communications biology

ARTICLE

https://doi.org/10.1038/s42003-022-03821-y OPEN

Predicting genes associated with RNA methylation pathways using machine learning

Georgia Tsagkogeorga[1,2✉], Helena Santos-Rosa[3], Andrej Alendar[3], Dan Leggate[1], Oliver Rausch[1], Tony Kouzarides[2,3], Hendrik Weisser[1,5✉] & Namshik Han[2,4,5✉]

RNA methylation plays an important role in functional regulation of RNAs, and has thus attracted an increasing interest in biology and drug discovery. Here, we collected and collated transcriptomic, proteomic, structural and physical interaction data from the Harmonizome database, and applied supervised machine learning to predict novel genes associated with RNA methylation pathways in human. We selected five types of classifiers, which we trained and evaluated using cross-validation on multiple training sets. The best models reached 88% accuracy based on cross-validation, and an average 91% accuracy on the test set. Using protein-protein interaction data, we propose six molecular sub-networks linking model predictions to previously known RNA methylation genes, with roles in mRNA methylation, tRNA processing, rRNA processing, but also protein and chromatin modifications. Our study exemplifies how access to large omics datasets joined by machine learning methods can be used to predict gene function.

[1] STORM Therapeutics Ltd, Babraham Research Campus, Cambridge, UK. [2] Milner Therapeutics Institute, University of Cambridge, Puddicombe Way, Cambridge, UK. [3] The Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge, UK. [4] Cambridge Centre for AI in Medicine, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. [5] These authors contributed equally: Hendrik Weisser and Namshik Han. ✉email: georgia.tsagkogeorga@stormtherapeutics.com; hendrik.weisser@stormtherapeutics.com; n.han@milner.cam.ac.uk

RNA modifications have been known since the 1960s, when the sequencing of the first transfer RNA (tRNA) from yeast revealed 10 chemically modified ribonucleosides, including pseudouridine (Ψ)[1]. Since then, the number of identified modifications has grown to over 150, found on both coding and non-coding RNAs across all three kingdoms of life[2]. Technological advances in the field have established that RNA modifications are widespread, reversible and dynamically regulated[1]. Methylation is the most abundant type, with methyl-groups decorating multiple RNA species, such as messenger RNA (mRNA), ribosomal RNA (rRNA) and tRNA, at different nucleosides and positions. So far, N6-methyladenosine (m^6A) is the most studied modification, commonly detected in mRNA, rRNA, long intergenic non-coding RNA (lincRNA), primary microRNA (pri-miRNA), and small nuclear RNAs (snRNA). Other methyl-marks include 5-methylcytosine (m^5C), N1-methyladenosine (m^1A), 7-methylguanosine (m^7G), 2'-O-dimethyladenosine (m^6Am) and 5-hydroxymethylcytosine (hm^5C)[3–5].

Deposition of methyl-marks on RNA is catalysed by writer enzymes, known as RNA methyltransferases. To date, there are 57 RNA methyltransferases identified in the human genome. Of these, five methylate mRNAs, six small RNAs, 14 rRNAs, and 22 tRNAs, whereas 12 remain with unknown substrates[6]. Most enzymes use S-adenosyl-methionine (SAM) as a methyl donor to the RNA substrate, while many also recruit accessory proteins, which are often essential for substrate binding, localisation, and stability. The most well-studied examples of RNA methylation writers are by far the complex METTL3-METTL14 complex responsible for the deposition of m^6A, followed by a NOL1/NOP2/Sun (NSUN) domain-containing family of tRNA-modifying enzymes depositing m^5C on tRNAs[7].

Dynamic regulation of RNAs via chemical modifications has recently attracted a rising interest in RNA modifying enzymes as new potential therapeutic targets[8]. This is because multiple lines of evidence suggest that RNA methylation plays a far more important role in cell functioning than previously thought. In line with this, several studies have shown that RNA methylation is a key modulator of transcript stability, gene expression, splicing and translation efficiency[9–11]. Furthermore, a growing body of data has demonstrated that changes in RNA methylation processes can be linked to a range of cancers, neurological disorders and various other diseases[12]. Surprisingly, despite this critical role in cellular homoeostasis and disease, RNA methylation pathways in general remain understudied[7]. Our current understanding of RNA modifications is also highly fragmentary, with an estimated 20% or more of RNA modifying enzymes still remaining unknown or unidentified[13].

Conventional approaches for studying novel gene functions include a range of labour-intensive wet-lab techniques, including mutagenesis, gene disruption or gene depletion (knocking-down/-out) for characterising gene-specific phenotypic effects, and chromatography and mass spectrometry for identifying molecular interactions. However, over the last two decades, access to large-scale omics data has enabled the use of "dry" computational methods for understanding biological functions. A wide array of bioinformatic tools have been developed under the umbrella of functional genomics, ranging from methods used to identify homologous genes with similar functionalities across species to genome-wide screens for specific sequence motifs and functional domains. Today, machine learning techniques are emerging as a powerful approach to harness the increasing wealth of large-scale biological data, allowing the discovery of hidden patterns and more reliable statistical predictions[14].

Here, we aimed to better understand the molecular pathways involved in RNA methylation in human using machine learning. To this end, we used publicly available human transcriptomic, proteomic, structural and protein–protein interaction data[15] and built a large machine learning dataset for supervised binary classification. We trained and evaluated five ensembles of predictive models: Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB) models. We employed the best models to predict genes functionally associated with RNA methylation pathways in the human genome.

Results and discussion

Data engineering and feature selection. Mining functional annotation databases in conjunction with extensive literature searches allowed us to identify 92 proteins involved in RNA methylation (Supplementary Data 1). These were either methyl-writers (known RNA methyltransferases[6] and their partner proteins in protein complexes), or enzymes previously annotated as putative RNA methyltransferases (see Methods). Genes encoding for these proteins constituted our positive class (Class 1) in machine learning analyses. To frame our predictive modelling as a binary classification problem, we assembled multiple stratified training and test datasets by randomly sampling a number of genes equal to our positive set from the remaining genome, ensuring that all genes of our initial dataset were sampled exactly once (Fig. 1). Our rationale was that this would allow machine learning models to be trained and tested across a diverse range of other gene functions, instead of just choosing one function for the negative set. In addition, this approach alleviates any putative bias that may arise from sampling a single negative set of genes from the human genome.

We initially pooled 50,176 features collected from publicly available and previously curated transcriptomic, proteomic, functional annotation, structural and physical interaction datasets (Supplementary Data 2). To identify features that were informative for classification and thereby useful for predicting genes associated with RNA methylation, we performed feature selection prior to model training, followed by feature ranking after training and cross-validation. To reduce the feature-to-sample ratio, first we eliminated features with excessive missing data in the training dataset. Second, we removed features with low variance, which resulted in a drastic dimensionality reduction to 1,505 features for the final dataset. Selected features used for classification were drawn from BioGPS[16] (35), Gene Ontology[17] (GO: 59), GTEx[18] (1,114), Human Protein Atlas[19] (HPA: 107), InterPro (1), Pathway Commons (PathCommons: 150) and TISSUES[20] (40) datasets.

During model training and cross-validation, we computed feature importance by using the GB importance measure as averaged across all training sets. The 50 most informative features and their relative importance in classification are shown in Supplementary Fig. 1. The features with the highest importance for the full feature set were mainly GO terms, such as GO:0032259, GO:0016740, GO:0003723, GO:0008168 and GO:0016070, all corresponding to methylation, transferase/methyltransferase activity and RNA metabolic processes. Equally, the InterPro domain IPR029063, which represents the S-adenosyl-L-methionine-dependent methyltransferase superfamily was ranked among the top 50 most informative features (Supplementary Fig. 1a). Although anticipated, the fact that the classifiers seemed to rely on RNA and methylation-related annotation features provides support that the models learn to classify genes with a strong link to RNA methylation processes.

