








communications biology

文章

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

打开

利用机器学习预测与RNA甲基化途径相关的基因

Georgia Tsagkogeorga¹, Helena Santos-Rosa³, Andrej Alendar³, Dan Leggate¹,
Oliver Rausch¹, Tony Kouzarides^{2,3}, Hendrik Weisser¹,⁵ & Namshik Han²,^{4,5}

RNA甲基化在RNA的功能调控中起着重要作用，因此在生物学和药物发现中引起了越来越多的兴趣。在此，我们从Harmonizome数据库中收集和整理了转录组、蛋白质组、结构和物理相互作用数据，并应用监督机器学习来预测与人类RNA甲基化途径相关的新基因。我们选择了五种分类器，在多个训练集上使用交叉验证法进行训练和评估。最好的模型在交叉验证的基础上达到88%的准确率，在测试集上的平均准确率为91%。利用蛋白质-蛋白质相互作用的数据，我们提出了六个分子子网络，将模型预测与先前已知的RNA甲基化基因联系起来，它们在mRNA甲基化、tRNA加工、rRNA加工以及蛋白质和染色质修饰中发挥作用。我们的研究证明了如何通过机器学习方法获取大型全息数据集来预测基因功能。

¹ STORM Therapeutics Ltd, Babraham Research Campus, Cambridge, UK. ² 剑桥大学Milner Therapeutics研究所, 英国剑桥, Puddicombe Way. ³ 剑桥大学Gurdon研究所, 英国剑桥, 网球场路。 ⁴ 剑桥大学应用数学和理论物理系剑桥医学人工智能中心, 英国剑桥。 ⁵ 这些作者贡献相同。亨德里克-韦瑟和韩南石。✉电子邮件: georgia.tsagkogeorga@stormtherapeutics.com; hendrik.weisser@stormtherapeutics.com; n.han@milner.cam.ac.uk

自20世纪60年代以来，人们就知道NA的修饰，当时对酵母的第一个转移RNA (tRNA) 进行了测序，发现了10种化学修饰的核糖核苷，包括假尿苷 (Ψ)¹。从那时起，已确定的修饰数量已增加到150多个，在所有三个生命王国的编码和非编码RNA上都有发现²。该领域的技术进步已经确定，RNA修饰是广泛的、可逆的和动态调节的¹。甲基化是，是最丰富的类型，甲基组装饰多个RNA物种，如信使RNA (mRNA)，核糖体RNA (rRNA) 和tRNA，在不同的核苷和位置。到目前为止，N6-甲基腺苷 (m6A) 是研究最多的修饰，通常在mRNA、rRNA、长基因间非编码RNA (lincRNA)、初级微RNA (pri-miRNA) 和小核RNA (snRNA) 中检测到。

其他甲基标记包括5-甲基胞嘧啶 (m5C)、N1-甲基腺苷 (m1A)、7-甲基鸟苷 (m7G)、2'-甲基鸟苷 (m2'G)、O-二甲基腺苷 (m6Am) 和 5-羟甲基胞嘧啶 (hm5C)³⁻⁵。

RNA上的甲基标记的沉积是由作家的酶催化的，被称为RNA甲基转移酶。到目前为止，在人类基因组中共发现了57种RNA甲基转移酶。其中，5种甲基化mRNAs，6种小RNAs，14种rRNAs和22种tRNAs，而12种仍有未知底物⁶。大多数酶使用S-腺苷酸-蛋氨酸 (SAM) 作为RNA底物的甲基供体，同时许多酶还招募附属蛋白，这些附属蛋白通常对底物的结合、定位和稳定性至关重要。研究最充分的RNA甲基化作者的例子是迄今为止负责沉积m6A的METTL3-METTL14复合物，其次是在tRNA上沉积m5C的含NOL1/NOP2/Sun (NSUN) 结构域的tRNA修饰酶家族⁷。

通过化学修饰对RNA进行动态调控，最近引起了人们对RNA修饰酶作为新的潜在治疗目标的兴趣⁸。这是因为多条证据表明，RNA甲基化在细胞功能中发挥的作用远比以前想象的要重要。与此相呼应，一些研究表明，RNA甲基化是转录本稳定性、基因表达、剪接和翻译效率的一个关键调节因素⁹⁻¹¹。此外，越来越多的数据表明，RNA甲基化过程的变化可能与一系列癌症、神经系统疾病和其他各种疾病有关¹²。令人惊讶的是，尽管RNA甲基化在细胞平衡和疾病中起着关键作用，但一般来说，RNA甲基化的途径仍然没有得到充分的研究⁷。我们目前对RNA修饰的理解也是非常零散的，估计有20%或更多的RNA修饰酶仍然是未知的或未确定的¹³。

研究新基因功能的常规方法

包括一系列劳动密集型的湿式实验室技术，包括用于描述基因特定表型效应的诱变、基因破坏或基因耗竭 (敲除/淘汰)，以及用于识别分子相互作用的色谱法和质谱法。然而，在过去的20年里，大规模的全基因组学数据的获取使人们能够使用“干”的计算方法来理解生物功能。在功能基因组学的框架下，已经开发了一系列的生物信息学工具，包括用于识别不同物种间具有相似功能的同源基因的方法，以及针对特定序列主题和功能域的全基因组筛选。今天，机器学习技术正在成为一种强大的方法来利用日益丰富的大规模生物数据，允许发现隐藏模式和更可靠的统计预测¹⁴。

在这里，我们旨在利用机器学习更好地了解人类RNA甲基化所涉及的分子途径。为此，我们使用了公开可用的人类转录组。

蛋白质组学、结构学和蛋白质相互作用数据¹⁵，并建立了一个大型机器学习数据集，用于监督下的二元分类。我们训练并评估了五种预测模型的组合。逻辑回归 (LR)、高斯奈夫贝叶斯 (GNB)、支持向量机 (SVM)、随机森林 (RF) 和梯度提升 (GB) 模型。我们采用最好的模型来预测与人类基因组中RNA甲基化途径相关的基因功能。

结果和讨论

数据工程和特征选择。通过挖掘功能注释数据库和广泛的文献搜索，我们确定了92个参与RNA甲基化的蛋白质 (补充数据1)。这些蛋白要么是写甲基的 (已知的RNA甲基转移酶6和它们在蛋白复合物中的伙伴蛋白)，要么是以前被注释为推定的RNA甲基转移酶的酶 (见方法)。在机器学习分析中，这些蛋白质的编码基因构成了我们的阳性类别 (1类)。为了将我们的预测建模框定为二元分类问题，我们通过从剩余的基因组中随机抽取与我们的阳性基因组数量相等的基因，组装了多个分层的训练和测试数据集，确保我们初始数据集的所有基因都正好被抽样一次 (图1)。我们的理由是，这将允许机器学习模型在不同的其他基因功能范围内进行训练和测试，而不是仅仅选择一个功能作为阴性集。此外，这种方法减轻了从人类基因组中抽取单一阴性基因集可能产生的任何假定的偏见。

我们最初收集了50,176个特征，这些特征来自公开的和以前策划的转录组、蛋白质组、功能注释、结构和物理相互作用数据集 (补充数据2)。为了确定对分类有参考价值的特征，从而有助于预测与RNA甲基化相关的基因，我们在模型训练前进行了特征选择，然后在训练和交叉验证后进行特征排序。为了减少特征与样本的比例，首先我们消除了训练数据集中有大量缺失数据的特征。其次，我们删除了具有低方差的特征，这导致最终数据集的维度大幅降低到1505个特征。用于分类的部分特征来自BioGPS16 (35)、Gene Ontology17 (GO : 59)、GTEx18 (1114)、Human Protein Atlas19 (HPA : 107)、InterPro (1)、Pathway Commons (PathCommons : 150) 和TISSUES20 (40) 数据集。

