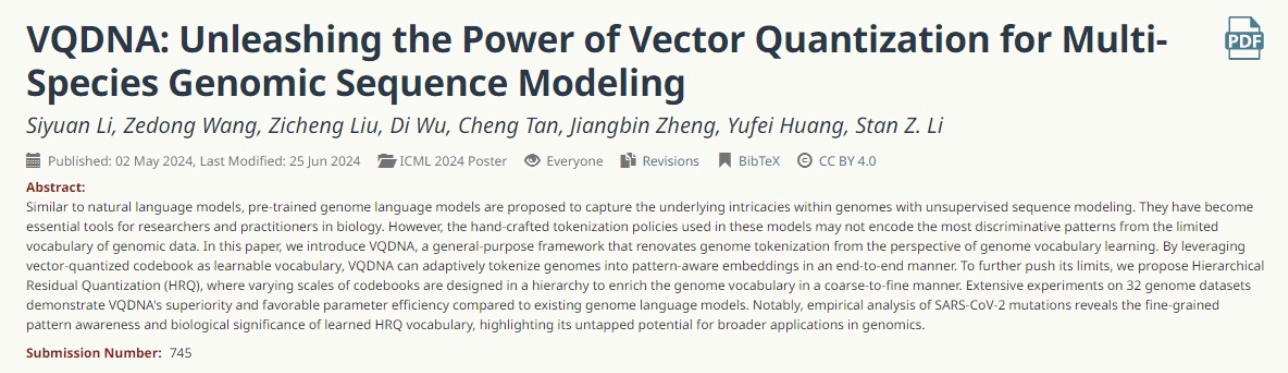
VQDNA: Unleashing the Power of Vector Quantization for Multi-Species Genomic Sequence Modeling



VQDNA: Unleashing the Power of Vector Quantization for Multi-Species Genomic Sequence Modeling

Siyuan Li, Zedong Wang, Zicheng Liu, Di Wu, Cheng Tan, Jiangbin Zheng, Yufei Huang, Stan Z. Li

Published: 02 May 2024, Last Modified: 25 Jun 2024 ICML 2024 Poster Everyone Revisions BibTeX CC BY 4.0

Abstract:
Similar to natural language models, pre-trained genome language models are proposed to capture the underlying intricacies within genomes with unsupervised sequence modeling. They have become essential tools for researchers and practitioners in biology. However, the hand-crafted tokenization policies used in these models may not encode the most discriminative patterns from the limited vocabulary of genomic data. In this paper, we introduce VQDNA, a general-purpose framework that renovates genome tokenization from the perspective of genome vocabulary learning. By leveraging vector-quantized codebook as learnable vocabulary, VQDNA can adaptively tokenize genomes into pattern-aware embeddings in an end-to-end manner. To further push its limits, we propose Hierarchical Residual Quantization (HRQ), where varying scales of codebooks are designed in a hierarchy to enrich the genome vocabulary in a coarse-to-fine manner. Extensive experiments on 32 genome datasets demonstrate VQDNA's superiority and favorable parameter efficiency compared to existing genome language models. Notably, empirical analysis of SARS-CoV-2 mutations reveals the fine-grained pattern awareness and biological significance of learned HRQ vocabulary, highlighting its untapped potential for broader applications in genomics.

Submission Number: 745

问题： 类似于自然语言模型，预训练的基因组语言模型通过无监督序列建模来捕捉基因组中的复杂性，成为生物学领域研究人员和从业者的基本工具。然而，这些模型中使用的手工设计的分词策略存在局限性，可能无法充分捕捉基因组数据中的关键模式，限制了模型的性能和适用性。

方法： 作者提出了一种名为VQDNA的通用框架，通过基因组词汇学习来改进基因组分词。**VQDNA利用向量量化码本作为可学习的词汇，以自适应的方式将基因组分词为模式感知的嵌入，并实现端到端训练。** 此外，提出了分层残差量化（HRQ）方法，通过设计层次结构的不同尺度的码本，以粗到细的方式丰富基因组词汇。在32个基因组数据集上的实验表明，VQDNA在性能和参数效率方面优于现有的基因组语言模型。特别是对SARS-CoV-2突变的实证分析显示了所学习的HRQ词汇的细粒度模式感知能力和生物学意义，突显了其在基因组学中更广泛应用的潜力。

1、背景

基因组学研究生物体内的基因组，即完整的DNA指令集合，使科学家能够深入了解生命的分子机制。这一领域提供了关于基因编码和表达的关键洞察，控制着生物体的发育、功能和繁殖，引发了生物学发现的范式转变，揭示了多因子性状、遗传疾病和进化的奥秘。伴随大规模基因组数据的积累，为提取通用模式提供了机会，这些模式可以直接用于微调各种下游任务。受到自然语言模型成功的启发，基因组语言模型通过将基因组表示为语言进行无监督序列建模。例如，DNABERT首次探索了人类基因组的语言模型预训练；Nucleotide Transformers在多物种基因组上进行预训练，增加了跨物种多样性；HyenaDNA针对超长序列问题，达成了精度和效率的良好平衡；DNABERT-2则首次引入了字节对编码（BPE）以合并可能具有基因组学意义的共现核苷酸。 在此背景下，分词已成为基因组语言模型的