Although GO annotations are informative, they may equally bias gene prediction towards pre-existing functional annotations. We assembled thus a second feature set of reduced dimensionality, by excluding GO and InterPro data types. When classifiers were

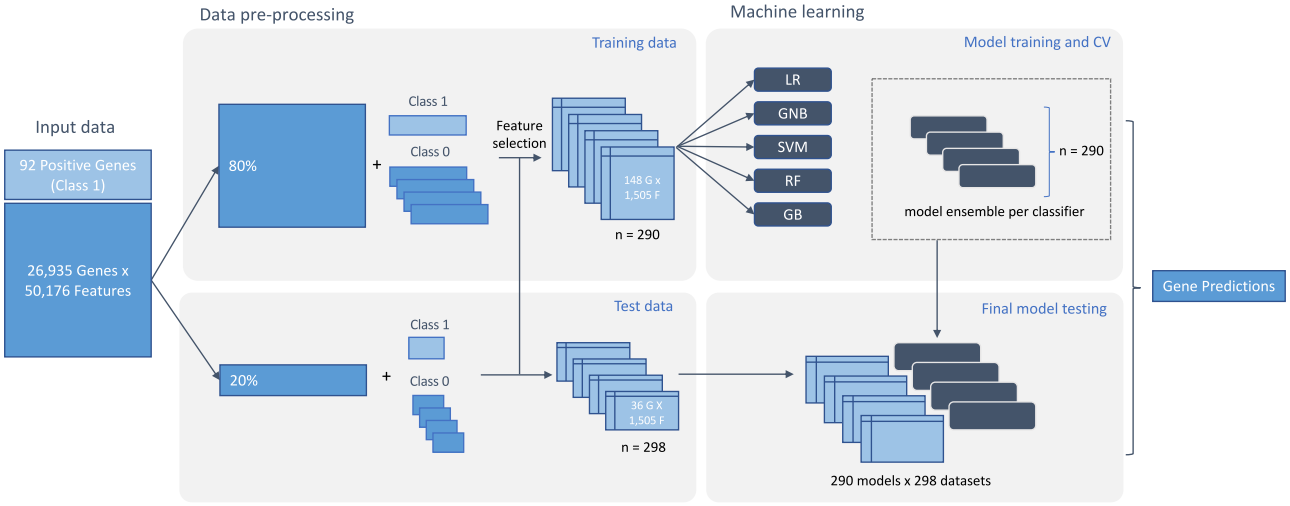

Fig. 1 Schematic representation of the analysis workflow. Previously known RNA methylation genes were used as positive samples (Class 1) and split into two sets comprising 80% of the data for training and 20% kept unseen for model testing. An analogous 80/20 split was performed for the remaining genes of the human genome, which were further divided into sets of equal size to the positive samples and used as negative samples (Class 0) to generate stratified sets for feature selection, training and testing. Following feature pre-filtering, five types of machine learning models for binary classification—Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB)—were trained on each of the training sets resulting in a classifier ensemble. Each model from the classifier ensemble was evaluated on each of the test datasets and overall performance was calculated by averaging results of all models across test sets. The best-performing ensemble was used to make predictions for the whole genome.

trained on this reduced feature set, the most informative types of features were mainly GTEx expression profiles (Supplementary Fig. 1b). The GTEx project aims to provide a comprehensive public resource of tissue-specific gene expression and regulation, so far including samples from 54 non-diseased tissues across nearly 1000 individuals[18]. Tissue sample expression data as integrated into Harmonizome and thus sampled here, consist of one-hot-encoded sets of genes with high or low expression in each tissue sample relative to other tissue samples from the GTEx tissue expression profiles dataset.

A possible interpretation of the high ranking of such GTEx expression profile features is that under specific biological conditions, i.e., in certain tissues, RNA methylation genes tend to be collectively down- or up-regulated as compared to other processes. Alternatively, a high ranking of GTEx features may be due to the high proportion of GTEx features in the feature set and noise originating from the high dimensionality of the training dataset with respect to the feature-to-sample ratio. To investigate this further, we calculated the relative frequency of GTEx features in the top hundred most informative features across models from all training sets (Supplementary Data 3). Notably, certain samples taken from the areas of blood, heart, pancreas, and brain were retrieved as informative by more than a hundred models.

Model performance. We selected five machine learning classifiers (LR, GNB, SVM, RF and GB) and trained each on training sets from the full and the reduced feature set, creating an ensemble of models per classifier and feature set. Overall, all five model ensembles showed very similar performance based on cross-validation (Table 1). Among classifiers trained using the full feature set, GB and RF models showed the highest average accuracy at 0.875 and 0.870, respectively, as well as a similarly high average precision of 0.895 and 0.870, respectively. The GB ensemble followed by that the RF models also yielded the highest AUROC score, with an average AUC estimated at 0.938 and 0.937, respectively.

The performance of the five classifiers for the reduced feature set without GO/InterPro annotations was diminished compared to the full dataset (Table 1). The model ensembles of SVM and RF outperformed the remaining three ensembles across almost all metrics. SVM models performed the best on the reduced feature set based on cross-validation, with an average prediction accuracy of 0.812, precision of 0.822 and AUROC of 0.864.

Model predictions. To evaluate results from different models and feature sets, we first compared the distribution of probability scores for all human genes, as predicted by models trained on the full feature set (Fig. 2a) and models derived from the reduced feature set (Fig. 2b). The prediction landscape appeared very similar across all five types of model ensembles, as exemplified by the extensive overlap of their respective distributions. Most genes were highly skewed towards an average probability score of zero, in line with the hypothesis that the majority of human genes are not expected to be directly involved in RNA methylation pathways. Note, however, that GNB models showed an aberrant prediction profile, with a notably higher peak near probability one, compared to that of all other classifiers.

Second, we assessed the degree of overlap among genes predicted with high confidence to be involved in RNA methylation by different machine learning models. Here, we defined as high confidence all genes in the top 1% of the probability distribution for Class 1. For most classifiers this fraction encompassed ~270 genes, except from GNB models that ambiguously assigned the same high probability score to a large number of genes, hence a larger set was considered (> 1,750 genes). When comparing top predictions from different models, their relative concordance was high for both feature sets (Supplementary Fig. 2). We identified over 280 genes that were commonly predicted to be involved in RNA methylation pathways, by at least three out of five types of model ensembles at the selected confidence cut-off (Supplementary Fig. 2a, b).

Finally, to get a high-level understanding of the predictions from different models, we performed exploratory GO enrichment analyses using the same high confidence genes as above. The top 10 enriched terms for each machine learning model are compared in Fig. 3. All model ensembles, independently of the dataset they

Table 1 Model performance based on 10-fold cross-validation.