在模型训练和交叉验证过程中，我们通过使用GB的重要性衡量标准来计算特征的重要性，因为所有训练集的平均值。50个信息量最大的特征及其在分类中的相对重要性显示在补充图1中。在全部特征集中具有最高重要性的特征主要是GO术语，如GO:0032259, GO:0016740, GO:0003723, GO:0008168和

GO:0016070，都对应于甲基化、转移酶/甲基转移酶活性和RNA代谢过程。同样，代表S-腺苷-L-蛋氨酸依赖性甲基转移酶超家族的InterPro结构域IPR029063也被列为信息量最大的前50个特征之一 (补充图1a)。虽然是预期的，但分类器似乎依赖于RNA和甲基化相关的注释特征的事实提供了支持，即模型学会了对与RNA甲基化过程有密切联系的基因进行分类。尽管GO注释是有信息量的，但它们同样可能使基因预测偏向于预先存在的功能注释。因此，我们通过排除GO和InterPro数据类型，组装了第二个降维的特征集。当分类器被

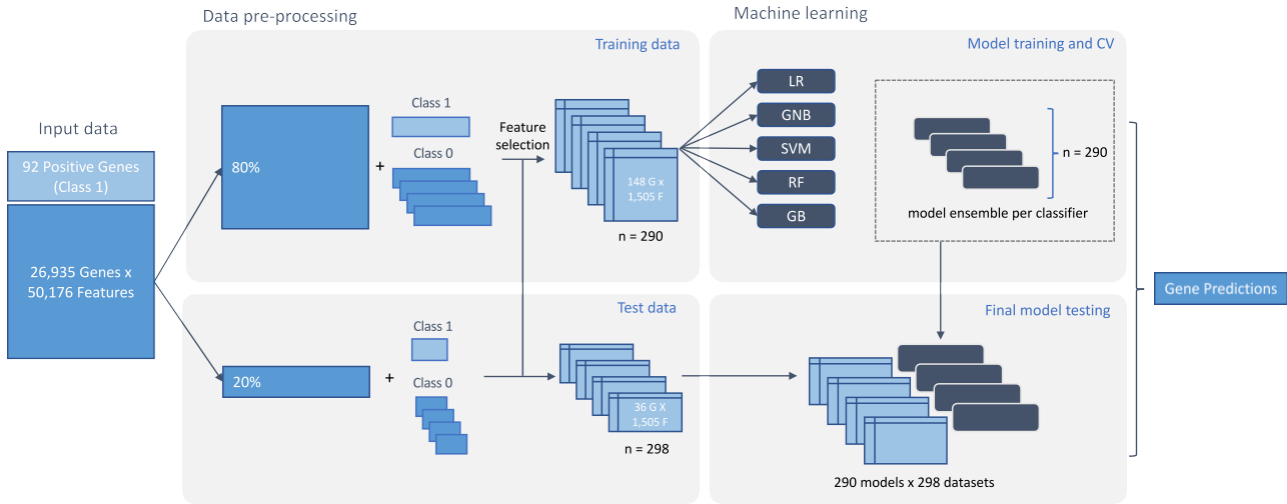


图1 分析工作流程示意图。以前已知的RNA甲基化基因被用作阳性样本（第1类），并分成两组，其中80%的数据用于训练，20%保留未见过的数据用于模型测试。对人类基因组的其余基因也进行了类似的80/20分割，这些基因被进一步分成与阳性样本同等大小的集合，作为阴性样本（0类），以产生分层的特征选择、训练和测试集合。在特征预筛选之后，五种用于二元分类的机器学习模型--逻辑回归（LR）、高斯奈夫贝叶斯（GNB）、支持向量机（SVM）、随机森林（RF）和梯度提升（GB）--被训练在每个训练集上，形成一个分类器集合。在每个测试数据集上对分类器组合中的每个模型进行评估，并通过对所有模型在测试集上的结果进行平均来计算总体性能。表现最好的组合被用来对整个基因组进行预测。

在这个减少的特征集上进行训练，信息量最大的特征类型主要是GTEx表达谱（补充图1b）。GTEx项目旨在提供一个全面的组织特异性基因表达和调控的公共资源，到目前为止，包括来自54个非疾病组织的样本，涉及近1000个人18。组织样本的表达数据被整合到Harmonizome中，因此在此取样，包括每个组织样本中相对于其他组织样本的高或低表达的GTEx组织表达谱数据集的单次编码的基因集。

这种GTEx表达谱特征的高排名的一个可能的解释是，在特定的生物条件下，即在某些组织中，RNA甲基化基因与其他过程相比，倾向于集体下降或上升。另外，GTEx特征的高排名可能是由于GTEx特征在特征集中的高比例和源于训练数据集的高维度的特征与样本比例的噪声。为了进一步研究这个问题，我们计算了GTEx特征在所有训练集的模型中最有信息量的前一百个特征中的相对频率（补充数据3）。值得注意的是，某些取自血液、心脏、胰腺和大脑区域的样本被一百多个模型检索为信息量大。

模型性能。我们选择了五种机器学习分类器（LR、GNB、SVM、RF和GB），并在完整的和减少的特征集的训练集上进行训练，为每个分类器和特征集创建一个模型集合。总的来说，所有五个模型的组合在交叉验证的基础上表现出非常相似的性能（表1）。在使用完整特征集训练的分类器中，GB和RF模型的平均精度最高，分别为0.875和0.870，平均精度也同样高，分别为0.895和0.870。紧随其后的GB模型和RF模型也产生了最高的AUC分数，其平均AUC分别为0.938和0.937。

在没有GO/InterPro注释的情况下，五种分类器对减少的特征集的性能比原来有所下降。

对完整的数据集（表1）。SVM和RF的模型组合在几乎所有指标上都优于其余三个组合。SVM模型在基于交叉验证的减少的特征集上表现最好，平均预测精度为0.812，精确度为0.822，AUROC为0.864。

模型预测。为了评估不同模型和特征集的结果，我们首先比较了所有人类基因的概率分数分布，即由完整特征集训练的模型（图2a）和由缩小的特征集得出的模型（图2b）预测的结果。在所有五种类型的模型组合中，预测情况似乎非常相似，这体现在它们各自分布的广泛重叠上。大多数基因的平均概率得分高度偏向于零，这与大多数人类基因不会直接参与RNA甲基化途径的假设相一致。然而，请注意，GNB模型显示了一个异常的预测曲线，与所有其他分类器相比，在概率一附近有一个明显较高的峰值。

其次，我们评估了不同机器学习模型预测的高置信度参与RNA甲基化的基因之间的重叠程度。在此，我们将第1类概率分布中前1%的所有基因定义为高可信度。对于大多数分类器来说，这部分基因包括了约270个基因，但GNB模型对大量的基因模糊地分配了相同的高概率分数，因此考虑了更大的集合（>1750个基因）。当比较不同模型的顶级预测时，它们的相对一致性对两个特征集都很高（补充图2）。我们确定了280多个基因，这些基因通常被预测为参与RNA甲基化途径，由五种模型组合中的至少三种在选定的一致性临界点上进行预测（补充图2a, b）。