关键部分，影响模型对基因组的理解。常用的k-mer方法使用滑动窗口将相邻的k长核苷酸片段组合，而BPE通过统计迭代合并频繁共现的片段。尽管有更精细的分词策略，但这些手工设计的方法可能无法充分表达基因组中有限的四种碱基（A, T, C, G）的信息，难以确保生成的词嵌入能够捕捉最具辨别力的基因组模式。作者提出了一种新的观点：如果能够从输入基因组中学习到一种记录最具辨别力模式的词汇，就可以作为工具进行模式感知的嵌入，助力后续预训练。

为此，作者将基因组分词重新构建为一个辨别性基因组词汇学习问题，提出了VQDNA框架，抛弃手工设计的方案，完全依赖VQ-VAE分词器，这个分词器使用VQ代码本计算模式感知嵌入，作为可在线优化的基因组词汇表。作者在GUE基准上使用28个数据集和4个额外的基因组数据集全面评估了VQDNA的有效性，输入序列长度从63到32k不等。进一步研究超长序列问题时，将VQDNA（HRQ）的输入长度扩展至32k，与HyenaDNA在物种分类任务中进行了公平比较。大量实验表明，VQDNA作为多物种基因组序列建模的通用框架，能够处理大规模和多样化的基因组分析任务，并在32个不同输入长度的数据集上达到了最先进的水平，同时实现了复杂性和精度的良好平衡。此外，对SARS-CoV-2的实证分析展示了HRQ词汇的细粒度模式感知和生物学意义，表明其在更广泛生物学应用中的潜力。

文章的主要贡献如下：

1. 从基因组词汇学习的新视角推进了基因组分词的边界，提出了VQDNA框架，用于端到端的模式感知基因组语言分词。
2. 设计了一种HRQ分词器，通过分层结构逐步丰富原有有限的基因组词汇，使用更少的参数达到与最先进模型相当的性能。
3. 广泛的实验验证了VQDNA的卓越通用性，并在SARS-CoV-2突变的实证研究中显示了其生物学意义和潜力。

2、研究内容

文章旨在开发一个通用框架，将向量量化（VQ）码本作为可学习的基因组词汇，用于自适应地将输入数据标记为模式感知的词嵌入，以实现基因组序列建模并支持多种下游任务。其核心思想是通过最近邻查找学习一种由离散编码嵌入组成的判别性基因组词汇，用于基因组标记化。通过优化该词汇以最小化量化目标，码本嵌入能够在完全自监督的范式下表示一种模式感知的数据簇字典。具体来说，作者介绍了一个三阶段的VQDNA训练框架（如图1所示），用于多物种基因组序列建模。首先，在标记化过程中引入著名的VQ-VAE模型，以替代手工设计的方法，实现模式感知的基因组词汇学习。此外，有限的基因组序列词汇可能会妨碍判别性码本的学习，导致原始四种核苷酸中的细粒度模式丢失。为解决这一问题，

作者提出了分层残差量化（HRQ）方法，通过逐级构建多尺度码本的层次结构，以粗到细的方式逐步丰富基因组词汇。最后，作者详细描述了VQDNA词汇学习的具体实现细节。

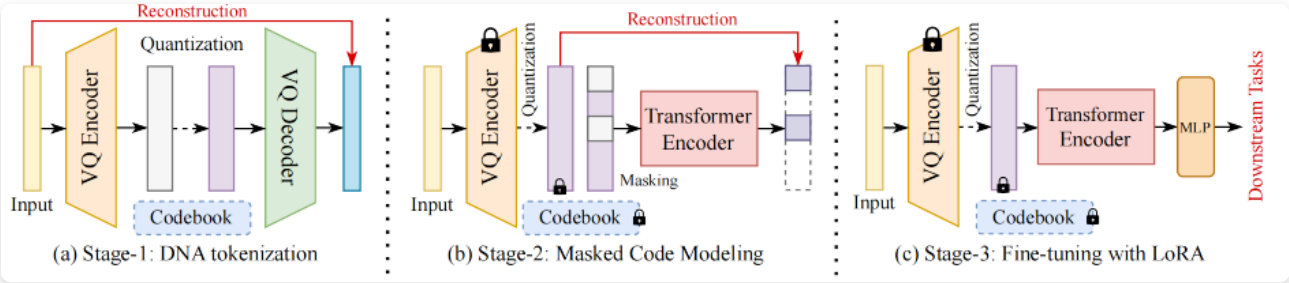


图1 VQDNA 三阶段训练流程概览。(a) 利用大规模多物种基因组序列学习 VQ 基因组词汇。(b) 利用冻结的基因组词汇对 Transformer 编码器进行屏蔽建模预训练。(c) 利用 MLP 头微调预训练编码器，以完成各种下游基因组分析任务。

2.1、矢量量化基因组词汇学习

给定输入基因组序列 $X \in \mathbb{R}^{L\times d}$，编码器 $E_\theta(\cdot)$ 使用参数 θ 将 X 映射到潜在空间 $Z = E_\theta(X) \in \mathbb{R}^{L\times D}$。使用有限的词汇 K 作为向量量化（VQ）码本，$\mathcal{C} = \{(k, e(k))\}_{k\in[K]}$，其中每个码（索引）$k$ 具有其可学习的码嵌入向量 $e(k) \in \mathbb{R}^D$，表示 Z 可以通过逐元素的码映射函数 $Q(\cdot,\cdot)$ 进行量化：

$$M_i = Q(Z_i; C) = \arg\min_{k\in[K]} \|Z_i - e(k)\|_2$$

其中 $1 \le i \le L, M \in [K]^L$ 表示码映射索引。因此，潜在变量 Z_i 可以通过距离最近的码本 $\mathcal{C}$ 中的 1-of-K 嵌入向量来索引和量化。