	Accuracy	Precision	Recall	F1	AUC
Full feature set					
GB	0.875 ± 0.025	0.895 ± 0.033	0.865 ± 0.031	0.872 ± 0.025	0.938 ± 0.015
GNB	0.851 ± 0.025	0.821 ± 0.032	0.924 ± 0.021	0.863 ± 0.021	0.862 ± 0.023
LR	0.859 ± 0.021	0.870 ± 0.025	0.859 ± 0.023	0.857 ± 0.021	0.921 ± 0.015
RF	0.870 ± 0.021	0.870 ± 0.026	0.886 ± 0.032	0.871 ± 0.022	0.937 ± 0.014
SVM	0.856 ± 0.022	0.876 ± 0.028	0.845 ± 0.027	0.852 ± 0.023	0.921 ± 0.017
Reduced feature set					
GB	0.799 ± 0.029	0.800 ± 0.035	0.819 ± 0.032	0.801 ± 0.029	0.860 ± 0.031
GNB	0.781 ± 0.022	0.765 ± 0.028	0.840 ± 0.043	0.792 ± 0.024	0.800 ± 0.021
LR	0.795 ± 0.030	0.797 ± 0.035	0.814 ± 0.030	0.797 ± 0.029	0.857 ± 0.032
RF	0.805 ± 0.024	0.802 ± 0.033	0.833 ± 0.023	0.809 ± 0.022	0.867 ± 0.025
SVM	0.812 ± 0.027	0.822 ± 0.036	0.816 ± 0.032	0.811 ± 0.027	0.864 ± 0.026

LR Logistic Regression, *GNB* Gaussian Naïve Bayes, *SVM* Support Vector Machine, *RF* Random Forest, *GB* Gradient Boosting.

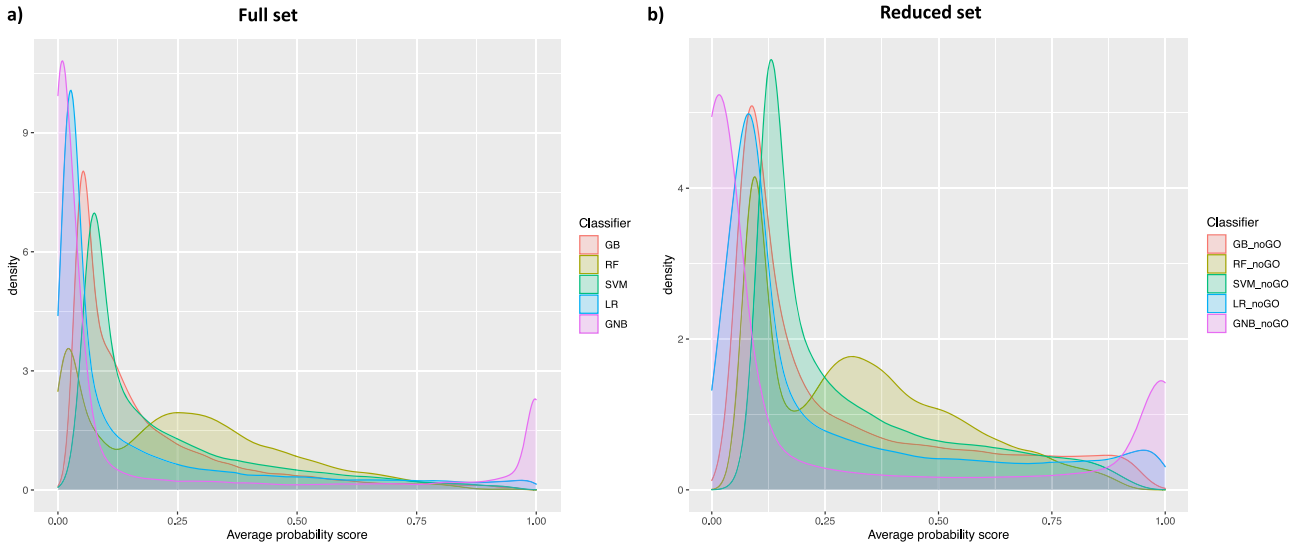

Fig. 2 Model predictions across all human genes. Probability score distributions for Class 1 as obtained in predictive modelling based on the (**a**) full and (**b**) reduced feature sets.

derived from, yielded predictions enriched in GO terms associated with RNA-binding and RNA catalytic activities. Note that top enrichment results for models trained on the full feature set included additionally terms associated with chromatin and protein methylation processes (Fig. 3). Of models trained on the reduced feature set, only GNB and SVM showed an enrichment in protein methylation. This may indicate a modelling artefact, i.e., predictions erroneously assigned to Class 1, that could be caused by the hierarchical nature of GO terms (e.g., "methylation" being the parent term of both "RNA methylation" and "protein methylation" processes). Yet, all models with the exception of GB predicted genes that were previously associated with a catalytic activity on DNA. Therefore, an alternative interpretation is that our models capture a putative functional link between modification pathways operating on different substrates. Overall, the functional annotation analyses provided a good qualitative control for model performance. The rationale here is that although we did not recover enrichment in the biological term "RNA methylation" per se (given that the models predict "novel" genes), features closely associated with the term should figure among the top GO results. A breakdown of the GO enrichment results by feature set is provided in Supplementary Fig. 3.

In silico validation. Of all classifiers, GB models that were trained on the full feature set showed the best performance based on cross-validation, and have thus been selected to apply on previously unseen test data. Model performance metrics for the stratified test datasets were calculated by averaging the values obtained for each model in the ensemble. The average test set accuracy for the GB ensemble was 0.905, precision 0.897, recall 0.923, and AUCROC 0.973.

It is known, however, that a high feature-to-sample ratio may lead to overfitting and overestimation of model performance. As our machine learning modelling was based on a small number of positive genes (Class 1), we were particularly interested in obtaining an estimate of the false positives in our predictions. To this goal, we pooled together the entire hold-out data (18 positive and 5,368 negative examples) and averaged the predicted probability score of each gene as estimated by each model in the GB ensemble. This allowed us to estimate the false positives, by counting the number of negative genes (Class 0) that were wrongfully predicted by our models as positive. Of the 5368 negative genes, 425 were erroneously classified, resulting to an estimated false positive rate of 0.079, defined as the number of False Positives (FP) divided by the sum of False Positives (FP) and True Negatives (TN).