最后，为了获得对预测的高层次理解根据不同的模型，我们使用上述相同的高可信度基因进行了探索性的GO富集分析。图3中比较了每个机器学习模型的前10个富集术语。所有的模型组合，独立于它们的数据集

| 表1 基于10倍交叉验证的模型性能。 | | | | | |
|---|----------------|---------------|----------------|---------------|---------------|
| | 准确度 | 精度 | 召回率 | F1 | AUC |
| 完整的功能设置 | | | | | |
| GB0 | .875 ± 0.0250 | .895 ± 0.0330 | .865 ± 0.031 | 0.872 ± 0.025 | 0.938 ± 0.015 |
| 卫星导航系统(GNB) | | 0.851 ± 0.025 | 0.821 ± 0.032 | 0.863 ± 0.021 | 0.862 ± 0.023 |
| | 0.924 ± 0.021 | | | | |
| LR | 0.859 ± 0.0210 | .870 ± 0.025 | 0.859 ± 0.023 | 0.857 ± 0.021 | 0.921 ± 0.015 |
| RF0 | .870 ± 0.0210 | .870 ± 0.026 | 0.886 ± 0.032 | 0.871 ± 0.022 | 0.937 ± 0.014 |
| 证券公司 | 0.856 ± 0.022 | 0.876 ± 0.028 | 0.845 ± 0.027 | 0.852 ± 0.023 | 0.921 ± 0.017 |
| 减少的功能集 | | | | | |
| GB | 0.799 ± 0.0290 | .800 ± 0.035 | 0.819 ± 0.032 | 0.801 ± 0.029 | 0.860 ± 0.031 |
| 卫星导航系统(GNB) | | 0.781 ± 0.022 | 0.765 ± 0.0280 | 0.792 ± 0.024 | 0.800 ± 0.021 |
| | .840 ± 0.043 | | | | |
| LR | 0.795 ± 0.030 | 0.797 ± 0.035 | 0.814 ± 0.030 | 0.797 ± 0.029 | 0.857 ± 0.032 |
| RF0 | .805 ± 0.0240 | .802 ± 0.033 | 0.833 ± 0.023 | 0.809 ± 0.022 | 0.867 ± 0.025 |
| 证券公司 | 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 Naive Bayes, SVM Support Vector Machine, RF Random Forest, GB Gradient Boosting. | | | | | |

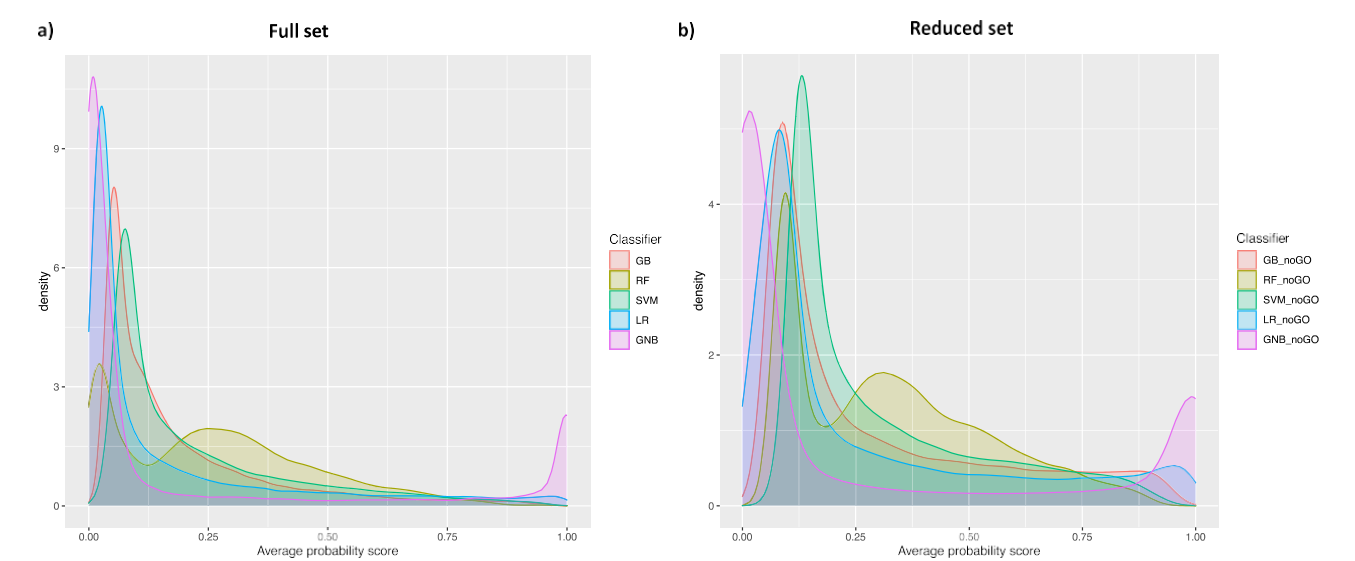


图2 所有人类基因的概率分数分布。第1类的概率分数分布，是基于(a)完整和(b)完整的预测模型而得到的。
(b) 减少的特征集。

衍生出来的，产生了富含与RNA结合和RNA催化活动相关的GO术语的预测。请注意，在完整的特征集上训练的模型的顶级富集结果还包括与染色质和蛋白质甲基化过程相关的术语（图3）。在缩小的特征集上训练的模型中，只有GNB和SVM显示了蛋白质甲基化的富集。这可能是一个建模的伪命题，即预测结果被错误地分配到第1类，这可能是由GO术语的层次性造成的（例如，“甲基化”是“RNA甲基化”和“蛋白质甲基化”过程的父术语）。然而，除GB以外的所有模型都预测了以前与DNA上的催化活动有关的基因。因此，另一种解释是，我们的模型捕捉到了在不同底物上运作的修饰途径之间的假定功能联系。总的来说，功能注释分析为模型性能提供了一个良好的定性控制。这里的理由是，尽管我们没有发现生物术语“RNA甲基化”本身的富集（鉴于模型预测的是“新”基因），但与该术语密切相关的特征应该在GO结果中名列前茅。补充图3提供了按特征集分类的GO富集结果。

硅验证。在所有的分类器中，根据交叉验证，在全部特征集上训练的GB模型表现出最好的性能，因此被选为应用于以前未见过的测试数据。分层测试数据集的模型性能指标是通过组合中每个模型获得的数值进行平均来计算的。GB合集的平均测试集精度为0.905，精度为0.897，召回率为0.923，AUCROC为0.973。

然而，众所周知，高特征与样本的比率可能会导致模型性能的过度和高估。由于我们的机器学习模型是基于少量的阳性基因（第1类），我们对获得我们预测中的假阳性基因的估计特别感兴趣。为了达到这个目的，我们将整个保留数据（18个阳性和5368个阴性例子）汇集在一起，并将GB集合中的每个模型所估计的每个基因的预测概率得分平均化。这样我们就可以通过计算被我们的模型错误地预测为阳性的阴性基因（0类）的数量来估计假阳性。在5368个阴性基因中，有425个被错误地分类，导致估计的假阳性率为0.079，被定义为假阳性（FP）的数量除以假阳性（FP）和真阴性（TN）的总和。

