基于机器学习框架QRIS研究影响逆转录病毒整合偏好性的因素

张厚煜^{1,2*}, 杨柳⁴, 徐进³, 颜光玟^{2,4*}

1. 华南理工大学生物科学与工程学院, 广州 510006;

2. 中国医学科学院&北京协和医学院血液病医院(血液学研究所), 实验血液学国家重点实验室, 天津 300020;

3. 华南理工大学医学院细胞发育系统生物学实验室, 广州 510006;

4. 南方医科大学基础医学院发育生物学教研室, 广州 510006

* 联系人, E-mail: hughiez047@gmail.com; kuangyuyuen@smu.edu.cn

收稿日期: 2021-10-06; 接受日期: 2021-10-16; 网络版发表日期: 2022-01-06

国家自然科学基金(批准号: 31522031, 31571526)资助

摘要 逆转录病毒的活动和多种疾病的发生密切相关, 其可通过整合自身基因组到宿主基因组中进行大量有害扩增. 与此同时, 宿主细胞可根据逆转录病毒的整合位置来启动相关通路抑制其活动. 因此整合位点的选择是决定逆转录病毒和宿主命运的关键步骤. 尽管国际上已有报道称逆转录病毒在宿主基因组上整合并不随机, 并发现了跨逆转录病毒种属的不同DNA基序偏好性, 但是这些不足以解释一些基序偏好较弱的逆转录病毒如何获得整合特异性. DNA结构特征, 如DNA形状被提出可以影响DNA结合蛋白与DNA结合的亲和力, 但DNA形状对不同逆转录病毒整合的影响机制尚不清楚. 本研究通过构建QRIS(quantify the retrovirus integration specificity)机器学习框架, 定量评估了DNA形状对逆转录病毒整合位点的影响. 通过系统地研究跨越四个种属的六种逆转录病毒, 本研究发现, DNA形状可以独立地或与DNA基序协同调控逆转录病毒的整合. 基于此, 可将六种逆转录病毒分为三类: “强兼性偏好型”“弱兼性偏好型”和“强形状偏好型”逆转录病毒. 此外本研究发现, 即使没有特定偏好的DNA基序, “强形状偏好型”逆转录病毒也可以通过DNA形状获得整合特异性. 本文基于机器学习对于大规模逆转录病毒插入位点特性的研究, 揭示了DNA结构特征在逆转录病毒致病位点选择中的重要作用. 该发现在临床方面为治疗逆转录病毒的策略提供了新靶点和新思路, 在应用方面有助于更精确地设计慢病毒载体用于基因编辑和基因治疗. 此外, 本文建立的机器学习框架也为其他种类病毒的整合特异性研究提供了新方法和新视角.

关键词 逆转录病毒, 整合偏好, DNA基序, DNA形状, 机器学习

逆转录病毒可以将其基因组整合到宿主基因组中并存储为原病毒(provirus)^[1], 其可扰乱宿主的正常基因组结构, 并转录其原病毒进行扩增来威胁宿主细胞.

同时, 宿主细胞根据逆转录病毒的整合位置发展出相应的机制来抑制原病毒转录, 从而消除危害^[1]. 因此, 整合位点的选择对于逆转录病毒的命运至关重要. 此

引用格式: 张厚煜, 杨柳, 徐进, 等. 基于机器学习框架QRIS研究影响逆转录病毒整合偏好性的因素. 中国科学: 生命科学, 2022, 52: 1540–1552
Zhang H Y, Yang L, Xu J, et al. QRIS: a machine learning framework to investigate the determinants of retroviral integration specificity (in Chinese). Sci Sin Vitae, 2022, 52: 1540–1552, doi: [10.1360/SSV-2021-0118](https://doi.org/10.1360/SSV-2021-0118)

前的研究表明, 逆转录病毒的整合位点选择不是随机的过程. 例如, 人类免疫缺陷病毒1型(human immunodeficiency virus 1, HIV-1)和人类T细胞白血病病毒1型(human T-lymphotropic virus 1, HTLV-1)倾向于整合在转录活跃区域^[2~4]; 鼠白血病病毒(murine leukemia virus, MLV)倾向于整合在转录起始位点附近^[2,4~6]; 而猴免疫缺陷病毒(simian immunodeficiency virus, SIV)则偏好于密集分布的基因区域^[7].

逆转录病毒的整合偏好性被认为和其偏好的DNA基序(DNA motif)有关. 例如, 多项研究表明, HIV-1偏好共有序列“TDGGTWACCHA”^[8~11]; MLV更偏好共有序列“DNSTVVTRBSAV”^[9,10]. 然而, 据报道HTLV-1和小鼠乳腺肿瘤病毒(mouse mammary tumor virus, MMTV)具有较弱的基序偏好性^[12,13], 但是为何其偏好特定的基因组区域尚不清楚.