量化后的嵌入码 M_i 可以表示为：

$$\tilde{Z}_i = e(M_i).$$

解码器 $G_\phi(\cdot)$ 使用参数 ϕ 将量化后的嵌入 $\tilde{Z}$ 映射回输入基因组序列空间，以重构 $\hat{X}$：

$$\hat{X} = G_\phi(\tilde{Z}) = G_\phi(e(M)).$$

由于量化过程不可微分，因此在反向传播计算中使用直通估计器（STE）作为梯度近似。为了优化整体框架，模型通过最小化 VQ-VAE 损失 $\mathcal{L}_{VQ}$ 来优化：

$$\mathcal{L}_{VQ} = \underbrace{\mathcal{L}_{CE}(X, \hat{X})}_{\mathcal{L}_{rec}} + \underbrace{\|\text{sg}[Z] - \tilde{Z}\|_2^2}_{\mathcal{L}_{code}} + \beta\underbrace{\|Z - \text{sg}[\tilde{Z}]\|_2^2}_{\mathcal{L}_{commit}},$$

其中 $\mathrm{sg}[\cdot]$ 表示前述的停止梯度操作符，$\beta \in [0,1]$ 是一个折中超参数（默认值为 0.5）。第一个术语 $\mathcal{L}_{rec}$ 表示重构损失，用于在 VQ-VAE 词汇学习过程中优化编码器和解码器（图1中的阶段1）。中间项 $\mathcal{L}_{code}$ 计算平方误差，作为码本损失，通过推送嵌入向量朝编码器输出方向来更新码嵌入。第三项 $\mathcal{L}_{commit}$ 是一个约束损失，确保在 VQ-VAE 词汇学习过程中，码映射函数 $Q(\cdot,\cdot)$ 的训练稳定性。

在本文中，作者通过指数移动平均（EMA）更新嵌入来优化码本 $\mathcal{C}$，而不是使用损失 $\mathcal{L}_{code}$：

$$\tilde{Z}_i = (1-\alpha)Z_i + \alpha\tilde{Z}_i,$$

其中 α 是动量系数。方程 (4) 中码本的 EMA 更新可以减少由不同批次中某些代码频率变化较大所引起的训练不稳定性。

在获得学习到的码本 $\mathcal{C}$ 后，作者将其作为现成的基因组词汇，用于将基因组序列标记化为模式感知的基因组嵌入，以用于语言模型的预训练。随后，将这些标记化的数据存储起来，用于第二阶段的预训练，并对VQDNA进行与DNABERT-2相同的掩码预训练和下游微调。通过VQ-VAE的基因组词汇学习视角重新定义了基因组标记化，这种方法具有明显的优势：**(i) 基因组数据的序列特性与VQ计算天然匹配。基因组内的核苷酸碱基对不仅能够形成局部的基序（如启动子元件），还能调节全局染色质状态，这种特性类似于图像像素在计算机视觉中的表现，而VQ在该领域已经取得了优势。量化后的后验证明在压缩复杂多模态分布时非常有效，能够在没有人为规则和偏差的限制下编码出最具判别性的基因组模式。** **(ii) 基因组上下文在基因组分析任务中起着至关重要的作用。与现有的仅关注序列内核苷酸更好合并的标记化方法不同，VQ标记化器通过将整个输入纳入其码本优化过程中，能够自然地记录基因组上下文，而不仅仅考虑序列内的依赖关系。** 第3.4节中的实证分析展示了VQ在类内和类间模式感知上的有效性。本节的其余部分将进一步扩展分层残差量化（HRQ），以进一步提升基因组词汇学习的极限。

2.2、分层残差量化

尽管VQ-VAE标记化器能够带来显著的优势，其主要通过扩大码本规模来增强其能力。然而，仅仅扩大码本规模并不高效，因为这会导致码本崩溃问题，更重要的是，这种方法可能不适合基因组数据的特性。基因组数据本质上是由四种核苷酸碱基（A、T、C、G）组成的序列，相比于图像或自然语言等其他模态，基因组的词汇空间非常有限。从VQ标记化的角度来看，这样受限的词汇空间可能过于粗糙，难以提供足够的细节来进行感知丰富的码本学习。因此，作者认为，有必要设计一种特定的策略来解开这些受限核苷酸中的潜在复杂性，以实现判别性的基因组词汇学习。

受到多尺度感知在视觉识别中成功应用的启发，作者可以将这种成功从计算机视觉迁移到基因组学中，即构建多尺度的码本作为多粒度的基因组词汇，并使用相应的词汇对不同层次的输入进行标记化，这些层次可以通过残差技术进行分层对齐。为此，作者提出了分层残差量化（HRQ），设计出一系列码本，以粗到细的方式扩展基因组词汇。