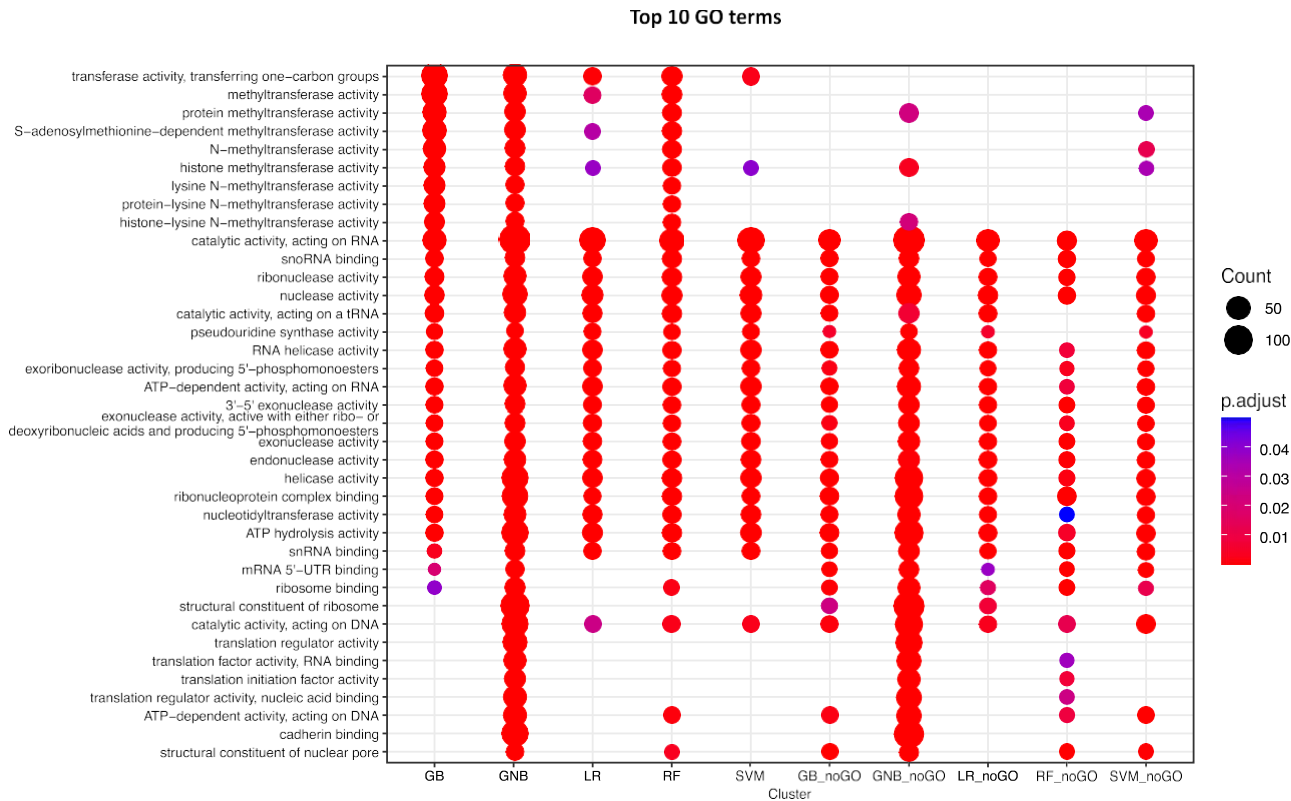


图3 高可信度预测的功能富集分析。比较每个模型组合的第1类概率分布中前1%的基因的GO富集结果。富集度最高的术语包括RNA结合和RNA催化活动等。对于在完整特征集上训练的模型，预测也与染色质和蛋白质甲基化过程有关。

关于新的预测，我们首先选择了GB模型预测的与RNA甲基化途径相关的前一百个基因作为进一步验证的候选基因（补充数据4）。为了评估这些预测与以前已知的RNA甲基化基因的关系，我们首先根据这里使用的机器学习数据对预测的加阳性（第1类）基因进行了分层聚类分析（补充图4）。正如预期的那样，已知和预测的基因很好地聚在一起，在已知和预测的RNA甲基化基因之间没有明显的分裂。然而，请注意，基于我们监督建模分析中使用的相同特征的所有人类基因的无监督聚类方法本身并不足以识别参与RNA甲基化途径的新基因，因为阳性基因并没有聚集在一个集群中（补充图5）。

第二，我们询问了STRING数据库²¹，以了解独立的我们对已知的RNA甲基化基因和人类基因组的其他基因的蛋白质-蛋白质相互作用（PPI）信息进行了研究。我们根据一致性分数在400分以上的相互作用建立了一个PPI网络，并从已知介导RNA甲基化的蛋白质开始进行随机漫步（第1类）。这使我们能够权衡网络中的所有其他蛋白质，并根据它们相对于我们的阳性基因组的重要性进行排序。为了评估我们的模型预测的基因是否在重要的相互作用者中排名很高，我们使用PageRank分数作为输入进行了基因集富集分析（GSEA）。我们得到了模型预测的强烈正向富集（NES=1.605， $P=0.0001$ ）（补充数据5），证实了它们与基于独立PPI证据的RNA甲基化途径的密切功能关联（图4）。

洞察新预测的作用。为了深入了解新预测的基因在先前注释的RNA甲基转移酶和相关蛋白方面的作用，我们在STRING数据库中查询了连接我们的模型预测和已知RNA甲基化基因的可用PPI数据。我们的搜索揭示了一个密集的相互作用网络（图5a），包括2450条边（置信度 ≥ 400 ）。为了进一步剖析这些PPI数据，并确定与特定途径相关的蛋白质子群，我们采用了群体检测的Louvain方法²²。我们总共确定了6个群落（图5b），我们用大量的功能注释资源对其进行注释²³。

社区1（C1，图5b）将大多数RNA甲基化基因从阳性组中分组，同时还有10个模型预测。*ctu2*、*fars2*、*hemk1*、*kars*、*mocs3*、*mtol1*、*n6amt1*、*pus1*。*PUS3*和*TRNT1*。群体成员的功能分析显示，组成该子网络的蛋白质在tRNA修饰（GO:0006400， $P=5.09E-70$ ）、tRNA甲基化（GO:0030488， $P=6.31E-66$ ），和tRNA加工（Reactome R-HSA-72306， $P=4.10E-45$ ）。事实上，集群中的四个预测，CTU2、MOCS3、PUS1和PUS3，是介导tRNA修饰的RNA修饰酶。CTU2和MOCS3参与tRNA摇摆位置的mcm5S2U的2-硫醇化，而PUS1和PUS3属于tRNA假尿苷合成酶TruA家族，分别在某些tRNA的27/28和38/39位置介导假尿苷的形成¹³。在同一家族的其他成员中，*TRNT1*基因编码线粒体CCA tRNA核苷酸转移酶1，负责将保守的3'-CCA序列添加到tRNAs上。它以前曾被

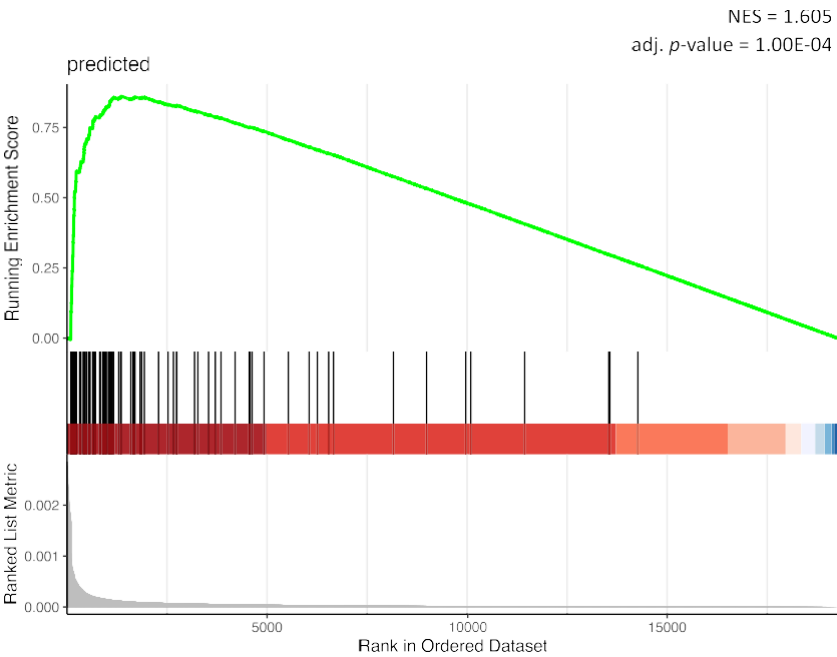


图4 基于PageRank分数的模型预测的GSEA分析。所有人类基因的个性化PageRank分数是使用STRING的PPI数据计算的，从以前已知的RNA甲基化基因开始。中间图中的黑色垂直线表示模型预测在结果排名中的位置。预测的基因获得了强烈的正向富集（NES=1.605， $P=0.0001$ ），证实了与RNA甲基化途径的密切功能关联。