除了DNA基序, 有研究表明, DNA结构相关特征可以影响逆转录病毒整合特异性, 例如DNA的弯曲度(curvature)及柔性(flexibility)可以显著影响逆转录病毒整合位点的选择^[14]. 最近一项生物物理研究表明, HIV-1更喜欢柔性DNA, 这进一步突出了DNA结构特征在逆转录病毒整合特异性中的作用^[15]. 鉴于不同的基因组区域具有本质上不同的DNA结构特征, 这可能会赋予全基因组逆转录病毒整合特异性^[16]. 然而目前国际上尚未有研究对DNA结构特征在逆转录病毒整合的影响进行系统和定量的评估, 主要因为: (i) 逆转录病毒整合位点数据较为有限; (ii) 没有高精度度的全基因组DNA形状(DNA shape)计算方法; (iii) 缺乏有效的生物信息学定量研究工具.

本研究旨在通过以下方法解决这些关键问题: (i) 在文献中搜索近期大规模测序的逆转录病毒整合位点; (ii) 利用最近发表的DNAShapeR软件^[17,18]计算全基因组水平上的14种DNA形状; (iii) 构建QRIS (quantify the retrovirus integration specificity)机器学习框架, 以定量衡量DNA形状和DNA基序对逆转录病毒整合的影响. 结合这些策略, 本文系统研究了DNA形状和基序对于跨越四个种属的以下六种逆转录病毒整合的影响: HIV-1, HTLV-1, MLV, SIV, 莫洛尼鼠白血病病毒(moloney murine leukemia virus, MoMLV)和MMTV. 通过定量研究DNA基序和DNA形状特征对不同种属逆转录病毒整合偏好性的影响, 期望为理解逆转录病毒如何特异性地整合到宿主基因组而引起致病

性提供新角度, 并为其他种类病毒的整合偏好性提供参考.

1 材料与方法

1.1 逆转录病毒整合位点数据信息

本研究使用的逆转录病毒整合位点均从发表文献的公开数据中下载所得, 其通过特定种类的逆转录病毒感染细胞系, 并通过高通量测序获得逆转录病毒的整合位点. 为了全面探究逆转录病毒的插入保守特性, 对于同一种逆转录病毒, 本研究搜寻其在不同细胞系中的整合数据进行比较, 具体数据集信息见表1^[19~24]. 随之将下载的原始数据通过liftover软件^[25]转换到统一的基因组装配版本, 其中人的基因组装配版本为hg38, 鼠为mm10, 斑马鱼为danRer11.

1.2 DNA基序对逆转录病毒整合的影响

获取每一种逆转录病毒的每一个数据集的整合位点后, 本研究使用bedtools软件^[26]拓展每个数据集中单个碱基(base pair, bp)整合位点到51 bp的范围, 随后在每个数据集中随机抽取10000个整合位点, 利用getfasta命令获取其DNA序列. 之后使用MEME软件^[27]在每个数据集中从头搜索DNA基序. MEME计算得到的位置权重矩阵(position weight matrix, PWM), 该矩阵使用51行存储了每种病毒偏好的基序模式, 以单碱基位置上每种核苷酸的概率量化用于下游分析.

1.3 DNA形状对逆转录病毒整合的影响

本研究使用Dickerson^[28]在1989年定义的DNA形状, 其中14种DNA形状可以使用DNAShapeR软件计算^[17]. 这些DNA形状可根据其衡量的指标分为三类: (i) 沿DNA骨架相邻碱基的空间取向: 移动(Shift), 滑动(Slide), 升起(Rise), 倾斜(Tilt), 滚动(Roll), 扭转(Helix Twist, HelT); (ii) 碱基配对的空间取向: 剪切(Shear), 拉伸(Stretch), 错开(Stagger), 弯曲(Buckle), 螺旋(Propeller Twist, ProT), 打开(Opening); (iii) 静电势(electrostatic potential, EP)和DNA小沟宽度(minor groove width, MGW). DNAShapeR通过蒙特卡洛模拟方法计算每种核苷酸五聚体的DNA形状, 其计算结果已被验证与DNA晶体结构的实验数据高度一致^[17]. 给定DNA序列, 其可以利用核苷酸五聚体的窗口计算每

表 1 本研究使用的逆转录病毒插入位点数据集信息

Table 1 Information of retroviruses integration dataset researched in this study

逆转录病毒类型	种属	感染的宿主细胞	整合位点数量	数据来源
HIV-1	lentivirus	Human HEK293T	943009	[19]
HIV-1	lentivirus	Human T cell	12430	[20]
HIV-1	lentivirus	Human PBMC	384382	[21]
HTLV-1	δ -retrovirus	Human Hematopoietic	12196	[22]
HTLV-1	δ -retrovirus	Human T cell	26849	[23]
MLV	γ -retrovirus	Human HepG2	2739814	[6]
MLV	γ -retrovirus	Human K562	233099	[6]
SIV	lentivirus	Human PBMC	159168	[21]
MoMLV	γ -retrovirus	Zebrafish Blastula	14922	[24]
MMTV	β -retrovirus	Mouse Mammary	176942	[13]