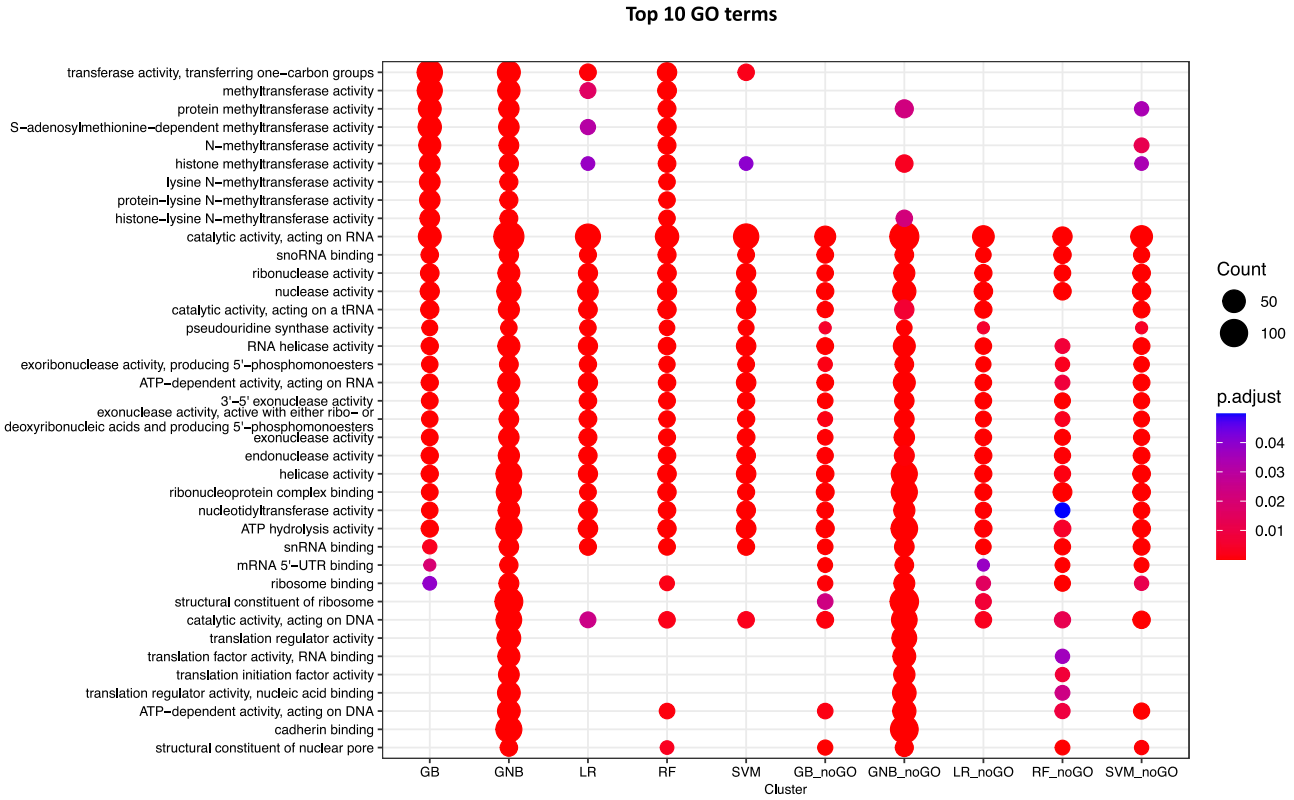

Fig. 3 Functional enrichment analyses of high-confidence predictions. Comparing GO enrichment results of genes in the top 1% of the probability distribution for Class 1 for each model ensemble. The top enriched terms include functions such as RNA-binding and RNA catalytic activities. For models trained on the full feature set, predictions were also associated with chromatin and protein methylation processes.

With regards to novel predictions, we first selected the top hundred genes predicted by the GB models to associate with RNA methylation pathways as candidates for further validation (Supplementary Data 4). To evaluate these predictions with respect to previously known RNA methylation genes, we first performed a hierarchical clustering analysis of predicted plus positive (Class 1) genes based on the machine learning data used here (Supplementary Fig. 4). As anticipated, known and predicted genes were well clustered together, with no evident split between known and predicted RNA methylation genes. Note, however, that an unsupervised clustering approach of all human genes based on the same features used in our supervised modelling analyses was not sufficient on its own for identifying novel genes involved in RNA methylation pathways, as positive genes did not group together in a single cluster (Supplementary Fig. 5).

Second, we interrogated the STRING database[21] for independent Protein–Protein Interaction (PPI) information on known RNA methylation genes and other genes of the human genome. We built a PPI network based on interactions with a confidence score of 400 or above, and performed Random Walks starting from proteins known to mediate methylation of RNAs (Class 1). This allowed us to weigh all other proteins in the network and rank them by their importance relative to our positive gene set. To evaluate whether genes predicted by our models were highly ranked among important interactors, we performed Gene Set Enrichment Analysis (GSEA) using the PageRank score as an input. We obtained a strong positive enrichment (NES = 1.605, $P = 0.0001$) for the model predictions (Supplementary Data 5), corroborating their close functional association with RNA methylation pathways based on independent PPI evidence (Fig. 4).

Insights into the role of new predictions. To gain functional insights into the role of newly predicted genes with regards to previously annotated RNA methyltransferases and associated proteins, we interrogated the STRING database for available PPI data connecting our model predictions to known RNA methylation genes. Our search unravelled a dense network of interactions (Fig. 5a), comprising 2,450 edges (confidence ≥ 400). To further dissect these PPI data and identify subgroups of proteins associated with specific pathways, we employed the Louvain method of community detection[22]. We identified six communities in total (Fig. 5b), which we annotated using a large collection of functional annotation resources[23].

Community 1 (C1, Fig. 5b) groups most RNA methylation genes from the positive set, together with 10 model predictions: *CTU2*, *FARS2*, *HEMK1*, *KARS*, *MOCS3*, *MTO1*, *N6AMT1*, *PUS1*, *PUS3* and *TRNT1*. Functional analysis of community members showed that proteins comprising this sub-network are significantly enriched in the functions of tRNA modification (GO:0006400, $P = 5.09E{-}70$), tRNA methylation (GO:0030488, $P = 6.31E{-}66$), and tRNA processing (Reactome R-HSA-72306, $P = 4.10E{-}45$). Indeed, four predictions in the cluster, CTU2, MOCS3, PUS1 and PUS3, are RNA modifying enzymes mediating tRNA modifications. CTU2 and MOCS3 are involved in 2-thiolation of mcm^5S^2U at wobble positions of tRNAs, whereas PUS1 and PUS3 belong to the tRNA pseudouridine synthase TruA family and mediate the formation of pseudouridine at positions 27/28 and 38/39 of certain tRNAs, respectively[13]. Among other members of the same community, the gene *TRNT1* encodes the mitochondrial CCA tRNA nucleotidyltransferase 1 responsible for the addition of the conserved 3′-CCA sequence to tRNAs. It has been previously

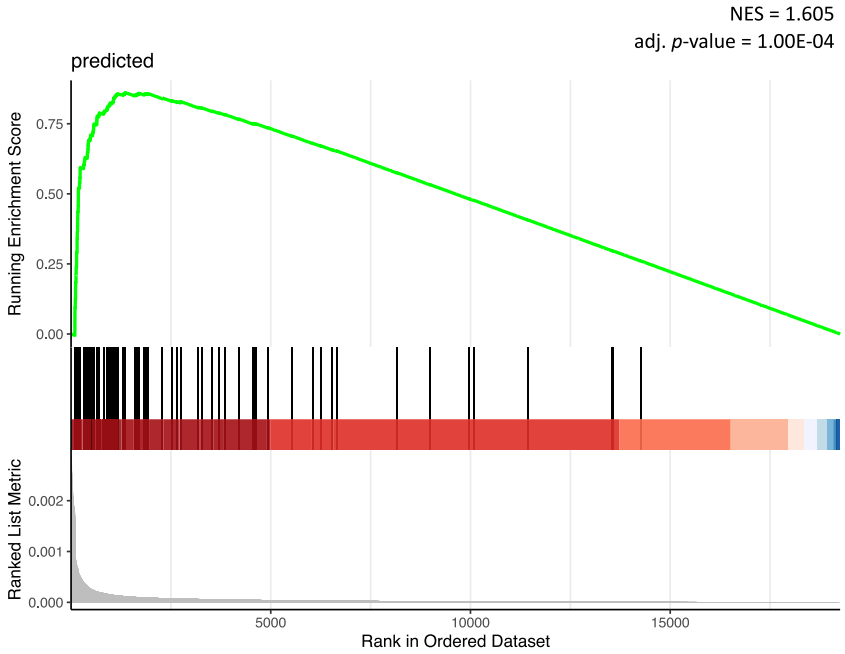

Fig. 4 GSEA analysis of model predictions based on PageRank score. Personalised PageRank score of all human genes was computed using PPI data from STRING, starting from previously known RNA methylation genes. Black vertical lines in the middle plot indicate the position of model predictions in the resulted ranking. A strong positive enrichment (NES = 1.605, $P = 0.0001$) was obtained for predicted genes, corroborating a close functional association with RNA methylation pathways.