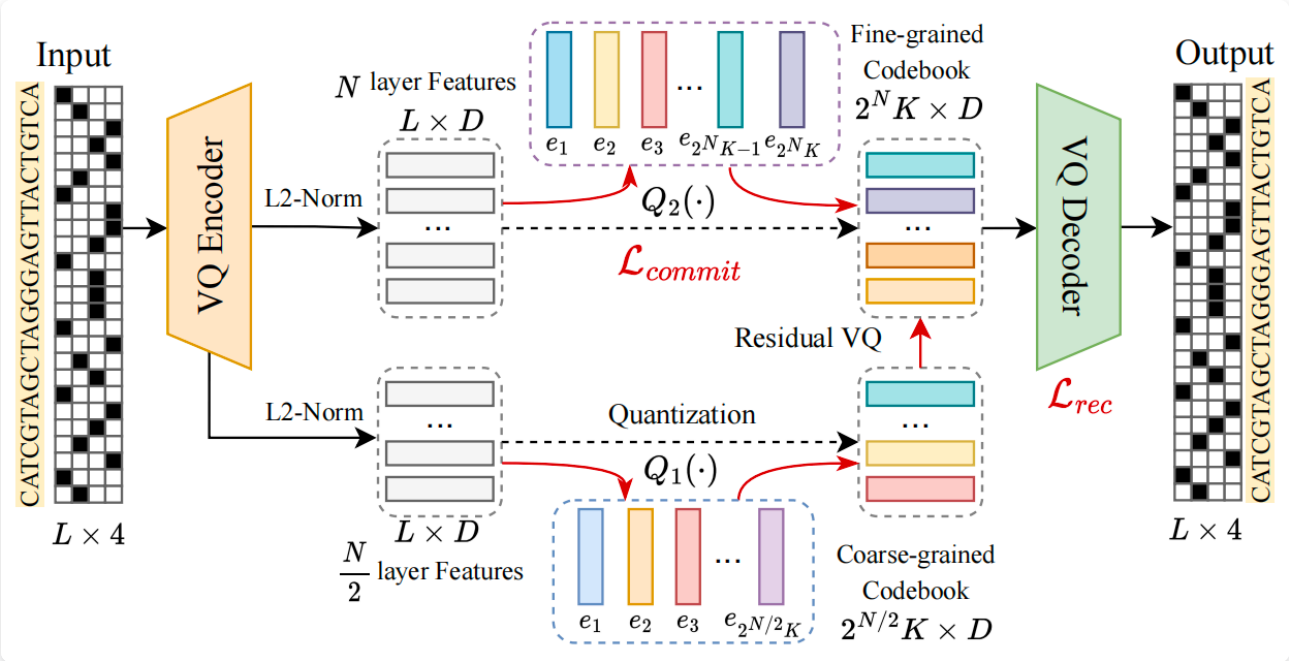


图 2. 分层残差量化（HRQ）作为 VQDNA 框架基因组词嵌入的示意图。在实际应用中，作者将 HRQ 实例化为一个 6 层编码器和解码器，在第 3 层和第 6 层的输出后分别有两个分层编码本。

在图2中所示，多尺度的码本设计成一种分层结构，较粗粒度的语义集中在较低层，而细粒度的细节集中在较高层。量化从编码器层1到N顺序执行。给定第n层编码器的分层输入 $H^{(n)} \in \mathbb{R}^{L \times D}$，对应的 $2^n \cdot K$ 大小的码本 $C^{(n)} = \{(k^{(n)}, e(k^{(n)}))\}_{k \in [2^n K]}$，其中每个码本嵌入向量 $e(k^{(n)}) \in \mathbb{R}^D$ 被定义。因此，每个表示 $H_i^{(n)}$ 通过相同的码本映射算子 $Q(\cdot, \cdot)$ 在公式 (1) 中被量化为：

$$M_i^{(n)} = Q(H_i^{(n)}, c^{(n)}) = \arg \min_{k \in [2^n K]} \|H_i^{(n)} - e(k^{(n)})\|_2,$$

其中 $1 \leq i \leq L, M^{(n)} \in [2^n K]^L$ 表示 HRQ 码本索引 $H^{(n)}$ 的映射。由此，作者构造了一个具有不同感知粒度的码本层次结构，用于从粗到细的层级编码化。对于给定的码本 $M_i^{(n)}$，第n层的潜在特征可以被量化为：

$$\hat{H}_i^{(n)} = e(M_i^{(n)}).$$

但是，一个挑战是，给定来自第n层编码器的输出 $Z^{(n)} \in \mathbb{R}^{L \times D}$，如何将分层输入 $H^{(n)}$ 与 $Z^{(n)}$ 相关联以形成一个统一的HRQ架构。虽然残差量化（RQ）以利用多个码本的训练，

但他们的方法本质上是为单一输入的递归量化而设计的，未解决多个输入的问题。为了解决这个问题，作者定义了一种策略，将分层输入 $H_i^{(n)}$ 和 $Z_i^{(n)}$ 关联起来，公式如下：

$$H_i^{(n)} = \begin{cases} 2Z_i^{(n)} - e(M_i^{(n-1)}), & \text{for } n = 2, \cdots, N, \\ e(M_i^{(1)}), & \text{otherwise.} \end{cases}$$