一个碱基位点的14种DNA形状值。基于此,对于每一种逆转录病毒的每一个数据集,本研究使用getShape()函数计算了1.2节中选取的整合位点侧翼51 bp的范围内的DNA形状值。此外,本研究随机抽取每个数据集对应物种基因组上相同数量的位点作为负对照,并以相同的方法计算其附近DNA形状值。

1.4 基于QRIS机器学习框架定量评估DNA形状和基序的影响

本文通过构建机器学习框架QRIS来定量评估DNA形状和DNA基序对逆转录病毒整合偏好的影响。在此框架中,第一步是使用glmnet软件的弹性网逻辑回归模型^[29]对逆转录病毒整合位点和随机基因组位点进行逻辑回归分类。在此步骤中,可以根据需求设计不同的输入向量用于模型训练。本研究采用三种类型的输入向量:(i)基序向量:仅计算逆转录病毒整合位点周围51 bp每个核苷酸的PWM得分;(ii)DNA形状向量:使用DNASHapeR的encodeSeqShape()函数,编码逆转录病毒整合位点侧翼51 bp的4种DNA形状值;(iii)基序+DNA形状向量:基序向量和4种或者14种DNA形状向量的组合。此训练的基本假设是,如果输入向量可以获得更好的模型精度来区分整合位点和随机位点,那么此输入向量对于逆转录病毒的影响更大。

单碱基精度基序的计算:对于每一种逆转录病毒的偏好PWM,该矩阵描述了每个碱基位点上出现每一种核苷酸的概率。因此对于每一个整合位点侧翼51 bp的序列,可使用遍历的方法计算每一个整合位点附近每一个碱基出现当前核苷酸的概率。单碱基精度形状

的计算:对于每一个整合位点侧翼51 bp的序列,使用DNASHapeR以5 bp的窗口计算每一个碱基的14种DNA形状值。为了最大程度地减少对机器学习算法在特定数据集集中的过拟合,本研究使用10折交叉验证方法来训练模型。

第二步为提取模型的标准化系数作为代理,来衡量逆转录病毒整合位点侧翼单碱基水平的DNA形状和基序重要性。本研究为了全面综合评估DNA基序和每一种DNA形状对于逆转录病毒的整合影响,选取了基序+14形状模型作为进一步评估的模型。对于整合位点侧翼51 bp的范围,共有765个预测变量(15种因素)输入模型。因为15种因素具有不同的计量单位,QRIS在训练模型之前使用最大最小标准化算法对这些预测变量进行归一化,因此该步骤返回的每一个预测变量的系数可以用于直接比较每一种预测变量对于该种逆转录病毒的单碱基精度的影响。随之,QRIS使用ComplexHeatmap^[30]绘制热图以直观地可视化单碱基精度DNA基序和14种DNA形状对于逆转录病毒的影响。

QRIS框架以及本研究所有源代码可在Github数据库中获取: <https://github.com/YenLab/QRIS>。

2 结果

2.1 逆转录病毒整合的DNA基序偏好性

逆转录病毒的整合酶可通过与特定的核苷酸排列相互作用而获得整合特异性,该排列被称为DNA基序^[31]。为了系统地探究逆转录病毒的基序偏好性,本研究搜集了来自6种逆转录病毒的10个数据集(表1)。

这些数据集均包含至少10000个整合位点, 是研究其整合特性的宝贵资源。

为了在不同逆转录病毒之间公平地比较其DNA基序偏好性, 本研究在每个数据集中随机选择了10000

个整合位点, 并使用MEME^[27]在每个数据集整合位点侧翼的51 bp窗口内搜索了对应的DNA基序(图1)。结果表明, 在三种不同宿主细胞类型中, HIV-1整合位点周围具有轻微不同的核苷酸保守性。这些基序与先前

图 1 逆转录病毒DNA基序的整合偏好性。逆转录病毒整合位点侧翼51 bp窗口内的基序。Bits衡量整合位点周围的序列保守性, 核心基序保守区域为灰色阴影部分

Figure 1 Retroviruses integration preference for DNA motif. Motifs within a 51 bp window around retroviruses integration sites. Bits measure the sequence conservation surrounding integration sites. The core conservative motif region was shaded grey

描述的共有序列“TDGGTWACCHA”^[9]一致, 说明HIV-1具有特定的DNA基序偏好性(图1)。在HepG2和K562细胞系中的MLV整合位点附近显示出一个强而对称的基序模式, 该富含AT核苷酸的基序与先前的报道一致^[9,10]。相比之下, SIV偏好于富含CG的基序模式。HTLV-1整合位点周围的核苷酸组成在造血细胞和T细胞中都更加平缓, 共同表明了其较弱的基序偏好。尽管MoMLV对富含AT的基序有着相似的偏好, 但MoMLV插入位点附近的核苷酸分布与MLV明显不同, 这表明粗略地使用AT或CG的含量来描述逆转录病毒的整合偏好性是不全面的。有趣的是, 本研究观察到MMTV几乎没有基序偏好性^[13,32](图1)。

