.g. RiboMethSeq and Nm-Seq at low [dNTP] [27,53]) may bring an acceptable solution to this problem.

Analysis of the observed False Positive (FP) and False Negative (FN) hits

Inspection of FP and FN hits obtained for the best-performing model “no_Surrounding_base_info”, clearly shows that most frequent FP hits are predictions of unmodified nucleotide as ψ, followed by some low frequency of FP Gm/Am/Um predictions. This is anticipated from the way of model’s construction and is an assumed attempt to make a distinction between Nm and potential ψ residue. In practice, the main goal of RiboMethSeq analysis is prediction of Nm candidates and if FP ψ predictions are sorted out for further validation by HydraPsiSeq [54,55], the total number of FP hits remains very low. Similarly, the most frequent FN hits also correspond to non-prediction of ψ, since many of those residues give only weak RiboMethSeq signal, undistinguishable from background. These minor limitations of the proposed RF predictive model must be considered in practical

applications. Alternative methods for ψ mapping, like HydraPsiSeq [54] or CMCT-based protocols [56,57] are better adapted to the precise mapping of those modified residues.

Applications to other RNA modifications or different mapping protocols

The approach described in this work may be further extended for analysis of other RNA modification generating any distinguishable signal in RiboMethSeq or other appropriate deep sequencing-based protocol for RNA analysis, like HydraPsiSeq [54], AlkAnilineSeq [58] or current BisulfiteSeq protocols [59]. With some minor adaptations, the RF model could be also applied to other sequencing detection methods such as MeRIP-seq, and to Nanopore sequencing data. Analysis of observed False Positive and False Negative hits for these two last methods would assess their detection robustness against different epitranscriptomics markers. The main difficulty remains the availability of appropriate and validated experimental datasets required for model's training and evaluation of performance.

Acknowledgements

This work was supported by ANR PRC grant "MethRibo" and Epi-ARN2018 and ViroMOD2020 FRCR grants from Grand Est Region, France to YM and Deutsche Forschungsgemeinschaft (DFG) grants to M.H. [HE3397/17-1, SPP1784, and TRR319 RMaP, TP C01].

References

[1] Y. Motorin, M. Helm, tRNA stabilization by modified nucleotides, Biochemistry. 49 (2010) 4934–4944. https://doi.org/10.1021/bi100408z.
[2] J. Liu, G. Jia, Methylation modifications in eukaryotic messenger RNA, J Genet Genomics. 41 (2014) 21–33. https://doi.org/10.1016/j.jgg.2013.10.002.
[3] K.D. Meyer, S.R. Jaffrey, The dynamic epitranscriptome: N6-methyladenosine and gene expression control, Nat. Rev. Mol. Cell Biol. 15 (2014) 313–326. https://doi.org/10.1038/nrm3785.
[4] B.S. Zhao, I.A. Roundtree, C. He, Post-transcriptional gene regulation by mRNA modifications, Nat. Rev. Mol. Cell Biol. 18 (2017) 31–42. https://doi.org/10.1038/nrm.2016.132.
[5] D.G. Dimitrova, L. Teysset, C. Carré, RNA 2'-O-Methylation (Nm) Modification in Human Diseases, Genes (Basel). 10 (2019). https://doi.org/10.3390/genes10020117.
[6] T. Sibbritt, H.R. Patel, T. Preiss, Mapping and significance of the mRNA methylome, Wiley Interdiscip Rev RNA. 4 (2013) 397–422. https://doi.org/10.1002/wrna.1166.
[7] M.A. O'Connell, N.M. Mannion, L.P. Keegan, The Epitranscriptome and Innate Immunity, PLoS Genet. 11 (2015) e1005687. https://doi.org/10.1371/journal.pgen.1005687.