其中 $1 \leq n \leq N, 1 \leq i \leq L$。

从最初的量化 $H_i^{(1)} = e(M_i^{(1)})$，HRQ计算公式 (6) 中的码本映射 $M_i^{(n)}$，它与 $Z_i^{(n+1)}$ 一起生成下一级量化的分层输入 $H_i^{(n+1)}$。这样做的目的是解决 $H_i^{(n)}$ 和 $Z_i^{(n)}$ 之间的一致性，同时保持HRQ层的比例一致性，因为这种一致性已经被证明在更好地利用多个码本方面非常有效。如图 3 所示，作者将所提出的HRQ策略与著名的 RQ 进行了比较。直观地说，通过加倍输入残差计算出的表征在各层之间表现出更有利的尺度一致性。

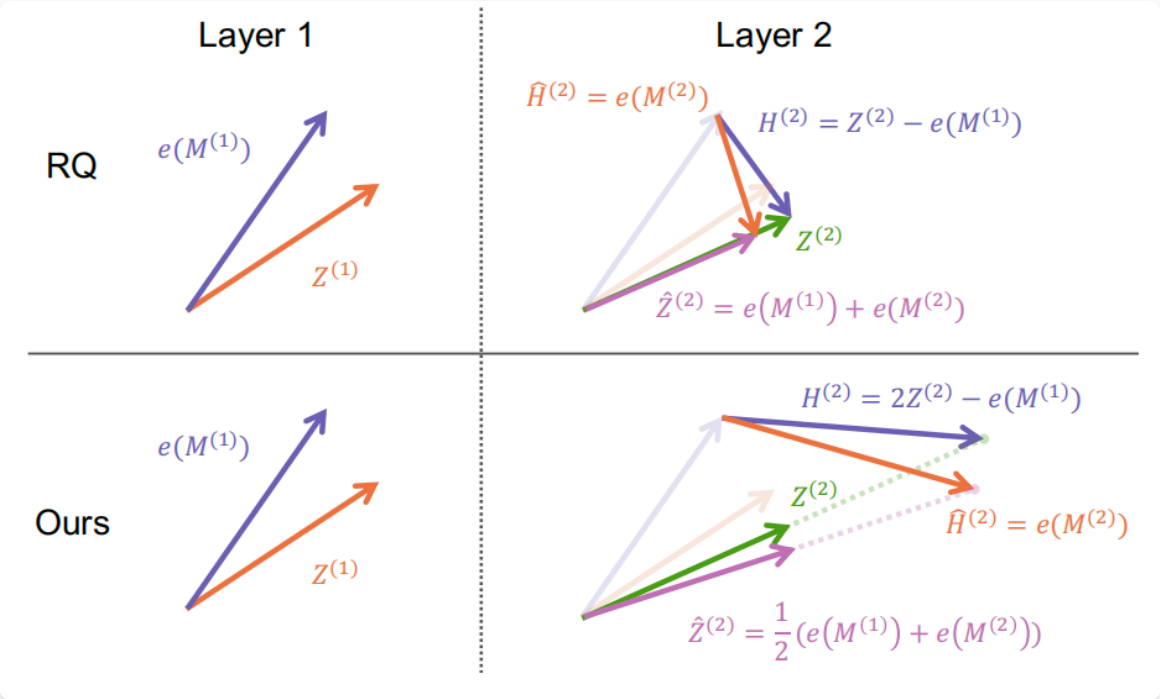


图 3. 在二维空间中展示RQ和作者提出的HRQ在两层量化情况下的示意图。紫色表示当前的分层输入 H^(n)（或在 RQ 中的残差），绿色表示第二层编码器输出 Z^(2)，橙色表示输入 Z^(1) 和每层的输出分层嵌入 ^H^(n)，浅兰花色表示在 n 层量化后得到的最终嵌入 ^Z^(n)。这里说明展示了如何在二维空间中用不同的颜色区分HRQ过程中的各个重要元素。

通过这种方式，作者获得了一个学习的码本层次结构。然后可以用作现成的基因组词汇，在量化的N层后，将输入基因组转化为一组层次嵌入：

$$HRQ(Z_i, C, N) = (\hat{H}_i^{(1)}, \cdots, \hat{H}_i^{(N)}),$$

其中 $\hat{H}_i^{(n)} = e(M_i^{(n)}) \in \mathbb{R}^{L\times D}$ 表示第n层的量化基因组嵌入。作者将HRQ的最终输出嵌入定义为：

$$\hat{Z}_i = \frac{1}{N}\left(\sum_{n=1}^{N}\hat{H}_i^{(n)}\right),$$

$\hat{Z}_i$汇总了所有N个量化层的分层嵌入 $\hat{H}_i^{(n)}$ 来保持比例一致性。

2.3、 HRQ的训练

作者提出的HRQ的总体学习目标定义如下：

$$\mathcal{L}_{HRQ} = \mathcal{L}_{CE}(X,\hat{X}) + \beta\sum_{n=1}^{N}\|Z^{(n)} - \mathrm{sg}[\hat{Z}^{(n)}]\|_2^2,$$

其中 $\beta > 0$ 是等式(4)中的超参数，第一项是重构损失 $\mathcal{L}_{rec}$。作者还使用了广泛应用的嵌入指数移动平均（EMA）方法来更新码本 C 代替公式(4)中的码本损失 $\mathcal{L}_{code}$。