值得注意的是, 逆转录病毒整合位点的基序偏好范围在不同病毒间有明显的差异(图1), 其中, HIV-1的基序偏好限制在15 bp的区域内, 而MoMLV的基序偏好限制区域为11 bp, MLV为19 bp。这提示不同种属病

毒的整合酶与宿主DNA结合过程中不同的机制^[1]。

2.2 DNA形状不同程度地影响逆转录病毒的整合偏好性

DNA形状描述了沿DNA骨架相邻碱基或碱基配对之间的空间取向。其已被广泛报道可以通过与DNA的间接相互作用影响DNA结合蛋白的亲和力^[33,34]。但DNA形状是否会影响逆转录病毒整合尚未有系统的研究。

本研究首先定性地测量了4种DNA形状——DNA小沟宽度, 扭转, 螺旋和滚动, 并发现DNA形状在逆转录病毒间有不同程度的影响(图2)。具体而言, 在三种不同宿主细胞类型中, HIV-1在中心15 bp区域更偏好宽的MGW, 较小的HelT, 较大的ProT和较大的Roll, 此范围与其偏好的基序长度相似(图2)。但是其中心基序侧翼内的DNA形状相比于随机位点有明显波动, 表明

图 2 DNA形状对逆转录病毒整合的影响。选择10000个人、小鼠和斑马鱼基因组随机位点作为负对照并用虚线表示。每种DNA形状的单位是MGW(埃), HelT(度), ProT(度), Roll(度)。使用Wilcoxon-Mann检验并以 P 值 <0.01 为阈值来检验整合位点和随机位点之间的DNA形状图谱的显著性。为了示意简洁, 仅使用相应的颜色标记了不显著的值

Figure 2 DNA shape effect on retroviruses integration. Exactly 10,000 randomly selected genomic sites for human, mouse, and zebrafish were selected respectively as negative controls and shown by dashed grey lines. The unit of each DNA shape is MGW (angstrom), HelT (degree), ProT (degree), Roll (degree). The significance of the DNA shape profile between integration sites and random sites was determined with P -value <0.01 as a threshold (Wilcoxon-Mann test). For simplicity, only not significant values were labeled using the corresponding color

DNA形状可以在更大范围内影响逆转录病毒的整合。值得注意的是, 更大的Roll与增强的DNA柔韧性相关^[33], 这与先前关于HIV-1偏好整合在柔性DNA的报道相一致^[14,15]。此外, 具有强烈基序偏好性的MLV也受到DNA形状的强烈影响, 如在HepG2和K562细胞系中其整合位点周围的DNA形状图谱所示(图2)。此外, 对于HTLV-1, 尽管其基序偏好较弱, 但其在造血细胞和T细胞中的整合位点周围的MGW, HelT和ProT图谱与基线明显不同, 这共同表明DNA形状影响了其整合过程(图2)。研究表明, 尽管MMTV没有基序偏好, 但其在研究的6种逆转录病毒中具有最强的DNA形状偏好性。其整合位点附近的DNA形状图谱与其他逆转录病毒明显相反: MMTV偏好较窄的MGW, 较大的HelT, 较小的ProT和较小的Roll, 这表明DNA形状在调节不同的逆转录病毒整合过程中甚至可以起到相反的作用。

基于DNA基序和形状的调控, 本研究对逆转录病毒整合偏好性进行分类: (i) “强兼性偏好型”逆转录病毒, 如HIV-1, MLV和SIV, 此类病毒在整合过程中同时受DNA基序和形状的强烈影响; (ii) “弱兼性偏好型”逆转录病毒, 如HTLV-1和MoMLV, 此类病毒在整合过程中同时受DNA基序和形状的较弱影响; (iii) “强形状偏好型”逆转录病毒, 如MMTV, 此类病毒在逆转录过程中展现出强烈的DNA形状偏好性, 而几乎没有DNA基序偏好。

2.3 基于QRIS框架评估DNA基序和形状对逆转录病毒整合的影响

为了进一步定量评估DNA基序和DNA形状对逆转录病毒的整合偏好性的影响, 本文构建了QRIS机器学习框架来统一定量地评估其作用(图3)。QRIS框架的第一步为使用弹性网逻辑回归算法(elastic-net logistic regression)建立预测模型^[29]。简而言之, 在每个数据集中随机选择10000个逆转录病毒整合位点和10000个对应的基因组位点, 之后将每个位点51 bp窗口内的基序和DNA形状值组合成四种类型的向量: 基序向量, 4形状向量, 基序+4形状向量, 基序+14形状向量。随之针对每一种向量训练特定的模型, 逐步探究基序和形状对于此逆转录病毒整合的定量作用。每种向量中编码的信息被用来对随机基因组位点和逆转录病毒整合位点进行逻辑回归, 并使用精度(accuracy)指标来评估每种模型的性能(图3)。

结果表明, 当仅使用DNA基序向量进行逻辑回归时(基序模型), HIV-1, MLV和SIV的模型精度更高, 这与它们具备更强的DNA基序偏好的结论相一致(图4A; 网络版附表1)。相反, HTLV-1和MoMLV的模型精度较低, 这与它们具备弱基序偏好性的结论一致。MMTV的模型精度最低, 这是由于它几乎没有基序偏好性。