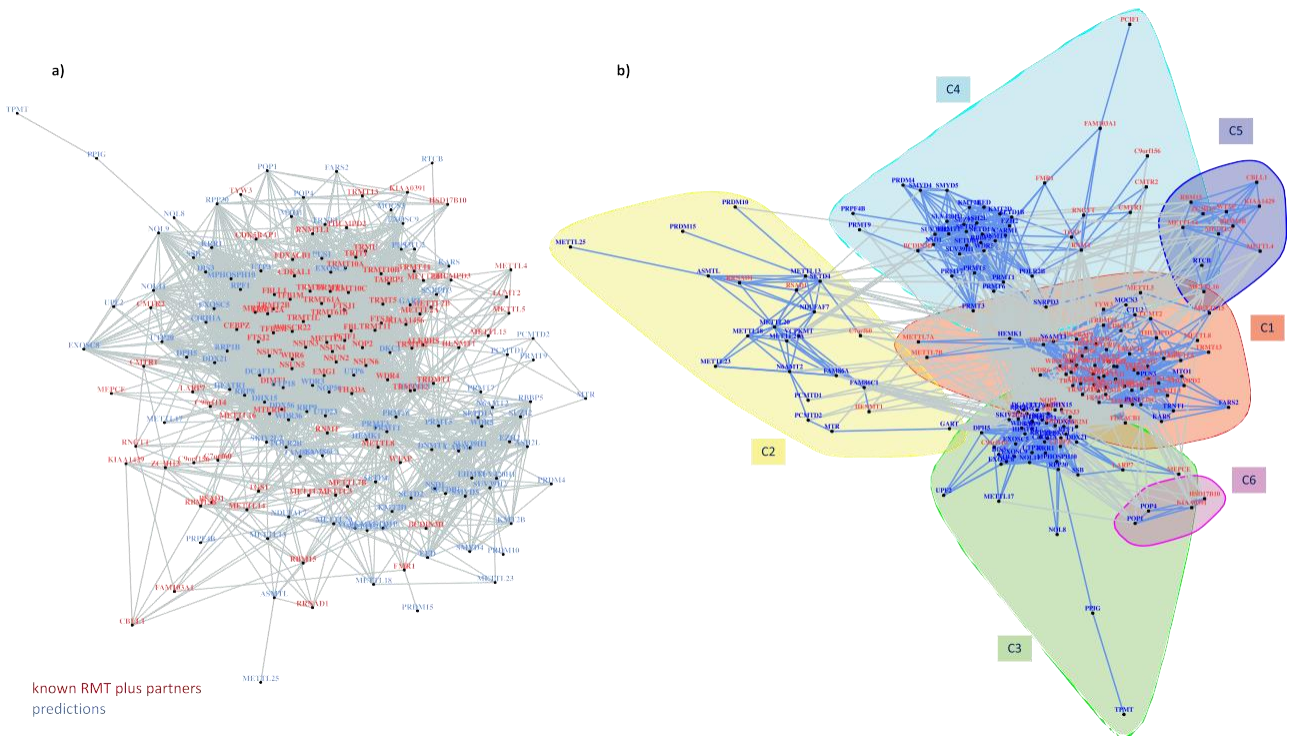


图5 参与RNA途径的已知和预测基因的PPI网络。a 基于现有PPI数据的网络，将新预测的基因与先前注释的RNA甲基转移酶和相关蛋白连接起来。

报告说，tRNA上3'-CCA尾巴的存在是tRNA甲基转移酶NSUN624识别目标所必需的，这可能是我们分析中TRNT1与RNA甲基化基因功能联系的基础。

同样，两个氨基酸-tRNA合成酶，FARS2和KARS，也被预测为与RNA密切相关。

甲基化途径，是社区1的一部分。FARS2是线粒体苯丙氨酸-tRNA连接酶，负责用苯丙氨酸对tRNA（Phe）进行充电。KARS编码一个赖氨酸-tRNA连接酶。尽管我们没有发现任何正交的证据将FARS2与RNA甲基化联系起来，但KARS以前被推断为与RNA的物理相互作用。

根据共同分馏数据（来源：BioGRID25），甲基转移酶 TRMT1。

同一子网络还包括两个 HemK 甲基转移酶，HEMK1 和 N6AMT1。前者是一个 N5-谷氨酰胺甲基转移酶，负责线粒体翻译释放因子 MTRF1L26 的 GGO 基序中谷氨酰胺残基的甲基化。N6AMT1 使真核翻译终止因子 1 (eRF1) 的 Gln-185 发生甲基化。值得注意的是，据报道，N6AMT1 与 RNA 甲基转移酶 TRMT11227 形成异源二聚体的催化亚基，这表明 RNA 甲基化和翻译因子的翻译后修饰之间存在功能上的相互作用。

我们的模型还预测 MTO1 是一个在功能上与 RNA 甲基化途径相关的基因。以前的研究表明，MTO1 编码的线粒体蛋白确实参与了线粒体 tRNAs 中晃动的尿苷碱基的 5-羧甲基化 (mnm5s2U34)，在翻译中具有关键作用²⁸。社区 2 (C2, 图 5b) 主要由新预测的基因组成，与阳性基因组中的四个基因相关。C7orf60、HENMT1、RRNAD1 和 RSAD1。基因 C7orf60 或 BMT2 可能编码一种 S-腺苷-L-蛋氨酸依赖的甲基转移酶。最近的研究表明，BMT2 (也称为 SAMTOR) 在人类中作为 mTOR 复合体 1 (mTORC1) 信号的抑制剂，是 SAM 传感器信号的蛋氨酸充足率²⁹。在酵母中，BMT2 负责 25S rRNA 的 m1A2142 修饰³⁰。同组的另外两个甲基转移酶基因是 RRNAD1 和 HENMT1。前者编码含有核糖体 RNA 腺嘌呤二甲甲基化酶域 1，但对其功能知之甚少。HENMT1 是一种小 RNA 甲基转移酶，在 piRNA 的 3' 端添加 2'-O-甲基，有助于维持成年生殖细胞中的可转义元素 (TE) 抑制³¹。这个群体的功能

注释表明肽基赖氨酸甲基化功能的富集 (GO:0018022, $P=1.92E-06$)，尽管这只是基于形成这个集群的 23 个蛋白质中的四个 (SETD4、VCPKMT、METTL21A 和 METTL18)。在这个群体的成员中，我们确定了在其他底物的甲基化中发挥作用的蛋白质。例如，FAM86A 催化延伸因子 2 (eEF2) 在 Lys-52532 的三甲基化。METTL13 也是一个甲基转移酶，负责延伸因子 1- α (eEF1A) 在两个位置 (Gly-2 和 Lys-55) 的双重翻译后甲基化，以特定密码的方式调控 mRNA 的翻译³³。这两个基因都参与修饰翻译延伸因子的残基，与上面提到的 N6AMT1 相同。因此，我们的结果表明，翻译因子的翻译后修饰和 RNA 上的表观转录组学变化在调控翻译效率方面可能是相互关联的。

社区 3 (C3, 图 5b) 包括 48 个蛋白质成员，其中 10 个是我们阳性集合的一部分，38 个是模型预测的。总的来说，我们发现与 ncRNA 加工 (GO:0034470, $P=6.79E-40$) 和 rRNA 加工 (R-HSA-72312, $P=1.03E-39$) 相关的功能术语有强烈的富集。这一发现与以前旨在预测 m6A 修饰位点功能作用的计算方法是一致的，这些方法独立地显示了 RNA 甲基化和 RNA 加工之间的紧密联系³⁴。例如，在社区 3 成员中，我们的预测包括五个编码核 RNA 外显子的基因，DIS3、EXOSC2、EXOSC5、EXOSC8 和 EXOSC9。众所周知，外显子参与各种细胞 RNA 处理和降解事件，防止异常 RNA 的核出口和/或翻译。因此，外显子体的功能可能与 RNA 上的外显子体标记相互关联。