3、实验结果

3.1、实验设置

在预训练阶段，作者遵循DNABERT-2的预训练策略，分别在包含2.75B核苷酸的人类基因组和包含32.49B核苷酸的多物种基因组上对VQ标记化器和BERT-Base Transformer编码器进行预训练。在第一阶段预训练中，VQDNA变体使用AdamW优化器进行一个epoch的训练，批大小为1024，基本学习率为1×10^{-4}，并由余弦调度器调整，使用8块GPU。在第二阶段预训练中，对标记化的VQ嵌入应用25%的随机掩码语言模型（MLM）训练，训练500k步，初始学习率为5×10^{-4}，批次大小为2048。

在第三阶段的下游任务适配中，作者遵循GUE基准中的微调评估设置。预训练的Transformer编码器在28个GUE数据集、3个EEP数据集以及物种分类数据集上，使用AdamW优化器结合LoRA进行微调。输入核苷酸序列的最大长度为512，并在此基础上报告GFLOPs。下游任务的评估指标包括Top-1准确率（Acc）、F1分数（F1）、Matthews相关系数（MCC）和斯皮尔曼相关系数（SC）。所有实验使用PyTorch、transformers库和NVIDIA A100 GPU进行，报告了3次试验的平均结果。

3.2 方法比较

作者将VQDNA变体与其他主流基因组语言模型进行了比较，包括DNABERT（3-mer）、Nucleotide Transformer（NT）变体和DNABERT-2（见表2），结果显示VQDNA变体在整体性能上取得了最佳和次佳的排名。在GUE基准测试中（见表5、表3和表4），作者也对VQDNA变体进行了评估，涉及7个常用的基因组任务，包括酵母的表观遗传标记预测（EMP）、小鼠和人类基因组的转录因子预测（TFP-M和TFP-H）、新冠病毒变异分类（CVC）、启动子检测（PD）、核心启动子检测（CPD）和剪接位点预测（SSP）。

Method	Tokenizer	Usage	Lin.	FT
DNABERT	6-mer	47	23.54	55.50
NT-2500M-1000g	6-mer (non)	47	23.54	66.73
HyenaDNA	one-hot	100	5.47	54.10
DNABERT-2	BPE (6-mer)	99	36.53	71.02
VQDNA	**VQVAE**	100	44.76	73.16
VQDNA	**HRQ**	100	**48.87**	**74.32**

表1 标记化效率分析：作者报告了在Covid变种分类任务中的标记化器类型、标记使用率（%）、线性探测（Lin.）和全量微调（FT）的宏F1分数（%），具体如第3.3节所述。需要注意的是，6-mer（non）使用的是不重叠的6-mer标记化，而BPE（6-mer）是通过迭代合并6-mer标记化中最常见的编码来实现的。标记使用率表示每种标记化器中所使用的标记占总标记的百分比。

Method	Date	Tokenizer	# Params. (M)	FLOPs (G)	Train (B)	Average Rank
DNABERT	BioInfo'2021	3-mer	86	3.3	122	5
NT-500M	biorxiv'2023	6-mer	480	3.2	50	6
NT-2500M	biorxiv'2023	6-mer	2537	19.4	300	4
DNABERT-2	ICLR'2024	BPE	117	1.0	262	3
VQDNA	**Ours**	**VQVAE**	86+16	1.1+0.5	262	2
VQDNA	**Ours**	**HRQ**	86+17	1.1+0.6	262	**1**

表2 在32个基因组下游任务上的平均性能排序、标记发生器类型、模型参数和工作流程，以及预训练标记。

Method	PD			CPD			TFP (Human)				
	all	notata	tata	all	notata	tata	0	1	2	3	4
DNABERT (3-mer)	90.44	93.61	69.83	70.92	69.82	78.15	67.95	70.90	60.51	53.03	69.76
NT-500M-1000g (6-mer)	89.76	91.75	78.23	66.70	67.17	73.52	63.64	70.17	52.73	45.24	62.82
NT-2500M-1000g (6-mer)	90.95	93.07	75.80	67.39	67.46	69.66	66.31	68.30	58.70	49.08	67.59
DNABERT-2 (BPE)	86.77	94.27	71.59	69.37	68.04	74.17	71.99	76.06	66.52	58.54	77.43
VQDNA	90.20	94.05	73.08	70.36	69.87	77.63	72.04	75.89	66.69	58.31	77.63
VQDNA (HRQ)	**90.75**	**94.48**	**74.52**	**71.02**	**70.58**	**78.50**	**72.48**	**76.43**	**66.85**	**58.92**	**78.10**

表3 MCC（%）启动子检测（PD）、核心启动子检测（CPD）和转录因子预测（TFP）任务的性能在GUE基准上进行微调。

Method	TFP (Mouse)					CVC	SSP	EEP (gRNA)		
	0	1	2	3	4	Covid	Reconstruction	K562	Jurkat	H1
DNABERT (3-mer)	42.31	79.10	69.90	55.40	41.97	62.23	84.14	88.63	86.89	62.72
NT-500M-1000g (6-mer)	39.26	75.49	64.70	33.07	34.01	52.06	80.97	90.58	88.94	63.80
NT-2500M-1000g (6-mer)	48.31	80.02	70.14	42.25	43.40	66.73	85.78	90.90	89.34	66.87
DNABERT-2 (BPE)	56.76	84.77	79.32	66.47	52.66	71.02	84.99	91.02	89.27	66.91
VQDNA	57.52	85.36	79.78	68.45	54.10	73.16	88.06	91.16	89.83	67.56
VQDNA (HRQ)	**58.34**	**85.81**	**80.39**	**69.72**	**54.73**	**74.32**	**89.53**	**91.53**	**90.12**	**67.98**