为了定量测试DNA形状对逆转录病毒整合的影响, 本研究首先使用图2中研究的4种DNA形状向量进行逻辑回归(4形状模型)。结果表明, 除SIV外, 其余逆转录病毒的“4形状模型”精度均大于“基序模型”(图4A; 网络版附表1)。对于较高SIV“基序模型”精度的一种可能性解释是在回归模型中, 因为随机抽取的位点来源于GC含量较低的人类基因组, 因此SIV富含GC的基序可以帮助其更好地区分整合位点和随机位点。此外, 较高的“4形状模型”精度表明, 这四种DNA形状共同编码了有助于逆转录病毒整合偏好的信息, 而这些信息并未完全被基序描述。为了验证这一猜想, 本研究使用基序+4形状向量进行逻辑回归(基序+4形状模型)(图4A; 网络版附表1)。结果表明, 在所有数据集中“基序+4形状模型”与“基序模型”和“4形状模型”相比, 精度都有明显的提高, 这提示着DNA形状和DNA基序编码互补的信息, 以协同调节逆转录病毒的整合特异性。此外, 当将所有14种DNA形状均输入模型时(基序+14形状模型), 模型精度达到最高。作为负对照, 本研究另外打乱整合位点序列并训练了“基序+14形状(shuffled seqs)模型”(图4A, 灰色点), 结果表明负对照的精度较低, 提示该框架评估的DNA基序和形状对逆转录病毒整合的影响具有实际生物学意义。

由于单个精度指标可能会对模型性能的评估产生偏差, 本研究计算了机器学习中常用的5个独立指标, 包括F分数(F-score), 查全率(Recall), 查准率(Precision), 受试者工作特征曲线下面积(auROC)和P-R曲线下面积(auPRC)。结果表明, 这些指标在训练不同向量时与精度指标有着相似的趋势。这进一步验证了本研究的结论: DNA形状对逆转录病毒的整合有影响(图4B; 网络版附表1)。

2.4 单碱基精度上衡量DNA基序和DNA形状在逆转录病毒整合位点附近的影响

QRIS框架的第二步中通过使用第一步的定量模型, 进一步解析逆转录病毒整合位点附近单碱基水平

步骤 1: 训练逻辑回归模型

Figure 3 Digram of QRIS framework. The first step is classifying retroviruses integration sites from random genomic sites using an Elastic-net logistic regression model. The second step is extracting the standardized coefficient from the model as a proxy to measure the single base DNA shape and DNA motif effect around retrovirus integration sites

评估DNA基序和每种DNA形状对于此逆转录病毒整合偏好的相对重要性.

分析结果表明, 对于HIV-1, MLV和SIV在内的“强兼性偏好型”逆转录病毒, 基序重要性排名第一(图5), 表明DNA基序在调控该类逆转录病毒的整合特异性时比所有14种DNA形状更为重要。但是它们偏好的DNA形状有所不同, 在三种宿主细胞类型中, HIV-1主

图 4 DNA基序模型性能和形状对逆转录病毒整合的影响的评估。A: 基于精度的模型性能评估。使用图3中所示的四种向量分别对模型进行训练。此外使用“基序+14形状(shuffled seqs)”向量作为负对照进行了训练, 以排除机器学习算法中内在的偏差。B: 评估模型的独立指标。训练过程与(A)相同, 但是使用auROC, auPRC, Precision, Recall和F分数来测量模型性能。左上方的面板中显示了每个雷达图中的图标

Figure 4 Evaluation of the model performance of DNA motif and shape effect on retroviruses integration. A: Accuracy (ACC) based model evaluation. Models were trained separately using four types of vectors illustrated in Figure 3. B: Independent metrics for evaluating the model. The training process is the same as (A), but using auROC, auPRC, Precision, Recall and F-score to measure the model performance. Labels in each radar plot are illustrated in the top-left panel

要受Rise, MGW, Shear和Shift影响; MLV主要受到Roll, Shift, Rise和EP的影响; SIV主要受到Shear, Shift, EP和Opening的影响。这些结果表明, 不同种类DNA形状可以特异性地影响不同种属逆转录病毒的整合。此外, DNA基序和DNA形状的重要性在整合位点附近的相对位置上有所不同, 突出了它们在逆转录病毒整合过程中的协同作用。

对于HTLV-1和MoMLV之类的“弱兼性偏好型”逆转录病毒, 结果表明基序在MoMLV整合位点附近排名第一, 而EP在HTLV-1整合中占主导地位(图5)。对于排名前5的DNA形状, HTLV-1特异性地受到EP, MGW影

响, 而MoMLV特异地受Tilt, Rise和Shear影响。这表明即使它们的DNA形状偏好性相比“强兼性偏好型”逆转录病毒而言较弱, DNA形状仍可以在该类逆转录病毒的整合中发挥重要而多样的作用, 甚至在某些种属中超过DNA基序的重要性。