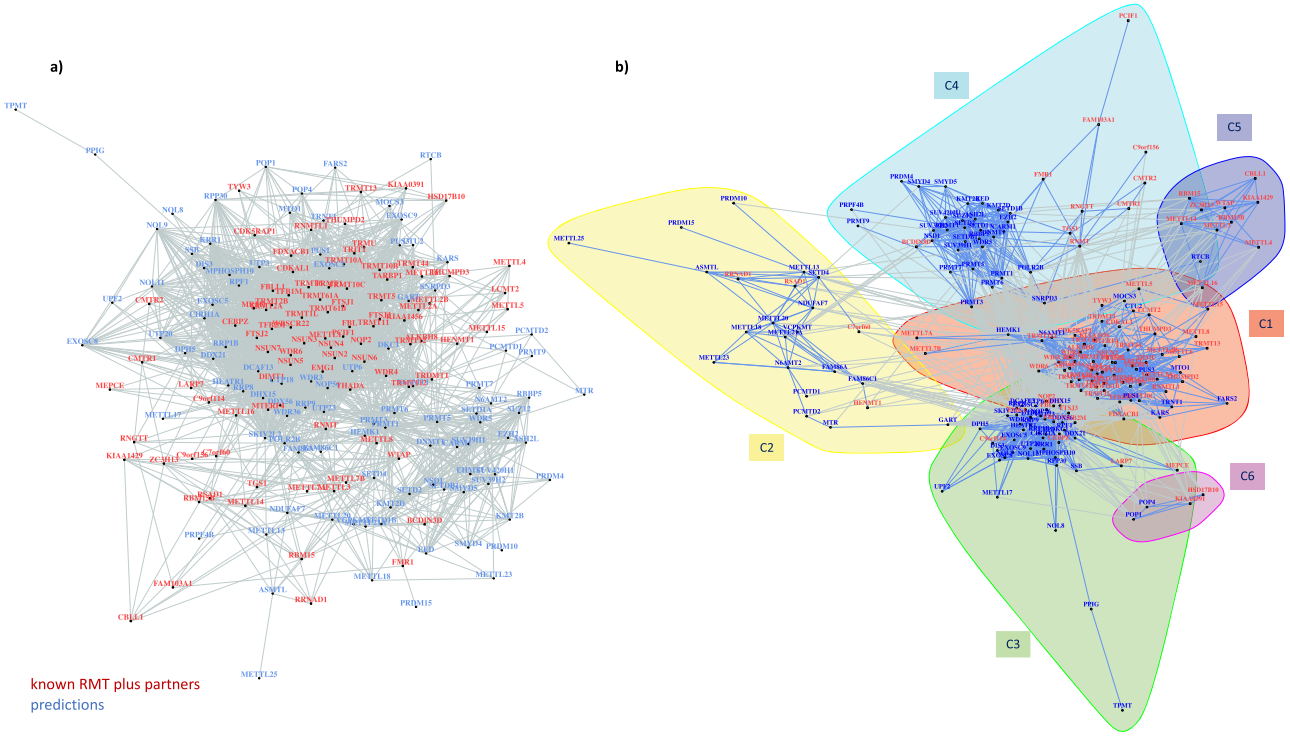

Fig. 5 PPI network of known and predicted genes involved in RNA pathways. a Network based on available PPI data connecting newly predicted genes with previously annotated RNA methyltransferases and associated proteins. **b** Subgroups of proteins associated with specific pathways, as inferred using the Louvain method of community detection.

reported that the presence of the 3′-CCA tail on tRNA is required for target recognition by the tRNA methyltransferase NSUN6[24], which could underlie the functional link of TRNT1 with RNA methylation genes in our analyses.

Likewise, two aminoacyl-tRNA synthetases, FARS2 and KARS, were also predicted to be closely associated with RNA methylation pathways and were part of Community 1. FARS2 is a mitochondrial Phenylalanine-tRNA ligase, responsible for the charging of tRNA(Phe) with phenylalanine. *KARS* encodes a Lysyl-tRNA ligase. Although, we have not found any orthogonal evidence linking FARS2 to RNA methylation, KARS has been previously inferred to physically interact with the RNA

methyltransferase TRMT1, based on co-fractionation data (source BioGRID[25]).

The same sub-network also included two HemK methyltransferases, HEMK1 and N6AMT1. The former is a N5-glutamine methyltransferase responsible for the methylation of the glutamine residue in the GGQ motif of the mitochondrial translation release factor MTRF1L[26]. N6AMT1 methylates the eukaryotic translation termination factor 1 (eRF1) on Gln-185. Notably, it has been reported that N6AMT1 forms the catalytic subunit of a heterodimer with the RNA methyltransferase TRMT112[27], suggestive of a functional interplay between RNA methylation and post-translational modifications of translation factors.

Our models also predicted that *MTO1* is a gene functionally associated with RNA methylation pathways. Previous studies have shown that *MTO1* encodes for a mitochondrial protein which is indeed involved in the 5-carboxymethylaminomethyl modification (mnm^5s^2U34) of the wobble uridine base in mitochondrial tRNAs, with a crucial role in translation fidelity[28].

Community 2 (C2, Fig. 5b) consists mainly of newly predicted genes, associated with four genes from the positive set: *C7orf60*, *HENMT1*, *RRNAD1* and *RSAD1*. The gene *C7orf60* or *BMT2* encodes a probable S-adenosyl-L-methionine-dependent methyltransferase. Recent studies have suggested that BMT2 (also known as SAMTOR) acts as an inhibitor of mTOR complex 1 (mTORC1) signalling in human, a SAM sensor signalling methionine sufficiency[29]. In yeast, BMT2 is responsible for the m^1A2142 modification of 25S rRNA[30]. Two other methyltransferase genes in the same cluster were *RRNAD1* and *HENMT1*. The former encodes for ribosomal RNA adenine dimethylase domain containing 1, but little is known about its function. HENMT1 is a small RNA methyltransferase that adds a 2′-O-methyl group at the 3′-end of piRNAs, contributing to the maintenance of Transposable Element (TE) repression in adult germ cells[31]. Functional annotation of this community indicated an enrichment in peptidyl-lysine methylation function (GO:0018022, $P = 1.92E{-}06$), albeit this was based on only four proteins out the 23 forming this cluster (SETD4, VCPKMT, METTL21A, and METTL18). Among members of this community, we identified proteins with a role in methylation of other substrates. For example, FAM86A catalyses the trimethylation of the elongation factor 2 (eEF2) at Lys-525[32]. METTL13 is also a methyltransferase responsible for the dual post-translational methylation of the elongation factor 1-alpha (eEF1A) at two positions (Gly-2 and Lys-55), modulating mRNA translation in a codon-specific manner[33]. Both genes are involved in modifying translation elongation factor residues, same as N6AMT1 mentioned above. Our results hence suggest that post-translational modifications of translation factors and epitranscriptomic changes on RNAs could be interconnected in modulating translational efficiency.

Community 3 (C3, Fig. 5b) comprises 48 protein members, of which 10 are part of our positive set and 38 were predicted by the models. Overall, we found a strong enrichment for functional terms linked to ncRNA processing (GO:0034470, $P = 6.79E{-}40$) and rRNA processing (R-HSA-72312, $P = 1.03E{-}39$). This finding is consistent with previous computational approaches aiming to predict the functional role of m^6A modification sites, which have independently shown a strong connection between RNA methylation and RNA processing[34]. Among Community 3 members for instance, our predictions include five genes encoding for members of the nuclear RNA exosome, *DIS3*, *EXOSC2*, *EXOSC5*, *EXOSC8* and *EXOSC9*. The exosome is known to participate in a wide variety of cellular RNA processing and degradation events preventing nuclear export and/or translation of aberrant RNAs. Exosome function is thus likely to be interlinked with epitranscriptomic marks on RNAs.