表4 转录因子预测（TFP）、Covid变异分类（CVC）、剪接位点预测（SSP）和编辑效率预测（EEP）任务的性能。TFP和SSP使用MCC（%），而CVS和EEP报告F1（%）和MCC（%）。

Method	Epigenetic Marks Prediction									
	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2	H3K4me3	H3K79me3	H3K9ac	H4	H4ac
DNABERT (3-mer)	74.15	42.07	48.49	42.95	31.34	28.92	60.12	50.48	78.27	38.60
NT-500M-1000g (6-mer)	72.52	39.37	45.58	40.45	31.05	26.16	59.33	49.29	76.29	36.79
NT-2500M-1000g (6-mer)	74.61	44.08	50.86	43.10	30.28	30.87	61.20	52.36	79.76	41.46
DNABERT-2 (BPE)	78.27	52.57	56.88	50.52	31.13	36.27	67.39	55.63	80.71	50.43
VQDNA	78.56	53.93	60.62	52.84	33.73	38.49	68.15	56.28	81.32	50.33
VQDNA (HRQ)	**79.21**	**54.46**	**61.75**	**53.28**	**34.05**	**39.10**	**68.47**	**56.63**	**81.84**	**50.69**

表5 在GUE基准测试上微调了不同数据集的表观遗传标记预测任务的MCC（%）性能。

Method	1k	20k	32k	250k	450k
HyenaDNA	61.13	87.42	93.42	97.90	99.40
DNABERT	39.61	76.21	91.93	N/A	N/A
DNABERT-2	61.04	86.83	99.28	N/A	N/A
VQDNA (HRQ)	**61.57**	**88.05**	**99.46**	N/A	N/A

表6 随着序列长度的增加，物种分类的Top-1精度（%），其中N/A表示内存不足。

结果表明，两种VQDNA版本在参数更少的情况下，一致优于先前的大规模模型NT-2500M 1000g和高效模型DNABERT-2，而VQDNA（HRQ）相较于VQDNA（VQVAE）有显著的性能提升。此外，在表4中，作者验证了VQDNA变体在编辑效率预测（EEP）任务中（短基因组序列）也能达到最先进的性能。随后，作者将序列长度扩展至与HyenaDNA相同的水平，并在表6中进行了五物种分类任务。尽管HyenaDNA可以对极长序列（如450k）进行微调，VQDNA（HRQ）在输入序列长度为32k时（使用FLASH Attention和梯度检查技术）取得了最佳准确率，这表明所学习到的VQDNA标记化器能够有效捕捉基因组分析中对极长依赖任务的有用上下文和模式。

3.3、 消融实验

作者对VQ码本的设置和MLM预训练的掩码比例进行了消融实验。由于在第一阶段标记化器上进行微调评估成本较高，作者报告了CVC数据集上VQDNA标记化序列的重构准确率和线性探测准确率。首先，作者对VQDNA和HRQ的码本维度和总大小进行了消融实验（见表7(a)）。结果显示，码本大小为512在重构能力和判别能力之间取得了良好的平衡，能够捕捉更多内在模式。接着，表7(b)显示码本维度对学习到的表示影响较小，因此作者选择384作为默认维度以提高效率。

Code size	VQDNA		+HRQ	
	Rec.	Lin.	Rec.	Lin.
128	98.2	42.1	98.4	42.8
256	98.8	43.6	99.1	47.7
512	99.5	**44.8**	99.6	**48.9**
1024	**99.6**	44.5	**99.8**	48.2

(a)

Code dim.	VQDNA		+HRQ	
	Rec.	Lin.	Rec.	Lin.
256	99.4	44.3	99.5	48.2
384	99.5	**44.8**	99.6	**48.9**
768	99.6	44.6	99.6	**48.9**
1024	99.8	44.7	99.7	48.8

(b)

Mask ratio	VQDNA		+HRQ	
	H3	CVC	H3	CVC
15%	77.9	72.6	78.3	73.7
20%	78.3	**73.4**	78.8	74.2
25%	**78.6**	73.2	**79.2**	**74.3**
30%	77.4	73.0	78.6	73.9

(c)

表7 (a) VQDNA标记化器中总码本大小的消融研究。(b) VQDNA 标记化器中的编码本维度 (dim.) 消融研究。(c) VQDNA中，第 2 阶段 MLM 预训练中的掩码比率分析。

随后，作者在表7(c)中分析了掩码比例，报告了H3和CVC数据集上的微调结果。作者发现25%的掩码比例能帮助VQDNA学到比之前模型（15%或20%）更好的表示。由此，作者推测，VQDNA标记化器可能学到了丰富的上下文信息，使得MLM能够使用更大的掩码比例来增加预测任务的难度。