有趣的是, 本研究发现Opening形状对“强形状偏好型”逆转录病毒MMTV的整合影响最大(图5)。另外Opening的系数总体呈负值, 提示MMTV偏好刚性DNA区域^[33], 这与HIV-1偏好柔性DNA形成了鲜明对比。此外, 总体上呈负值的MGW系数表明MMTV偏好较窄的MGW, 这与本研究先前的结论一致(图2)。值得

图 5 逆转录病毒整合位点附近的DNA基序和形状的相对重要性。DNA基序和形状的重要性由基序+14形状模型返回的系数所表示。热图中的红色表示在逆转录病毒整合中发挥了促进作用, 蓝色表示逆转录病毒整合中的抑制作用。元素重要性由每行的绝对值加和算得并在左侧显示为圆圈尺寸。位置重要性由每列的绝对值加和算得并在底部显示为绿色的热图。为了示意简洁, 仅显示了总共15个元素中的前8个元素

Figure 5 Relative importance of DNA motif and shape surrounding retroviruses integration sites. The contributions of these features are represented by the coefficients returned by the motif+14 shapes model. The red color in the heatmap indicates a positive role in retroviruses integration; the blue color indicates a negative role in retroviruses integration. Element importance was calculated by summing the absolute value across each row and shown as circle size in the left. Position importance was calculated by summing the absolute value across each column and shown as a green heatmap bar at the bottom

注意的是, DNA形状和基序对于MMTV的影响跨越了51 bp范围, 表明其整合过程相对于其他逆转录病毒更为复杂。

总体来说, 本研究的定性和定量分析揭示了逆转录病毒的整合偏好性可以通过不同的机制赋予:

(i) “强兼性偏好型”逆转录病毒受DNA基序的影响较大; (ii) “弱兼性偏好型”逆转录病毒, 根据种属而异可以主要受到DNA形状或DNA基序的调控; (iii) “强形状偏好型”逆转录病毒特异性地受到DNA形状影响。

3 讨论

本文通过构建机器学习框架QRIS对大规模逆转录病毒整合位点进行训练,在单碱基精度上定量评估了DNA基序和14种DNA形状对于其整合的影响。相比于常见的大范围病毒整合机制的研究策略,本文基于机器学习的单碱基精度的定量研究可为以后的精准病毒致病机制理解提供更精细的视角。例如,人类基因组中存在大量的单核苷酸多样性(single nucleotide polymorphism, SNP),且被报道影响转录因子等调控蛋白的结合特异性^[35,36]。基于单碱基精度的DNA基序和形状的研究,可以帮助后续理解SNP是否会影响不同人群中病毒整合和感染的异质性,以为精准病毒学的研究提供有力的工具。此外,因为进化的分歧导致不同的逆转录病毒有差异的整合酶结构^[1],因此在整合过程中某些氨基酸会对于特定DNA基序模式有特异性,某些蛋白质结构可能会对特定的DNA形状产生亲和力。在缺乏结构生物学的晶体结构支持下,本文单碱基精度的基序和形状影响评估可以帮助研究者推测不同病毒的整合酶功能相似性;在有晶体结构的支持下,结合单碱基精度的基序和形状影响评估可以帮助研究者更好地理解病毒获得整合特异性的功能位点和整合机制。

本文所研究的HIV-1和SIV均来自慢病毒属(lentivirus genus),其已被广泛改造为慢病毒载体用于基因编辑和基因治疗^[37-39]。慢病毒载体的使用扩大了基因编辑的领域,并被最近启动的体细胞基因组编辑联盟(Somatic Cell Gene Editing Consortium, SCGE)认为是有广阔应用前景的递送系统^[40]。在基因治疗方面,使用慢病毒载体最成功的案例是嵌合抗原受体T细胞(chimeric antigen receptor T cells, CAR-T)的癌症治

疗^[37]。这些慢病毒应用中的关键步骤是通过慢病毒载体将外部基因传递到特定的宿主基因组位点,但慢病毒整合常常发生在错误的位置,并可能导致基因突变,甚至进一步发展为癌症^[41,42]。本研究表明,HIV-1和SIV都受DNA基序和形状的强烈调控,由此推测此调控模式适用于慢病毒种属。本研究提示,在未来设计慢病毒载体时应综合考虑包括DNA形状在内的影响因素,以更安全有效地用于基因编辑和基因疗法^[43,44]。

最近有报道称,COVID-19可以使用HIV逆转录酶逆转录其RNA,并将该cDNA整合到人类宿主基因组中^[45],但目前尚不清楚哪个整合酶负责其整合过程。对于COVID-19整合位点附近DNA结构和序列特性的进一步研究有助于推测行使功能的整合酶,并协助设计干预策略以避免复发性感染。