We also identified a sub-cluster within the community connecting DIMT1, EMG1, FBL and NOP2 with 15 proteins predicted by our models. All members of the sub-cluster are RNA-binding proteins involved in rRNA modification in the nucleus (R-HSA-6790901, $P = 5.44E{-}36$). *EMG1* encodes for an RNA methyltransferase that methylates pseudouridine at position 1248 in 18S rRNA[35]. Pathway annotation data further suggest that EMG1 together with eight new predictions (CIRH1A, DCAF13, HEATR1, NOL11, UTP3, UTP6, UTP20 and WDR3) are required in pre-18S rRNA processing and ribosome biogenesis. Of these, the *NOL11* gene encodes a nucleolar protein contributing to pre-rRNA transcription and processing[36]. Partial evidence furthermore suggests that NOL11 interacts with the rRNA 2′-O-methyltransferase fibrillarin, FBL, which is involved in pre-rRNA processing by catalysing the site-specific 2′-hydroxyl methylation of pre-ribosomal RNAs[36]. FBL together with RRP9 and NOP56 are part of the box C/D RNP complex catalysing the ribose-2′-O-methylation of target RNAs.

Finally, three novel gene predictions within this community, *DPH5*, *TPMT* and *RRP8*, were previously reported to have SAM-dependent methyltransferase activity. *DPH5* is coding for a methyltransferase that catalyses the trimethylation of the eEF2 as part of the diphthamide biosynthesis pathway, whereas *TPMT* encodes an enzyme that metabolises thiopurine drugs. We cannot rule out that these may be false positives cases, i.e., erroneous predictions that stem from the presence of the SAM-binding domain in the protein. Yet genes mediating post-translational modifications were repeatedly classified as components of RNA methylation pathways by our machine learning models (e.g., *FAM86A* in Community 2). A noteworthy case is RRP8, which in human is reported to bind to H3K9me2 and to probably act as a methyltransferase, yet studies in yeast have shown that the RRP8 homologue is responsible for installing m1A in the peptidyl transfer centre of the ribosome (m^1A645 in 25S)[37].

Community 4 (C4, Fig. 5b) constitutes a large cluster of 42 proteins. Functional analysis of the group indicates that most community members are chromatin modifying enzymes (R-HSA-3247509, $P = 8.74E{-}29$), or are associated in general with chromatin organisation (R-HSA-4839726, $P = 8.74E{-}29$) and histone modification (WP2369, $P = 1.08E{-}23$). Previously known RNA methylation genes in this community were mainly involved in RNA-capping pathways, e.g., *RNMT*, *CMTR1*, *CMTR2*, *FAM103A1*, *TGS1* and *RNGTT*. Recent studies have suggested that there is indeed extensive crosstalk between RNA modifications and epigenetic mechanisms of gene regulation[7,38,39].

Community 5 (C5) and Community 6 (C6) encompass fewer members than the other communities. Community 5 consists of 10 proteins creating a small sub-network of RNA methyltransferases and partner proteins involved in RNA methylation (GO:0001510, $P = 1.91E{-}17$) and mRNA methylation, in particular (GO:0080009, $P = 6.26E{-}16$). Notably, this community captures proteins involved in the m6A pathway, including the m^6A writer complex of METTL3-METTL14 with co-factor WTAP, METTL16 and ZC3H13, as well as the m^6Am writer METTL4[40]. Community 6 is the smallest of all communities with only four protein members, two previously annotated RNA methylation genes, *HSD17B10* and *KIAA0391*, and two predicted genes *POP1* and *POP4*. Functional analysis suggests that all four proteins contribute to tRNA processing (R-HSA-72306, $P = 5.97E{-}09$) and three of them are involved in tRNA 5′-end processing (GO:0099116, $P = 5.32E{-}08$). The *HSD17B10* gene encodes the 3-hydroxyacyl-CoA dehydrogenase type-2, which is involved in mitochondrial fatty acid beta-oxidation. *HSD17B10* is involved in tRNA processing as it also forms a subcomplex of the mitochondrial ribonuclease P together with TRMT10C/MRPP1[41]. This subcomplex, named MRPP1-MRPP2, catalyses the formation of N1-methylguanine and N1-methyladenine at position 9 (m^1G9 and m^1A9, respectively) in tRNAs. *KIAA0391*,

also known as *PRORP*, encodes a catalytic ribonuclease component of mitochondrial ribonuclease P. It appears that POP1 and POP2 are also components of ribonuclease P and contribute to tRNA maturation via 5′-end cleavage.

False positive discoveries. Our machine learning models and analyses have provided a wealth of new information on putative gene networks underpinning RNA methylation in human. However, it is worth noting the limitations of our approach. First, it is uncertain whether employing previous knowledge from functional annotations may have biased model predictions. We addressed this caveat to an extent by using a reduced feature set without annotation features, such as GO terms. When looking at predictions based on models trained on this dataset, machine learning models point to a recurrent theme of this study: that RNA methylation is functionally interconnected to a range of other core cellular functions (Fig. 3 and Supplementary Fig. 3). For example, we found genes encoding chromatin modifiers among the top candidates. The key question here is whether these genes represent false positives, spurred by the hierarchical structure of GO terms or the shared SAM-binding domain. These ambiguous predictions should be subject to further validation, although multiple lines of evidence suggest that this could well be a biologically meaningful result echoing the crosstalk between DNA, RNA and post-transcriptional modification processes.

Second, there is no trivial way to control for false positives. Because only a few writer enzymes are to date known to deposit methyl-marks on RNA[6], we started from a very limited number of positive (and by consequence negative) samples to use for machine learning. Even though model performance based on test data was good, the small sample sizes may have hampered how well our models generalise. To better illustrate this limitation, we revisit our model predictions on test data, comprised of 18 positive and 5,368 negative genes in total. When considering the top 25 predictions (an equivalent fraction to the top 100 out of all predictions reported above), the False Positive Rate is very low at 0.002 (Supplementary Fig. 6). Nonetheless, because the number of total positives in our test data is also very low, the False Discovery Rate, defined as the false positives (FP) divided by the sum of false positives (FP) and true positives (TP), at this probability window is 0.56. Our machine learning models thus overpredict genes associated with RNA methylation pathways, where only a small fraction of the human genes plays a role in RNA methylation. To address this important caveat, we sought for independent evidence by mining human PPI data to corroborate that newly predicted genes are indeed associated with RNA methylation pathways.

For the afore-mentioned reasons and as a guidance for the interpretation of our results, for each candidate RNA methylation gene, we provide its predicted probability score across all machine learning models -derived with or without using functional annotation data-, as well as its PageRank score from the PPI network analysis (Supplementary Data 5). A gene with a consistently high probability score across multiple models, along with a high rank in the human PPI network, is less likely to represent a modelling artefact.

Conclusions

RNA methylation is a key modulator of transcript stability, splicing and translation efficiency, playing a critical role in cellular homoeostasis and disease[4]. Yet, its molecular underpinnings remain to date poorly understood[11]. Here, we aimed to gain novel insights into genes associated with RNA methylation pathways in human using machine learning approaches. Specifically, we analysed available transcriptomic, proteomic, structural and protein–protein interaction data in a supervised machine learning framework.