3.4、 SARS-CoV-2分析

SARS-CoV-2导致了COVID-19疫情，这场危机成为本世纪最严重的公共卫生事件之一。随着病毒在全球迅速传播，2020至2021年间出现了多种SARS-CoV-2变种，如Alpha（B.1.1.7）、Beta（B.1.351）、Delta（B.1.617.2）、Eta（B.1.525）、Iota（B.1.526）、Kappa（B.1.617.1）、Lambda（C.37）、Gamma（P.1）和Zeta（P.2），每种变种携带独特的突变，由Pango谱系标识符标记（O'Toole等，2022）。这些变种在短时间内快速变异，可能导致逃避免疫反应和抵抗现有疫苗及治疗手段，构成了严峻的挑战。

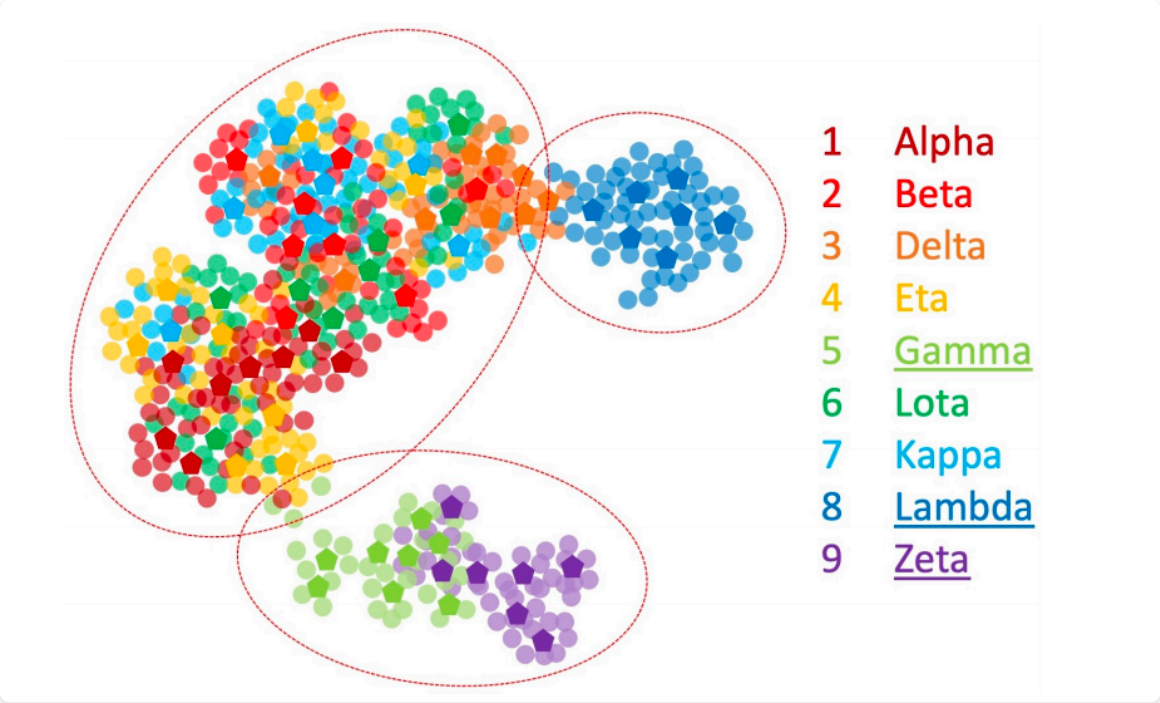


图4 通过UMAP对CVC数据集上的HRQ码本进行可视化。每个码本标签是通过计算线性分类器（基于HRQ标记化序列学习）中最相关的类别，并使用Grad-CAM得到的。五边形点代表第3层码本的编码，浅色圆点代表第6层码本的编码。结果显示，HRQ词汇在类内外的模式感知能力非常出色。

鉴于其现实意义，作者对这一问题进行了实证分析，以验证VQ标记化器的有效性。图4显示，作者的HRQ标记化器能够学习判别性基因组嵌入，相同谱系的变种聚类在一起，而不同谱系的变种分离开来，展示了其对谱系内外模式的识别能力。此外，扩展的码本成功捕捉到了细粒度的模式。例如，Lambda变种由Delta变异而来，虽然有部分相似特征，但属于不同谱系。图5中的Lambda簇靠近Delta簇，体现了HRQ的生物学意义。

4、 总结

文章提出了VQDNA，一种利用VQ码本作为可学习基因组词汇的创新框架，避免了手工偏差，实现了模式感知的基因组标记化。为进一步提升VQ标记化器的性能，作者提出了分层残差量化（HRQ），通过多尺度码本的层次设计丰富基因组词汇。大量实验显示，VQDNA在32个数据集上表现出色，具有优异的泛化能力和生物学意义。

这篇文章的局限性包括：(1) VQDNA需要额外的训练阶段，增加了计算成本，仍需进一步降低计算开销以提高在多组学中的适用性；(2) 受限于计算资源，VQDNA的模型规模尚未达到最大化，未来可以通过增加模型参数和预训练数据扩展其优势；(3) 鉴于HRQ词汇在SARS-CoV-2突变中的生物学意义，应进一步探索其在基因组学中更广泛的应用，如生成任务等。这些方向为未来研究提供了机会。