本研究从DNA层面上探究了DNA基序和DNA形状对逆转录病毒整合的影响,有报道显示宿主基因组的表观修改状态^[46]和调控因子(如LEDGF/p75^[20], ZAP^[47])可以影响逆转录病毒的整合及转录,这表明逆转录病毒的活动周期受到多层级的精确调控。除逆转录病毒外,其他种类的病毒,如乙型肝炎病毒(hepatitis B virus, HBV),人类乳头瘤病毒(Human papillomavirus, HPV)也被报道偏好性地整合到基因组特定区域^[48,49]。虽然本文将所研究的六种逆转录病毒分为三大类,即“强兼性偏好型”“弱兼性偏好型”和“强形状偏好型”,但是否所有种类的病毒均遵从类似的偏好性规则或是否存在“强基序偏好型”病毒,仍需后续对更多种类病毒的研究。基于此,本研究所使用的机器学习框架为未来综合研究病毒的整合机制提供了新策略,这不仅可以增进对病毒相关疾病的理解,还有助于更准确地预测病毒的病原性整合热点以进行精准干预。

参考文献

- 1 Lesbats P, Engelman A N, Cherepanov P. Retroviral DNA integration. *Chem Rev*, 2016, 116: 12730–12757
- 2 Mitchell R S, Beitzel B F, Schroder A R W, et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol*, 2004, 2: e234
- 3 Schröder A R W, Shinn P, Chen H, et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, 2002, 110: 521–529
- 4 Kim S, Kim N, Dong B, et al. Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer. *J Virol*, 2008, 82: 9964–9977
- 5 Wu X, Li Y, Crise B, et al. Transcription start regions in the human genome are favored targets for MLV integration. *Science*, 2003, 300: 1749–1751

- 6 LaFave M C, Varshney G K, Gildea D E, et al. MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Res*, 2014, 42: 4257–4269
- 7 Hematti P, Hong B K, Ferguson C, et al. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol*, 2004, 2: e423
- 8 Wang G P, Ciuffi A, Leipzig J, et al. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res*, 2007, 17: 1186–1194
- 9 Holman A G, Coffin J M. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc Natl Acad Sci USA*, 2005, 102: 6103–6107
- 10 Berry C, Hannehalli S, Leipzig J, et al. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol*, 2006, 2: e157
- 11 Serrao E, Krishnan L, Shun M C, et al. Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. *Nucleic Acids Res*, 2014, 42: 5164–5176
- 12 Derse D, Crise B, Li Y, et al. Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. *J Virol*, 2007, 81: 6731–6741
- 13 de Jong J, Akhtar W, Badhai J, et al. Chromatin landscapes of retroviral and transposon integration profiles. *PLoS Genet*, 2014, 10: e1004250
- 14 Pruss D, Reeves R, Bushman F D, et al. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J Biol Chem*, 1994, 269: 25031–25041
- 15 Michieletto D, Lusic M, Marenduzzo D, et al. Physical principles of retroviral integration in the human genome. *Nat Commun*, 2019, 10: 575
- 16 Pataskar A, Vanderlinden W, Emmerig J, et al. Deciphering the gene regulatory landscape encoded in DNA biophysical features. *iScience*, 2019, 21: 638–649
- 17 Chiu T P, Comoglio F, Zhou T, et al. DNASHapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 2016, 32: 1211–1213
- 18 Li J, Sagendorf J M, Chiu T P, et al. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res*, 2017, 45: 12877–12887
- 19 Singh P K, Plumb M R, Ferris A L, et al. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev*, 2015, 29: 2287–2297
- 20 Sherrill-Mix S, Lewinski M K, Famiglietti M, et al. HIV latency and integration site placement in five cell-based models. *Retrovirology*, 2013, 10: 90
- 21 Ferris A L, Wells D W, Guo S, et al. Clonal expansion of SIV-infected cells in macaques on antiretroviral therapy is similar to that of HIV-infected cells in humans. *PLoS Pathog*, 2019, 15: e1007869
- 22 Furuta R, Yasunaga J I, Miura M, et al. Human T-cell leukemia virus type 1 infects multiple lineage hematopoietic cells *in vivo*. *PLoS Pathog*, 2017, 13: e1006722
- 23 Artesi M, Marçais A, Durkin K, et al. Monitoring molecular response in adult T-cell leukemia by high-throughput sequencing analysis of HTLV-1 clonality. *Leukemia*, 2017, 31: 2532–2535
- 24 Varshney G K, Lu J, Gildea D E, et al. A large-scale zebrafish gene knockout resource for the genome-wide study of gene function. *Genome Res*, 2013, 23: 727–735
- 25 Lee C M, Barber G P, Casper J, et al. UCSC Genome Browser enters 20th year. *Nucleic Acids Res*, 2020, 48: D756–D761
- 26 Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010, 26: 841–842
- 27 Bailey T L, Johnson J, Grant C E, et al. The MEME suite. *Nucleic Acids Res*, 2015, 43: W39–W49
- 28 Dickerson R E. Definitions and nomenclature of nucleic acid structure components. *Nucl Acids Res*, 1989, 17: 1797–1803
- 29 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Soft*, 2010, 33: 1–22
- 30 Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 2016, 32: 2847–2849
- 31 Stormo G D. DNA binding sites: representation and discovery. *Bioinformatics*, 2000, 16: 16–23
- 32 Serrao E, Ballandras-Colas A, Cherepanov P, et al. Key determinants of target DNA recognition by retroviral intasomes. *Retrovirology*, 2015, 12: 39