我们还确定了社区内的一个子集群将 DIMT1、EMG1、FBL 和 NOP2 与 15 种蛋白质联系起来

我们的模型所预测的。该亚群的所有成员都是参与核内 rRNA 修饰的 RNA 结合蛋白 (R-HSA-6790901, $P=5.44E-36$)。EMG1 编码的是一种 RNA 甲基转移酶，能使 18S rRNA35 中 1248 位的假尿苷发生甲基化。路径注释数据进一步表明，EMG1 与八个新的预测 (CIRH1A、DCAF13、HEATR1、NOL11、UTP3、UTP6、UTP20 和 WDR3) 一起，在 18S 前 rRNA 的处理和核糖体的生物生成中需要这些基因。其中，NOL11 基因编码一种有助于前 rRNA 转录和加工的核极蛋白 36。部分证据进一步表明，NOL11 与 rRNA 2'-O-甲基转移酶 FBL 相互作用，FBL 通过催化前核糖体 RNA 的定点 2'-羟基甲基化而参与前 rRNA 的加工³⁶。FBL 与 RRP9 和 NOP56 一起是核 C/D RNP 复合物的一部分，催化目标 RNA 的核糖-2'-O-甲基化。

最后，这个群体中的三个新的基因预测，DPH5、TPMT 和 RRP8，以前被报道具有 SAM 依赖性甲基转移酶活性。DPH5 编码的是一种甲基转移酶，催化 eEF2 的三甲基化，作为双氢胺生物合成途径的一部分，而 TPMT 编码的是一种代谢硫嘌呤药物的酶。我们不能排除这些可能是假阳性的情况，即由于蛋白质中存在 SAM 结合域而导致的错误预测。然而，介导翻译后修饰的基因多次被我们的机器学习模型归类为 RNA 甲基化途径的组成部分 (例如，社区 2 中的 FAM86A)。一个值得注意的例子是 RRP8，据报道它在人类中与 H3K9me2 结合，并可能作为一个甲基转移酶，但在酵母中的研究表明，RRP8 的同源物负责在核糖体的肽基转移中心安装 m1A (25S 中的 m1A645)³⁷。

群落 4 (C4, 图 5b) 构成了一个由 42 个蛋白质组成的大群。该群落的功能分析表明，大多数群落成员是染色质修饰酶 (R-HSA-3247509, $P=8.74E-29$)，或与染色质组织 (R-HSA-4839726, $P=8.74E-29$) 和组蛋白修饰 (WP2369, $P=1.08E-23$) 普遍相关。此前该群体中已知的 RNA 甲基化基因主要参与 RNA 盖帽途径，如 RNMT、CMTR1、CMTR2、FAM103A1、TGS1 和 RNGTT。最近的研究表明，RNA 修饰和基因调控的表观遗传机制之间确实存在广泛的串扰^{7,38,39}。

社区 5 (C5) 和社区 6 (C6) 包含了较少的群体。社区 5 由 10 个蛋白质组成，形成了 RNA 甲基化酶和参与 RNA 甲基化 (GO:0001510) 的伙伴蛋白的小型子网络。群落 5 由 10 个蛋白质组成，形成了一个 RNA 甲基化酶和参与 RNA 甲基化 (GO:0001510, $P=1.91E-17$) 和 mRNA 甲基化 (GO:0080009, $P=6.26E-16$) 的伙伴蛋白的小型子网络。值得注意的是，这个群落捕获了参与 m6A 途径的蛋白质，包括 METTL3-METTL14 与共同因子 WTAP、METTL16 和 ZC3H13 的 m6A 书写者复合物，以及 m6Am 书写者 METTL440。社区 6 是所有社区中最小的，只有四个蛋白质成员，两个以前被注释的 RNA 甲基化基因，HSD17B10 和 KIAA0391，以及两个预测基因 POP1 和 POP4。功能分析表明，所有四个蛋白质都有助于 tRNA 的加工 (R-HSA-72306, $P=5.97E-09$)，其中三个参与了 tRNA 5' 端加工 (GO:0099116, $P=5.32E-08$)。HSD17B10 基因编码 3-羟基酰基-CoA 脱氢酶-2 型，参与线粒体脂肪酸 β -氧化。HSD17B10 参与 tRNA 的加工，因为它还与 TRMT10C/MRPP141 一起形成线粒体核糖核酸酶 P 的一个亚复合物。这个亚复合物被命名为 MRPP1-MRPP2，催化 tRNA 中第 9 位的 N1-甲基鸟嘌呤和 N1-甲基腺嘌呤的形成 (分别为 m1G9 和 m1A9)。KIAA0391，

POP1和POP2似乎也是核糖核酸酶P的组成部分，并通过5'端裂解对tRNA的成熟作出贡献。

假阳性的发现。我们的机器学习模型和分析提供了大量关于支撑人类RNA甲基化的推测基因网络的新信息。然而，值得注意的是我们方法的局限性。首先，不确定采用以前的功能注释知识是否会使模型预测产生偏差。我们在一定程度上解决了这个问题，因为我们使用了一个没有注释特征的缩小的特征集，如GO术语。在观察基于该数据集训练的模型的预测时，机器学习模型指出了本研究的一个反复出现的主题：RNA甲基化在功能上与一系列其他核心细胞功能相互关联（图3和补充图3）。例如，我们发现编码染色质修饰剂的基因是最主要的候选基因。这里的关键问题是这些基因是否代表假阳性，由GO术语的层次结构或共享的SAM结合域所刺激。这些模棱两可的预测应该得到进一步的验证，尽管多种证据表明这很可能是一个有生物学意义的结果，反映了DNA、RNA和转录后修饰过程之间的串联。

第二，没有琐碎的方法来控制假阳性。

因为到目前为止，只有少数作家的酶知道在RNA6上沉积甲基标记，所以我们从数量非常有限的阳性（以及随之而来的阴性）样本开始，用于机器学习。尽管基于测试数据的模型性能很好，但小的样本量可能阻碍了我们的模型的普及。为了更好地说明这一局限性，我们重新审视了我们在测试数据上的模型预测，这些数据总共包括18个阳性基因和5368个阴性基因。当考虑到前25个预测时（相当于上述所有预测中的前100个），假阳性率非常低，为

0.002（补充图6）。尽管如此，由于我们的测试数据中的总阳性数也很低，在这个概率窗口，假发现率，即假阳性（FP）除以假阳性（FP）和真阳性（TP）之和，是0.56。因此，我们的机器学习模型过度预测了与RNA甲基化途径相关的基因，其中只有一小部分人类基因在RNA甲基化中起作用。为了解决这一重要的注意事项，我们通过挖掘人类PPI数据来寻求独立的证据，以证实新预测的基因确实与RNA甲基化途径有关。

出于上述原因，并作为解释我们结果的指导，对于每个候选RNA甲基化基因，我们提供其在所有机器学习模型中的预测概率得分（无论是否使用功能注释数据得出），以及其在PPI网络分析中的PageRank得分（补充数据5）。一个在多个模型中具有持续高概率分数的基因，同时在人类PPI网络中具有较高的排名，就不太可能代表建模的假象。

结论

RNA甲基化是转录本稳定性、剪接和翻译效率的关键调控因素，在细胞的平衡和疾病中起着关键作用⁴。然而，其分子基础至今仍不为人所知¹¹。在此，我们旨在利用机器学习的方法获得对人类RNA甲基化途径相关基因的新认识。具体而言，我们分析了现有的转录组、蛋白质组、结构和基因组。

在有监督的机器学习框架下，蛋白质-蛋白质相互作用数据。

我们的机器学习模型在未见过的测试数据上显示出非常好的性能，达到了很高的准确率（91%）、精确度（90%）和召回率（92%）。先验的基因知识（如GO注释）与表达数据一起构成了预测模型中最有信息量的数据类型。值得注意的是，在某些组织中，如血液、心脏、胰腺和大脑，介导RNA甲基化的基因似乎显示了一个上升或下降的表达谱系。