[8] E.K. Borchardt, N.M. Martinez, W.V. Gilbert, Regulation and Function of RNA Pseudouridylation in Human Cells, Annu Rev Genet. 54 (2020) 309–336. https://doi.org/10.1146/annurev-genet-112618-043830.
[9] B. Linder, S.R. Jaffrey, Discovering and Mapping the Modified Nucleotides That Comprise the Epitranscriptome of mRNA, Cold Spring Harb Perspect Biol. 11 (2019). https://doi.org/10.1101/cshperspect.a032201.
[10] S. Nachtergaele, C. He, Chemical Modifications in the Life of an mRNA Transcript, Annu. Rev. Genet. 52 (2018) 349–372. https://doi.org/10.1146/annurev-genet-120417-031522.
[11] T.P. Hoernes, M.D. Erlacher, Translating the epitranscriptome, Wiley Interdiscip Rev RNA. 8 (2017). https://doi.org/10.1002/wrna.1375.
[12] T. Suzuki, H. Ueda, S. Okada, M. Sakurai, Transcriptome-wide identification of adenosine-to-inosine editing using the ICE-seq method, Nat Protoc. 10 (2015) 715–732. https://doi.org/10.1038/nprot.2015.037.
[13] A.M. Kietrys, W.A. Velema, E.T. Kool, Fingerprints of Modified RNA Bases from Deep Sequencing Profiles, Journal of the American Chemical Society. 139 (2017) 17074–17081. https://doi.org/10.1021/jacs.7b07914.
[14] V. Marchand, F. Pichot, K. Thüring, L. Ayadi, I. Freund, A. Dalpke, M. Helm, Y. Motorin, Next-Generation Sequencing-Based RiboMethSeq Protocol for Analysis of tRNA 2'-O-Methylation, Biomolecules. 7 (2017). https://doi.org/10.3390/biom7010013.
[15] Y. Motorin, M. Helm, Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies, Genes (Basel). 10 (2019). https://doi.org/10.3390/genes10010035.
[16] Y. Motorin, V. Marchand, Analysis of RNA Modifications by Second- and Third-Generation Deep Sequencing: 2020 Update, Genes (Basel). 12 (2021) 278. https://doi.org/10.3390/genes12020278.
[17] A.V. Grozhik, S.R. Jaffrey, Distinguishing RNA modifications from noise in epitranscriptome maps, Nat. Chem. Biol. 14 (2018) 215–225. https://doi.org/10.1038/nchembio.2546.
[18] A.B.R. McIntyre, N.S. Gokhale, L. Cerchietti, S.R. Jaffrey, S.M. Horner, C.E. Mason, Limits in the detection of m6A changes using MeRIP/m6A-seq, Scientific Reports. 10 (2020) 6590. https://doi.org/10.1038/s41598-020-63355-3.
[19] A.V. Grozhik, S.R. Jaffrey, Epitranscriptomics: Shrinking maps of RNA modifications, Nature. 551 (2017) 174–176. https://doi.org/10.1038/nature24156.
[20] D. Wiener, S. Schwartz, The epitranscriptome beyond m6A, Nat Rev Genet. (2020). https://doi.org/10.1038/s41576-020-00295-8.
[21] A. Sas-Chen, S. Schwartz, Misincorporation signatures for detecting modifications in mRNA: Not as simple as it sounds, Methods (San Diego, Calif.). 156 (2019) 53–59. https://doi.org/10.1016/j.ymeth.2018.10.011.
[22] V. Marchand, F. Blanloeil-Oillo, M. Helm, Y. Motorin, Illumina-based RiboMethSeq approach for mapping of 2'-O-Me residues in RNA, Nucleic Acids Res. 44 (2016) e135. https://doi.org/10.1093/nar/gkw547.
[23] J. Erales, V. Marchand, B. Panthu, S. Gillot, S. Belin, S.E. Ghayad, M. Garcia, F. Laforêts, V. Marcel, A. Baudin-Baillieu, P. Bertin, Y. Couté, A. Adrait, M. Meyer, G. Therizols, M. Yusupov, O. Namy, T. Ohlmann, Y. Motorin, F. Catez, J.-J. Diaz, Evidence for rRNA 2'-O-methylation plasticity: Control of intrinsic translational capabilities of human ribosomes, Proceedings of the National Academy of Sciences of the United States of America. 114 (2017) 12934–12939. https://doi.org/10.1073/pnas.1707674114.
[24] M. Ringeard, V. Marchand, E. Decroly, Y. Motorin, Y. Bennasser, FTSJ3 is an RNA 2'-O-methyltransferase recruited by HIV to avoid innate immune sensing, Nature. 565 (2019) 500–504. https://doi.org/10.1038/s41586-018-0841-4.
[25] M.T. Angelova, D.G. Dimitrova, B. Da Silva, V. Marchand, C. Jacquier, C. Achour, M. Brazane, C. Goyenvalle, V. Bourguignon-Igel, S. Shehzada, S. Khouider, T. Lence, V. Guerineau, J.-Y.