- 33 Rohs R, Jin X, West S M, et al. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*, 2010, 79: 233–269
- 34 Abe N, Dror I, Yang L, et al. Deconvolving the recognition of DNA shape from sequence. *Cell*, 2015, 161: 307–318
- 35 Halow J M, Byron R, Hogan M S, et al. Tissue context determines the penetrance of regulatory DNA variation. *Nat Commun*, 2021, 12: 2850
- 36 Yan J, Qiu Y, Ribeiro Dos Santos A M, et al. Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, 2021, 591: 147–151
- 37 Naldini L. Gene therapy returns to centre stage. *Nature*, 2015, 526: 351–360
- 38 Lu J, Chen F, Xu Z, et al. Membrane immunoglobulin expressed by retroviral vector gene transfer mimics partial function of the B-cell receptor *in vivo*. *Sci China Life Sci*, 2016, 59: 49–58
- 39 Zhang J P, Cheng T, Zhang X B. Hematopoietic stem cell gene therapy: progress and challenges (in Chinese). *Sci Sin Vitae*, 2017, 47: 1323–1335 [张健萍, 程涛, 张孝兵. 造血干细胞基因治疗: 进展与挑战. 中国科学: 生命科学, 2017, 47: 1323–1335]
- 40 Saha K, Sontheimer E J, Brooks P J, et al. The NIH somatic cell genome editing program. *Nature*, 2021, 592: 195–204
- 41 Vranckx L S, Demeulemeester J, Debyser Z, et al. Towards a safer, more randomized lentiviral vector integration profile exploring artificial LEDGF chimeras. *PLoS ONE*, 2016, 11: e0164167
- 42 Woods N B, Bottero V, Schmidt M, et al. Therapeutic gene causing lymphoma. *Nature*, 2006, 440: 1123
- 43 Papayannakos C, Daniel R. Understanding lentiviral vector chromatin targeting: working to reduce insertional mutagenic potential for gene therapy. *Gene Ther*, 2013, 20: 581–588
- 44 He Z Y, Men K, Qin Z, et al. Non-viral and viral delivery systems for CRISPR-Cas9 technology in the biomedical field. *Sci China Life Sci*, 2017, 60: 458–467
- 45 Zhang L, Richards A, Barrasa M I, et al. Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues. *Proc Natl Acad Sci USA*, 2021, 118: e2105968118
- 46 Vansant G, Chen H C, Zorita E, et al. The chromatin landscape at the HIV-1 provirus integration site determines viral expression. *Nucleic Acids Res*, 2020, 48: 7801–7817
- 47 Zhu Y, Chen G, Lv F, et al. Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. *Proc Natl Acad Sci USA*, 2011, 108: 15834–15839
- 48 Zhao L H, Liu X, Yan H X, et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nat Commun*, 2016, 7: 12992
- 49 Hu Z, Zhu D, Wang W, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet*, 2015, 47: 158–163

QRIS: a machine learning framework to investigate the determinants of retroviral integration specificity

ZHANG HouYu^{1,2}, YANG Liu⁴, XU Jin³ & YEN KuangYu^{2,4}

1 School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China;

2 State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Institute of Hematology & Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin 300020, China;

3 Division of Cell, Developmental and Integrative Biology, School of Medicine, South China University of Technology, Guangzhou 510006, China;

4 Department of Developmental Biology, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510006, China

Retroviral activity is linked with various diseases. Retrovirus can integrate a copy of its genomic DNA into the host genome and store it as the provirus. Thus, it threatens the host cell by interrupting the host genome architecture and transcribing its provirus for detrimental expansion. However, depending on where the retrovirus integrates, the host has corresponding defense mechanisms to repress the transcription of the provirus and eliminate harm. Therefore, the integration site selection is a vital process for retrovirus' fate. Many efforts have been made to understand its integration specificity and find diverse DNA motif preferences across retrovirus genera. However, the effect of genome-wide DNA structural properties, such as DNA shapes, on retroviral integration is unclear. Therefore, we systematically investigated this issue on six retroviruses that are representative of four genera. We devised QRIS (quantify the retrovirus integration specificity), a machine learning framework to assess the effect of DNA shape on large-scale retroviral integration sites. We observed that the DNA shape could independently or cooperatively work with the DNA motif to regulate retroviral integration. Based on this, we classified these retroviruses into three categories: StrongFavor retrovirus, WeakFavor retrovirus, and Strongshape retrovirus. Interestingly, the Strongshape retrovirus can gain specificity through DNA shapes without a particular DNA motif. Our qualitative and quantitative evaluation of DNA shape and motif revealed their diverse roles in regulating the retroviral integration specificity. Our findings may help control the lentivirus vector more precisely for gene therapy and disturb retroviral integration during pathogenesis.

retrovirus, integration preference, DNA motif, DNA shape, machine learning

doi: [10.1360/SSV-2021-0118](https://doi.org/10.1360/SSV-2021-0118)