使用独立的PPI数据，我们通过证实与以前已知的RNA甲基化基因的密切功能联系，正交验证了顶级模型的预测。社区检测划定了六个分子子网络，在tRNA加工（C1，C6）、rRNA加工（C3）、mRNA甲基化（C5）以及蛋白质（C2）和染色质修饰（C4）中具有不同的作用。网络分析表明，甲基标记在tRNA上的沉积与其他修饰过程，如2-硫醇化和假尿苷的形成，是共同协调的。同样，rRNA甲基转移酶似乎与参与rRNA加工和核糖体生物生成的几个基因有功能上的联系。有趣的是，RNA盖帽酶与染色质修饰剂聚集在一起，提出了这两个过程之间存在串扰的假设。我们的研究结果进一步表明，翻译因子的翻译后修饰和RNA上的表观转录组变化在调节翻译效率方面是相互交织的。总的来说，我们的研究体现了如何通过机器学习方法获取全息数据集来推断分子途径和新基因功能。

方法

数据集组装和预处理。为了组建一个机器学习数据集来预测人类基因组中参与RNA甲基化过程的基因，我们首先策划了一个先前已知的RNA甲基化基因的列表。为此，我们在标准的功能注释资源中进行了搜索，如ExPASy ENZYME (<https://enzyme.expasy.org/>)、InterPro (<https://www.ebi.ac.uk/interpro/>) 和GO资源 (<http://geneontology.org/>)，同时对注释的RNA甲基转移酶进行了全面的文献回顾，以跟进Schapira⁶的开创性论文。这使我们能够确定92个涉及到的蛋白质-或推测参与RNA甲基化，以用于机器学习建模（补充数据1）。

为了获得对基因功能进行分类的信息特征，我们查询了Harmonizome数据库¹⁵。Harmonizome提供了一个基因和蛋白质的预处理数据集的大集合，有来自70多个主要在线资源的约7200万个属性（功能关联）。特别是，数据最初被标准化为这样的连续值数据集，范围从0到1，或-1到1，其中1表示强烈的正基因-特征关联，0表示没有观察到基因-特征关联，-1表示强烈的负基因-特征关联（例如，基因表达数据集的下调）。然后，通过保留最强的10%的基因-特征关联，得出二元和三元数据集。我们从这四大类中选择了15个一键编码的数据集：(i) 转录组学；(ii) 蛋白质组学；(iii) 结构或功能注释；以及(iv) 物理相互作用（补充数据2）。特别是，从全能实验中，我们对BioGPS¹⁶、GTEx¹⁸、HPA¹⁹和TISSUES²⁰的基因和蛋白表达谱数据进行了采样。从功能数据集中，我们考虑了GO注释和InterPro结构域。最后，从物理相互作用数据集中，我们选择了KEGG和Reactome Pathways，以及Hub Proteins和Pathway Commons。整理这些数据产生了一个包含26935个基因和50176个单次编码特征的初始矩阵（“完整特征集”）。此外，我们通过排除所有5148个GO和InterPro注释的特征，编制了第二个降维的数据集（“降维特征集”）。

问题框架、模型定义、训练和评估。为了估计一个基因与RNA甲基化相关的概率，我们使用标准的机器学习方法进行二元分类。我们将92个事先已知的RNA甲基化基因标记为阳性样本（第1类），并将其分成两组，包括：(i) 80%的数据用于训练和交叉验证（ $n = 74$ ）；(ii) 20%的数据保持未见，用于模型测试（ $n = 18$ ）。我们认为人类基因组的其余基因是阴性样本（0类），并按类似的80/20比例分成训练/交叉验证（ $n = 21,476$ ）和测试集（ $n = 5368$ ）。这里的基本假设是，人类基因组中的绝大多数基因都有其他功能，因此训练数据中的假阴性基因数量应该非常少。

为了产生平衡的训练样本集，并在以后通过平均化减少我们最终模型的方差，为训练而保留的阴性基因 ($n = 21,476$) 被进一步分为74组，与训练的阳性样本数量相等。因此，我们产生了290个训练集，其中阳性类保持不变，阴性类由随机抽取基因的等量，每个基因采样一次。

从290个训练集和未经处理的Harmonizome数据，包括50,176个特征，我们接下来进行筛选，以去除低信息量的特征。我们删除了以下特征：(i)在每个训练组中超过70%的样本，或(ii)在至少一个训练组中小于16%的差异。然后将290个训练集中的每一个选定的特征合并成一个最终的列表，用于模型训练和测试。我们对缩小的特征集也采取了完全相同的选择过程。

我们接下来考虑了五种用于二进制的机器学习模型。Logistic回归 (LR)、高斯奈夫贝叶斯 (GNB)、支持向量机 (SVM)、随机森林 (RF) 和梯度提升 (GB) 模型。我们在每个训练集上使用网格搜索和3倍交叉验证来调整SVM的核函数 (线性或RBF)、成本参数和核带宽 (仅RBF核) 的超参数。对于RF，我们用网格搜索来确定森林中的最佳树数，然后用随机搜索来选择，这些参数包括分割节点时考虑的最大特征数、每个决策树的最大层数、在节点被分割前放在节点中的最小数据点，以及叶子节点中允许的最小数据点数。同样，对于GB模型，我们进行了网格搜索，以优化学习率和森林中的树木数量，并随后进行随机搜索，以调整其余的决策树参数 (见RF)。我们分别从完整的和缩小的特征集中的每一个训练集上训练了所有五种预测模型。所有分类器的性能采用10倍交叉验证，使用标准性能指标：准确度、精确度、召回率 (灵敏度)、F1得分和接收者操作特征曲线下的面积 (AUROC)。最后，我们使用GB特征排名来确定前100个信息量最大的特征，分别是完整的和缩小的特征集的训练集合。

在测试数据集和全基因组预测上的最终模型测试。一旦在交叉验证的基础上选择了最佳模型组合，我们就在未见过的数据上测试其性能。类似于上面描述的训练数据的程序，我们产生了298个测试数据集，将用于测试的阴性基因分成18个基因的，并将它们与预先保留的18个阳性样本相结合。在每个测试数据集上，使用准确度、精确度、召回率、F1得分和AUROC对分类器组合中的每个模型进行了评估。总体，通过对所有模型的测试结果进行平均计算。

同样，每个人类基因的预测概率也是通过，对合集的所有模型中第1类的概率分数进行平均计算。大多数非第1类基因 (除测试案例外的所有基因) 是训练数据中的阴性样本的一部分，但由于模型数量多 (290个)，这对最终预测的影响预计可以。

所有的可视化和元分析都是使用R软件环境 (4.0.5版)⁴²进行的。使用R软件包pheatmap生成了已知和预测的RNA甲基化基因的热图，包括用于机器学习的所有特征。使用clusterProfiler软件包对“生物过程”领域内的预测基因进行GO富集分析，对模型预测进行了进一步的验证⁴³。人类的蛋白质-蛋白质相互作用 (PPI) 数据来自STRING (v.11.0)²¹，并对综合得分在400分以上的相互作用进行了筛选。所有的网络分析都是使用igraph R软件包44进行的。PPI群落的功能注释是用EnrichR23进行的。

统计学和可重复性。每项分析中使用的样本量和统计参数在相关的方法和结果部分都有说明，适用时在图例中也有。所有的统计分析都在R (v4.0.5) 中进行。

报告摘要。关于研究设计的更多信息可在本文链接的《自然》研究报告摘要中找到。

数据可用性

本研究中使用的所有数据集都列在补充数据2中，并在Harmonizome数据库 (<https://maayanlab.cloud/Harmonizome/>) 中公开。

代码可用性

用于机器学习分析的Python脚本可在GitHub https://github.com/storm-therapeutics/ML_RNA_methylation。