Roignant, C. Antoniewski, L. Teysset, D. Bregeon, Y. Motorin, M.R. Schaefer, C. Carré, tRNA 2'-O-methylation by a duo of TRM7/FTSJ1 proteins modulates small RNA silencing in Drosophila, Nucleic Acids Res. 48 (2020) 2050–2072. https://doi.org/10.1093/nar/gkaa002.
[26] R. Gumienny, D.J. Jedlinski, A. Schmidt, F. Gypas, G. Martin, A. Vina-Vilaseca, M. Zavolan, High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq, Nucleic Acids Res. 45 (2017) 2341–2353. https://doi.org/10.1093/nar/gkw1321.
[27] J. Azevedo-Favory, C. Gaspin, L. Ayadi, C. Montacié, V. Marchand, E. Jobet, M. Rompais, C. Carapito, Y. Motorin, J. Sáez-Vásquez, Mapping rRNA 2'-O-methylations and identification of C/D snoRNAs in Arabidopsis thaliana plants, RNA Biol. 18 (2021) 1760–1777. https://doi.org/10.1080/15476286.2020.1869892.
[28] J. Delhermite, L. Tafforeau, S. Sharma, V. Marchand, L. Wacheul, R. Lattuca, S. Desiderio, Y. Motorin, E. Bellefroid, D.L.J. Lafontaine, Systematic mapping of rRNA 2'-O methylation during frog development and involvement of the methyltransferase Fibrillarin in eye and craniofacial development in Xenopus laevis, PLoS Genet. 18 (2022) e1010012. https://doi.org/10.1371/journal.pgen.1010012.
[29] F. Pichot, V. Marchand, L. Ayadi, V. Bourguignon-Igel, M. Helm, Y. Motorin, Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2'-O-Methylations in RNA by RiboMethSeq, Front Genet. 11 (2020) 38. https://doi.org/10.3389/fgene.2020.00038.
[30] Y. Motorin, V. Marchand, Detection and Analysis of RNA Ribose 2'-O-Methylations: Challenges and Solutions, Genes (Basel). 9 (2018). https://doi.org/10.3390/genes9120642.
[31] V. Marchand, L. Ayadi, A. El Hajj, F. Blanloeil-Oillo, M. Helm, Y. Motorin, High-Throughput Mapping of 2'-O-Me Residues in RNA Using Next-Generation Sequencing (Illumina RiboMethSeq Protocol), Methods Mol. Biol. 1562 (2017) 171–187. https://doi.org/10.1007/978-1-4939-6807-7_12.
[32] U. Birkedal, M. Christensen-Dalsgaard, N. Krogh, R. Sabarinathan, J. Gorodkin, H. Nielsen, Profiling of ribose methylations in RNA by high-throughput sequencing, Angew. Chem. Int. Ed. Engl. 54 (2015) 451–455. https://doi.org/10.1002/anie.201408362.
[33] N. Krogh, U. Birkedal, H. Nielsen, RiboMeth-seq: Profiling of 2'-O-Me in RNA, Methods Mol. Biol. 1562 (2017) 189–209. https://doi.org/10.1007/978-1-4939-6807-7_13.
[34] Y. Motorin, M. Quinternet, W. Rhalloussi, V. Marchand, Constitutive and variable 2'-O-methylation (Nm) in human ribosomal RNA, RNA Biol. 18 (2021) 88–97. https://doi.org/10.1080/15476286.2021.1974750.
[35] M. Nadif, F. Role, Unsupervised and self-supervised deep learning approaches for biomedical text mining, Brief Bioinform. 22 (2021) 1592–1603. https://doi.org/10.1093/bib/bbab016.
[36] P.S. Reel, S. Reel, E. Pearson, E. Trucco, E. Jefferson, Using machine learning approaches for multi-omics data analysis: A review, Biotechnol Adv. 49 (2021) 107739. https://doi.org/10.1016/j.biotechadv.2021.107739.
[37] B. Schmidt, A. Hildebrandt, Deep learning in next-generation sequencing, Drug Discov Today. 26 (2021) 173–180. https://doi.org/10.1016/j.drudis.2020.10.002.
[38] Y. Zhang, M. Hamada, DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning, BMC Bioinformatics. 19 (2018) 524. https://doi.org/10.1186/s12859-018-2516-4.
[39] K.-T. Lin, A.R. Krainer, PSI-Sigma: a comprehensive splicing-detection method for short-read and long-read RNA-seq analysis, Bioinformatics. 35 (2019) 5048–5054. https://doi.org/10.1093/bioinformatics/btz438.
[40] O. Begik, M.C. Lucas, L.P. Pryszcz, J.M. Ramirez, R. Medina, I. Milenkovic, S. Cruciani, H. Liu, H.G.S. Vieira, A. Sas-Chen, J.S. Mattick, S. Schwartz, E.M. Novoa, Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing, Nat Biotechnol. 39 (2021) 1278–1291. https://doi.org/10.1038/s41587-021-00915-6.