Our machine learning models showed very good performance on unseen test data, reaching high accuracy (91%), precision (90%) and recall (92%). A priori gene knowledge (e.g., GO annotations) together with expression data constituted the most informative data types in predictive modelling. Notably, in certain tissues, such as blood, heart, pancreas and brain, genes mediating RNA methylation seemed to show an up- or down-regulated expression profile.

Using independent PPI data, we orthogonally validated top model predictions by corroborating close functional links to previously known RNA methylation genes. Community detection delineated six molecular subnetworks, with distinct roles in tRNA processing (C1, C6), rRNA processing (C3), mRNA methylation (C5), but also protein (C2) and chromatin modifications (C4). Network analyses suggested that deposition of methyl marks on tRNAs is co-orchestrated with other modification processes, such as 2-thiolation and pseudouridine formation. Similarly, rRNA methyltransferases appeared functionally linked to several genes involved in rRNA processing and ribosomal biogenesis. Intriguingly, RNA-capping enzymes were clustered with chromatin modifiers, raising the hypothesis of a crosstalk between the two processes. Our results further indicate that post-translational modifications of translation factors and epitranscriptomic changes on RNAs are intertwined in modulating translational efficiency. Overall, our study exemplifies how access to omics datasets joined by machine learning methods can be used to infer molecular pathways and novel gene function.

Methods

Dataset assembly and pre-processing. To assemble a machine learning dataset for predicting genes involved in RNA methylation process in the human genome, we first curated a list of previously known RNA methylation genes. For this, we performed searches in standard functional annotation resources, such as ExPASy ENZYME (https://enzyme.expasy.org/), InterPro (https://www.ebi.ac.uk/interpro/) and the GO Resource (http://geneontology.org/), in conjunction with a comprehensive literature review for annotated RNA methyltransferases following up on the pioneering paper of Schapira[6]. This allowed us to identify 92 proteins involved —or putatively involved—in RNA methylation to use for machine learning modelling (Supplementary Data 1).

To obtain informative features for classifying gene functions, we interrogated the Harmonizome database[15]. Harmonizome provides a large collection of the pre-processed datasets for genes and proteins, with ~72 million attributes (functional associations) from over 70 major online resources. In particular, data were initially standardised as such continuous-valued datasets to range from 0 to 1, or −1 to 1, where 1 indicates strong positive gene-feature association, 0 indicates no observed gene-feature association, and -1 indicates strong negative gene-feature association (e.g. down-regulation for gene expression datasets). Binary and tertiary datasets were then derived by retaining the top 10% of the strongest gene-feature associations. We selected 15 of these one-hot-encoded datasets from four broad categories: (i) transcriptomics; (ii) proteomics; (iii) structural or functional annotations; and (iv) physical interactions (Supplementary Data 2). In particular, from omics experiments, we sampled BioGPS[16], GTEx[18], HPA[19] and TISSUES[20] gene and protein expression profile data. From functional datasets, we considered GO annotations and InterPro structural domains. Finally, from physical interactions datasets, we selected KEGG and Reactome Pathways, as well as Hub Proteins and Pathway Commons. Collating these data yielded an initial matrix of 26,935 genes and 50,176 one-hot-encoded features ("full feature set"). In addition, we compiled a second dataset of reduced dimensionality, by excluding all 5,148 GO and InterPro annotation features ("reduced feature set").

Problem framing, model definition, training and evaluation. To estimate the probability of a gene being associated with RNA methylation, we used standard machine learning approaches for binary classification. We labelled the 92 previously known RNA methylation genes as positive samples (Class 1), and split them into two sets comprising: (i) 80% of the data for training and cross-validation ($n = 74$) and (ii) 20% kept unseen for model testing ($n = 18$). We considered the remaining genes of the human genome as negative samples (Class 0) and performed an analogous 80/20 split into training/cross-validation ($n = 21,476$) and test sets ($n = 5368$). The underlying assumption here is that the vast majority of genes in the human genome serve other functions, thus the number of false negatives in the training data should be very small.

To produce balanced sets of training samples, and to later reduce the variance of our final models through averaging, negative genes kept for training ($n = 21,476$) were further divided into sets of 74—equal to the number of positive samples for training. We thus generated 290 training sets, where the positive class remained fixed and the negative class was represented by a random draw of an equal number of genes from the rest of the genome, sampling each gene once.

Starting with 290 training sets and our unprocessed Harmonizome data comprising 50,176 features, we next performed filtering to remove low-information features. We removed features with (i) zero values in more than 70% of the samples in each training set, or (ii) less than 16% variance in at least one training set. The selected features for each of the 290 training sets were then merged into a final list of features for model training and testing. We followed the exact same selection process for the reduced feature set as well.

We next considered five types of machine learning models for binary classification: Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB) models. We used grid search and 3-fold cross-validation on each training set for the SVM hyperparameter tuning of the kernel function (linear or RBF), cost parameter, and kernel bandwidth (RBF kernel only). For RF, we used grid search to determine the optimal number of trees in the forest, followed by a randomised search to select the best parameters for maximum number of features considered for splitting a node, maximum number of levels in each decision tree, minimum number of data points placed in a node before the node is split, and minimum number of data points allowed in a leaf node. Likewise, for the GB model, we performed grid search to optimise the learning rate and number of trees in the forest, and subsequently performed a randomised search to tune the remaining decision tree parameters (see RF). We trained all five predictive models on each of the training sets from the full and reduced feature sets, respectively. The performance of all classifiers was estimated using 10-fold cross-validation using standard performance metrics: accuracy, precision, recall (sensitivity), F1 score and Area Under the Receiver Operating Characteristic Curve (AUROC). Finally, we used GB feature ranking to determine the top 100 most informative features across the ensemble of training sets for the full and reduced feature sets, respectively.

Final model testing on test dataset and genome-wide prediction. Once the best model ensemble was selected based on cross-validation, we tested its performance on unseen data. Analogous to the procedure described above for training data, we generated 298 testing datasets, by splitting the negative genes kept for testing into equal sets of 18 genes, and combining them with the 18 of positive samples previously retained. Each model from the classifier ensemble was evaluated on each of the test datasets using accuracy, precision, recall, F1 score and AUROC. Overall performance was calculated by averaging results of all models across test sets.

Likewise, the prediction probability of each human gene was calculated by averaging probability scores for Class 1 across all models of the ensemble. Most non-Class 1 genes (all except the test cases) were part of the negative samples in the training data of exactly one model in the ensemble; however, due to the high number of models (290) the effects of this on the final predictions is expected to be negligible.

All visualisations and meta-analyses were performed using the R software environment (v. 4.0.5)[42]. A heatmap of known and predicted RNA methylation genes across all features used for machine learning was generated using the R package pheatmap. Further in silico validation of model predictions was performed using GO enrichment analyses of predicted genes within the domain "Biological Process" using the package clusterProfiler[43]. Protein–Protein Interaction (PPI) data for human were obtained from STRING (v.11.0)[21] and filtered to interactions with a combined score of 400 and above. All network analyses were performed using the igraph R package[44]. Functional annotation of PPI communities was performed using EnrichR[23].

Statistics and reproducibility. Sample sizes and statistical parameters used in each analysis are indicated in the relevant methods and results sections, as well as in the figure legends when applicable. All statistical analyses were performed in R (v4.0.5).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All datasets used in this study are listed in the Supplementary Data 2, and are publicly available in the Harmonizome database (https://maayanlab.cloud/Harmonizome/).