收到的。10 January 2022; Accepted:2022年8月8日。

Published online: 25 August 2022

参考文献

- Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187-1200 (2017).
- Boccaletto, P. et al. MODOMICS: a database of RNA modification pathways.2017年更新。 *Nucleic Acids Res.* 46, D303-D307 (2018)。
- Barbieri, I. & Kouzarides, T. Role of RNA modifications in cancer. *Nat.Rev. Cancer* 20, 303-322 (2020).
- Huang, H., Weng, H., Deng, X. & Chen, J. RNA modifications in cancer:功能、机制和治疗意义。 *Annu.Rev. Cancer Biol.* 4, 221-240 (2020).
- Delatte, B. et al. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* 351, 282-285 (2016).
- Schapiro, M. Human RNA methyltransferases的结构化学。 *ACS Chem.Biol.* 11, 575-582 (2016).
- Tzelepis, K., Rausch, O. & Kouzarides, T. RNA-modifying enzymes and their function in a chromatin context. *Nat.Struct.Mol.Biol.* 26, 858-862 (2019).
- Copeland, R. A., Olhava, E. J. & Scott, M. P. 瞄准表观遗传酶进行药物开发。 *Curr.Opin.Chem.Biol.* 14, 505-510 (2010).
- Shi, H., Chai, P., Jia, R. & Fan, X. Novel insight into the regulatory roles of diverse RNA modifications:重新定义转录和翻译之间的桥梁。 *Mol.Cancer* 19, 78 (2020).
- 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)。
- 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).
- Jonkhout, N.等人，人类疾病中的RNA修饰景观。 *RNA* 23, 1754-1769 (2017).
- de Crécy-Lagard, V. et al.将人类的tRNA修饰与其已知和预测的酶相匹配。 *Nucleic Acids Res.* 47, 2143-2159 (2019).
- Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* 10, 35 (2017).
- Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *数据库* 2016, baw100 (2016)。
- Wu, C. et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* 10, R130 (2009).
- Gene Ontology Consortium.基因本体资源：丰富一个Gold矿。 *Nucleic Acids Res.* 49, D325-D334 (2021).
- Aguet, F. et al. Genetic effects on gene expression across human tissues. *自然》* 550, 204-213 (2017)。
- Uhlén, M.等人，基于组织的人类蛋白质组地图。 *Science* 347, 1260419 (2015).
- Palasca, O., Santos, A., Stolte, C., Gorodkin, J. & Jensen, L. J. TISSUES 2.0: an integrated web resource on mammalian tissue expression. *Database* 2018, bay003 (2018).
- Szklarczyk, D. et al. STRING v11: 蛋白质-蛋白质关联网络的覆盖率增加，支持全基因组实验数据集的功能发现。 *Nucleic Acids Res.* 47, D607-D613 (2019).
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat.Mech.* 2008, P10008 (2008).
- Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* 14, 128 (2013).
- Haag, S. et al. NSUN6是一种人类RNA甲基转移酶，可催化特定tRNA中m5C72的形成。 *RNA* 21, 1532-1543 (2015).
- Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat.Rev. Genet.* 15, 829-845 (2014).
- Ishizawa, T., Nozaki, Y., Ueda, T. & Takeuchi, N. The human mitochondrial translation release factor HMRFL1 is methylated in the GGQ motif by the methyltransferase HMPPrmC. *Biochem.Biophys.Res. Commun.* 373, 99-103 (2008).
- 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).
- Tischner, C. et al. MTO1通过tRNA修饰和翻译优化介导OXPHOS缺陷的组织特异性，这可以通过饮食干预绕过。 *Hum.Mol.Genet.* 24, 2247-2266 (2015).
- Gu, X. et al. SAMTOR是mTORC1途径的一个S-腺苷蛋氨酸传感器。 *Science* 358, 813-818 (2017).
- Sharma, S., Watzinger, P., Kötter, P. & Entian, K.-D.一种新型甲基转移酶 Bmt2的鉴定，它负责N-1-甲基腺苷碱的转移。

- 酵母菌中25S rRNA的修饰。 *Nucleic Acids Res.* 41, 5428-5443 (2013)。
31. Lim, S. L. et al. HENMT1和piRNA的稳定性是成年男性生殖细胞转座子抑制和确定小鼠精子生成程序所必需的。 *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. 双重甲基转移酶METTL13靶向eEF1A的N端和Lys55，并调控编码特定的翻译率。 *Nat. Commun.* 9, 1-15 (2018).
 34. Wu, X. et al. m6Acomet: 来自RNA共甲基化网络的单个m6A RNA甲基化位点的大规模功能预测。 *BMC Bioinform.* 20, 223 (2019).
 35. Meyer, B. et al. Bowen-Conradi综合征蛋白Nep1 (Emg1)在真核生物体生物生成中具有双重作用，既是一个基本的组装因子，又是酵母18S rRNA中Ψ1191的甲基化因子。 *Nucleic Acids Res.* 39, 1526-1537 (2011).
 36. Freed, E. F., Prieto, J.-L., McCann, K. L., McStay, B. & Baserga, S. J. NOL11，与北美印第安儿童肝硬化的发病机制有关，是Pre-rRNA转录和处理所需的。 *PLOS Genet.* 8, e1002892 (2012).
 37. Shima, H. & Igarashi, K. N1-methyladenosine (m1A) RNA modification: the key to ribosome control. *J. Biochem.(东京)* 167, 535-539 (2020)。
 38. Kan, R. L., Chen, J. & Sallam, T. 基因调控中表观转录组和表观遗传机制之间的串扰。 *Trends Genet.* 38, 182-193 (2021).
 39. Huang, H. et al. Histone H3 trimethylation at lysine 36 guide m6A RNA modification co-transcriptionally. *Nature* 567, 414-419 (2019).
 40. Chen, H. et al. METTL4是一种调节RNA剪接的snRNA m6Am甲基转移酶。 *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核心团队。 *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).

鸣谢

作者感谢 Adrián Rodríguez-Bazaga 对机器学习分析的宝贵意见，并感谢 Woochang Hwang 对网络分析的反馈。

作者贡献

G.T.设计并进行了分析，由H.W.和N.H.指导，H.W.和N.H.构思研究。H.S.S.、A.A.、D.L.、O.R.和T.K.参与了数据收集和解释。在H.W.、D.L.、N.H.的帮助下，G.T.撰写并修改了论文，其他作者也提供了意见。所有作者都阅读并批准了该论文。

竞争性利益

G.T.、D.L.、O.R.和H.W.是Storm Therapeutics公司的员工。N.H.是KURE.ai和CardiaTec Biosciences的联合创始人。T.K.是Abcam和Storm Therapeutics的共同创始人。

其他信息

补充信息 在线版本包含补充材料，可在<https://doi.org/10.1038/s42003-022-03821-y>。

通讯和资料索取请联系 Georgia Tsagkogeorga, Hendrik Weisser 或 Namshik Han。

同行评审信息 《通信生物学》感谢 Achraf El Allali、张超和陈坤奇对这项工作的同行评审做出的贡献。主要处理编辑。Gene Chong。

重印和许可信息可在<http://www.nature.com/reprints>。

出版商说明 《斯普林格-自然》对已出版的地图和机构联盟中的管辖权要求保持中立。



开放存取 本文采用知识共享署名4.0国际许可协议进行许可，允许使用、分享。

只要你适当注明原作者和来源，提供知识共享协议的链接，并说明是否做了修改，就可以在任何媒介或格式中改编、分发和复制。本文中的图片或其他第三方材料都包含在文章的知识共享协议中，除非在材料的信用行中另有说明。如果材料没有包括在文章的知识共享协议中，而你的预期用途不被法定条例所允许，或者超出了允许的用途，你需要直接从版权持有人那里获得许可。要查看该许可证的副本，请访问<http://creativecommons.org/licenses/by/4.0/>。

© 作者：2022年