[41] D.A. Lorenz, S. Sathe, J.M. Einstein, G.W. Yeo, Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution, RNA (New York, N.Y.). 26 (2020) 19–28. https://doi.org/10.1261/rna.072785.119.
[42] J. Silvestre-Ryan, I. Holmes, Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing, Genome Biol. 22 (2021) 38. https://doi.org/10.1186/s13059-020-02255-1.
[43] R.R. Wick, L.M. Judd, K.E. Holt, Performance of neural network basecalling tools for Oxford Nanopore sequencing, Genome Biol. 20 (2019) 129. https://doi.org/10.1186/s13059-019-1727-y.
[44] H. Qin, L. Ou, J. Gao, L. Chen, J.-W. Wang, P. Hao, X. Li, DENA: training an authentic neural network model using Nanopore sequencing data of Arabidopsis transcripts for detection and quantification of N6-methyladenosine on RNA, Genome Biol. 23 (2022) 25. https://doi.org/10.1186/s13059-021-02598-3.
[45] R. Hauenschild, L. Tserovski, K. Schmid, K. Thüring, M.-L. Winz, S. Sharma, K.-D. Entian, L. Wacheul, D.L.J. Lafontaine, J. Anderson, J. Alfonzo, A. Hildebrandt, A. Jäschke, Y. Motorin, M. Helm, The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent, Nucleic Acids Res. 43 (2015) 9950–9964. https://doi.org/10.1093/nar/gkv895.
[46] S. Werner, L. Schmidt, V. Marchand, T. Kemmer, C. Falschlunger, M.V. Sednev, G. Bec, E. Ennifar, C. Höbartner, R. Micura, Y. Motorin, A. Hildebrandt, M. Helm, Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes, Nucleic Acids Res. 48 (2020) 3734–3746. https://doi.org/10.1093/nar/gkaa113.
[47] L. Ayadi, Y. Motorin, V. Marchand, Quantification of 2'-O-Me residues in RNA using next-generation sequencing (Illumina RiboMethSeq protocol), Methods Mol. Biol. 1649 (2018) 29–48. https://doi.org/10.1007/978-1-4939-7213-5_2.
[48] L. Tserovski, V. Marchand, R. Hauenschild, F. Blanloeil-Oillo, M. Helm, Y. Motorin, High-throughput sequencing for 1-methyladenosine (m(1)A) mapping in RNA, Methods. 107 (2016) 110–121. https://doi.org/10.1016/j.ymeth.2016.02.012.
[49] C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinformatics. 8 (2007) 25. https://doi.org/10.1186/1471-2105-8-25.
[50] Y. Zhou, Q. Cui, Y. Zhou, NmSEER V2.0: a prediction tool for 2'-O-methylation sites based on random forest and multi-encoding combination, BMC Bioinformatics. 20 (2019) 690. https://doi.org/10.1186/s12859-019-3265-8.
[51] C. Ao, Q. Zou, L. Yu, NmRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences, Brief Bioinform. 23 (2022) bbab480. https://doi.org/10.1093/bib/bbab480.
[52] H. Li, L. Chen, Z. Huang, X. Luo, H. Li, J. Ren, Y. Xie, DeepOMe: A Web Server for the Prediction of 2'-O-Me Sites Based on the Hybrid CNN and BLSTM Architecture, Front Cell Dev Biol. 9 (2021) 686894. https://doi.org/10.3389/fcell.2021.686894.
[53] D. Incarnato, F. Anselmi, E. Morandi, F. Neri, M. Maldotti, S. Rapelli, C. Parlato, G. Basile, S. Oliviero, High-throughput single-base resolution mapping of RNA 2′-O-methylated residues, Nucleic Acids Res. 45 (2017) 1433–1441. https://doi.org/10.1093/nar/gkw810.
[54] V. Marchand, F. Pichot, P. Neybecker, L. Ayadi, V. Bourguignon-Igel, L. Wacheul, D.L.J. Lafontaine, A. Pinzano, M. Helm, Y. Motorin, HydraPsiSeq: a method for systematic and quantitative mapping of pseudouridines in RNA, Nucleic Acids Research. 48 (2020) e110. https://doi.org/10.1093/nar/gkaa769.
[55] V. Marchand, V. Bourguignon-Igel, M. Helm, Y. Motorin, Analysis of pseudouridines and other RNA modifications using HydraPsiSeq protocol, Methods. (2021) S1046-2023(21)00206–1. https://doi.org/10.1016/j.ymeth.2021.08.008.