Code availability

The Python scripts used for the machine learning analyses are available at GitHub https://github.com/storm-therapeutics/ML_RNA_methylation.

Received: 10 January 2022; Accepted: 8 August 2022;
Published online: 25 August 2022

References

1. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. *Cell* **169**, 1187–1200 (2017).
2. Boccaletto, P. et al. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* **46**, D303–D307 (2018).
3. Barbieri, I. & Kouzarides, T. Role of RNA modifications in cancer. *Nat. Rev. Cancer* 20, 303–322 (2020).
4. Huang, H., Weng, H., Deng, X. & Chen, J. RNA modifications in cancer: Functions, mechanisms, and therapeutic implications. *Annu. Rev. Cancer Biol.* **4**, 221–240 (2020).
5. Delatte, B. et al. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* **351**, 282–285 (2016).
6. Schapira, M. Structural chemistry of human RNA methyltransferases. *ACS Chem. Biol.* **11**, 575–582 (2016).
7. Tzelepis, K., Rausch, O. & Kouzarides, T. RNA-modifying enzymes and their function in a chromatin context. *Nat. Struct. Mol. Biol.* **26**, 858–862 (2019).
8. Copeland, R. A., Olhava, E. J. & Scott, M. P. Targeting epigenetic enzymes for drug discovery. *Curr. Opin. Chem. Biol.* **14**, 505–510 (2010).
9. Shi, H., Chai, P., Jia, R. & Fan, X. Novel insight into the regulatory roles of diverse RNA modifications: Re-defining the bridge between transcription and translation. *Mol. Cancer* **19**, 78 (2020).
10. Chou, H.-J., Donnard, E., Gustafsson, H. T., Garber, M. & Rando, O. J. Transcriptome-wide analysis of roles for tRNA modifications in translational regulation. *Mol. Cell* **68**, 978–992.e4. (2017).
11. Frye, M., Jaffrey, S. R., Pan, T., Rechavi, G. & Suzuki, T. RNA modifications: what have we learned and where are we headed? *Nat. Rev. Genet.* **17**, 365–372 (2016).
12. Jonkhout, N. et al. The RNA modification landscape in human disease. *RNA* **23**, 1754–1769 (2017).
13. de Crécy-Lagard, V. et al. Matching tRNA modifications in humans to their known and predicted enzymes. *Nucleic Acids Res.* **47**, 2143–2159 (2019).
14. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* **10**, 35 (2017).
15. Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, baw100 (2016).
16. Wu, C. et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10**, R130 (2009).
17. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
18. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
19. Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
20. Palasca, O., Santos, A., Stolte, C., Gorodkin, J. & Jensen, L. J. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database* **2018**, bay003 (2018).
21. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
22. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
23. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
24. Haag, S. et al. NSUN6 is a human RNA methyltransferase that catalyzes formation of m5C72 in specific tRNAs. *RNA* **21**, 1532–1543 (2015).
25. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
26. Ishizawa, T., Nozaki, Y., Ueda, T. & Takeuchi, N. The human mitochondrial translation release factor HMRF1L is methylated in the GGQ motif by the methyltransferase HMPrmC. *Biochem. Biophys. Res. Commun.* **373**, 99–103 (2008).
27. Li, W., Shi, Y., Zhang, T., Ye, J. & Ding, J. Structural insight into human N6amt1–Trm112 complex functioning as a protein methyltransferase. *Cell Discov.* **5**, 1–13 (2019).
28. Tischner, C. et al. MTO1 mediates tissue specificity of OXPHOS defects via tRNA modification and translation optimization, which can be bypassed by dietary intervention. *Hum. Mol. Genet.* **24**, 2247–2266 (2015).
29. Gu, X. et al. SAMTOR is an S-adenosylmethionine sensor for the mTORC1 pathway. *Science* **358**, 813–818 (2017).
30. Sharma, S., Watzinger, P., Kötter, P. & Entian, K.-D. Identification of a novel methyltransferase, Bmt2, responsible for the N-1-methyl-adenosine base

modification of 25S rRNA in Saccharomyces cerevisiae. *Nucleic Acids Res.* **41**, 5428–5443 (2013).
31. Lim, S. L. et al. HENMT1 and piRNA stability are required for adult male germ cell transposon repression and to define the spermatogenic program in the mouse. *PLOS Genet.* **11**, e1005620 (2015).
32. Davydova, E. et al. Identification and characterization of a novel evolutionarily conserved Lysine-specific methyltransferase targeting eukaryotic translation elongation factor 2 (eEF2) *. *J. Biol. Chem.* **289**, 30499–30510 (2014).
33. Jakobsson, M. E. et al. The dual methyltransferase METTL13 targets N terminus and Lys55 of eEF1A and modulates codon-specific translation rates. *Nat. Commun.* **9**, 1–15 (2018).
34. Wu, X. et al. m6Acomet: large-scale functional prediction of individual m6A RNA methylation sites from an RNA co-methylation network. *BMC Bioinform.* **20**, 223 (2019).
35. Meyer, B. et al. The Bowen–Conradi syndrome protein Nep1 (Emg1) has a dual role in eukaryotic ribosome biogenesis, as an essential assembly factor and in the methylation of Ψ1191 in yeast 18S rRNA. *Nucleic Acids Res.* **39**, 1526–1537 (2011).
36. Freed, E. F., Prieto, J.-L., McCann, K. L., McStay, B. & Baserga, S. J. NOL11, Implicated in the pathogenesis of North American Indian childhood cirrhosis, Is required for Pre-rRNA transcription and processing. *PLOS Genet.* **8**, e1002892 (2012).
37. Shima, H. & Igarashi, K. N1-methyladenosine (m1A) RNA modification: the key to ribosome control. *J. Biochem. (Tokyo)* **167**, 535–539 (2020).
38. Kan, R. L., Chen, J. & Sallam, T. Crosstalk between epitranscriptomic and epigenetic mechanisms in gene regulation. *Trends Genet.* **38**, 182–193 (2021).
39. Huang, H. et al. Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. *Nature* **567**, 414–419 (2019).
40. Chen, H. et al. METTL4 is an snRNA m6Am methyltransferase that regulates RNA splicing. *Cell Res* **30**, 544–547 (2020).
41. Vilardo, E. et al. A subcomplex of human mitochondrial RNase P is a bifunctional methyltransferase—extensive moonlighting in mitochondrial tRNA biogenesis. *Nucleic Acids Res.* **40**, 11583–11593 (2012).
42. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
43. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).
44. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* Complex Systems **1695**, 1–9 (2006).

Acknowledgements

The authors are thankful to Adrián Rodríguez-Bazaga for his valuable input on the machine learning analyses, and Woochang Hwang for his feedback on the network analyses.

Author contributions

G.T. designed and performed the analysis, supervised by H.W. and N.H. H.W. and N.H. conceived the study. H.S.S., A.A., D.L., O.R., and T.K. contributed to data collection and data interpretation. G.T. wrote and revised the paper, with help from H.W., D.L., N.H., and input from all other authors. All authors read and approved the paper.

Competing interests

G.T., D.L., O.R. and H.W. are employees of Storm Therapeutics. N.H. is a co-founder of KURE.ai and CardiaTec Biosciences. T.K. is a co-founder of Abcam and Storm Therapeutics.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03821-y.

Correspondence and requests for materials should be addressed to Georgia Tsagkogeorga, Hendrik Weisser or Namshik Han.

Peer review information *Communications Biology* thanks Achraf El Allali, Chao Zhang and Kunqi Chen for their contribution to the peer review of this work. Primary Handling Editor: Gene Chong.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.