[56] S. Schwartz, D.A. Bernstein, M.R. Mumbach, M. Jovanovic, R.H. Herbst, B.X. León-Ricardo, J.M. Engreitz, M. Guttman, R. Satija, E.S. Lander, G. Fink, A. Regev, Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA, Cell. 159 (2014) 148–162. https://doi.org/10.1016/j.cell.2014.08.028.
[57] T.M. Carlile, M.F. Rojas-Duran, B. Zinshteyn, H. Shin, K.M. Bartoli, W.V. Gilbert, Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells, Nature. 515 (2014) 143–146. https://doi.org/10.1038/nature13802.
[58] V. Marchand, L. Ayadi, F.G.M. Ernst, J. Hertler, V. Bourguignon-Igel, A. Galvanin, A. Kotter, M. Helm, D.L.J. Lafontaine, Y. Motorin, AlkAniline-Seq: Profiling of m7 G and m3 C RNA Modifications at Single Nucleotide Resolution, Angew. Chem. Int. Ed. Engl. 57 (2018) 16785–16790. https://doi.org/10.1002/anie.201810946.
[59] V. Khoddami, A. Yerra, T.L. Mosbruger, A.M. Fleming, C.J. Burrows, B.R. Cairns, Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution, Proc. Natl. Acad. Sci. U.S.A. 116 (2019) 6784–6789. https://doi.org/10.1073/pnas.1817334116.

Figure legends

Figure 1 Scatter plot for scores mean and A2 for randomly selected RiboMethSeq sample of human 28S rRNA. Unmodified residues (A, C, G and U) are shown as dots with pastel color, known modified sites are in solid color. Commonly used combination of thresholds for score A2 (> 0.5) and score mean (> 0.92) is shown as dashed red lines. Positions holding exceeding values for both scores are considered to be 2'-O-methylated candidates.

Figure 2 Typical RiboMethSeq profiles for human 5.8S rRNA. Top panel - 5'-end count (blue), middle panel - 3'-end count (red) and merged profile (bottom, green). Positions of two known Nm residues (Um14 and Gm75) in 5.8S rRNA are shown. Numbering is in 0->n-1 format (from *.bed format), so Um14 and Gm75 signals appear at the nucleotide positions N and not at N+1.

Figure 3 Importance of RF model's parameters and performance of "no_Surrounding_Base_info" model in detection of Nm and pseudouridine (Psi/ψ) residues (other models and their performance are illustrated in **Supp Figures S2-S6**). **Panel A** – Relative importance for each feature has been calculated by normalizing Gini importance (or mean decrease impurity) across all features. **Panel B** shows the influence of the RF model on the precision (left) and sensitivity (right) of Nm and ψ detection. Identity of the models is shown on the bottom (Full model, no surrounding base info, no surrounding info). **Panel C** shows the results of predictions for confirmed and invariable Nm residues (Am, Cm, Gm, Um), variable positions Nm (vAm, vCm, vGm, vUm) and some low modified positions (_Um, _Gm, _Cm). No low modified _Am residues were present in the training dataset. Unmod stands for unmodified nucleotide in the sequence. **Panels D** and **E** show distribution of False-positive (FP) and False-negative (FN) hits for predictions using internal training dataset (20% of human RiboMethSeq datasets kept aside for validation, see Materials&Methods).

Figure 4 Validation of RF models using RiboMethSeq datasets for *S. cerevisiae* and *A. thaliana* rRNAs. Only Nm and Ψ residues were reliably mapped in *A. thaliana* rRNAs and this limited modification profile was used as a reference. Results of prediction for Nm, ψ and other RNA modifications are shown for "no_Surrounding_Base_info" model applied for *S. cerevisiae* (**A**) and *A. thaliana* (**D**) RiboMethSeq datasets. Identity of prediction is shown by color. **Panels B** and **F** show FP hits for *S. cerevisiae* (**B**) and *A. thaliana* (**E**) datasets, while **panels C** and **F** list FN predictions for both.

Figure 5 Performance of the reduced RF model for prediction of Nm and ψ residues. Relative importance of different features included in the reduced model (**A**). Prediction of human rRNA modifications using the internal dataset (**B**). Prediction of yeast S. cerevisiae rRNA

modifications (**C**). Number and distribution of FP (**D**) and FN (**E**) hits. Notice that Cm and Gm residues are not at all predicted to be modified by the reduced RF model.

Figure 6. Outline of the RF model construction and validation. RiboMethSeq datasets from various human cell lines have been used as training and internal validation. To reduce sample selection biases, 80% of the data (356 970 of sites randomly selected from 62 samples) has been used as a training dataset, while the last 20% (89 243 sites) served as an internal testing dataset for initial assessment of models' performance. Further validation of the model and evaluation of its performance was done using yeast *S. cerevisiae* and plant *A. thaliana* RiboMethSeq datasets.