

一种多重混样直接 RNA 纳米孔测序方法及系统

优先权申请

[0001] 本申请要求 2023 年 3 月 17 日提交的中国发明专利申请 CN2023102647091 “一种新型直接 RNA 纳米孔测序方法及系统”的优先权,该优先权发明专利申请以引用方式全文并入。

技术领域

[0002] 本发明涉及测序领域,具体涉及一种多重混样直接 RNA 纳米孔测序方法及系统。

背景技术

[0003] 三代测序技术的发展彻底改变了基因组和转录组的长读测序的能力。它能够在无需碎片化和扩增的情况下,直接对 DNA 和 RNA 及其修饰进行测序。牛津纳米孔技术(ONT)的直接 RNA 测序(DRS)技术可以通过一系列蛋白质纳米孔在合成膜上直接对 RNA 分子进行测序。

[0004] 当马达酶缓慢地将单个 RNA 分子通过纳米孔时,不同的核苷酸阻塞了纳米孔时会导致电流波动,这可以通过碱基调用算法记录下来以确定 RNA 分子的序列。直接 RNA 测序技术可以产生长读序列(长达 21 kb),旨在获得完整的转录本或接近完整的 RNA 基因组。完整的测序使其在测量 RNA Poly(A)(聚(A))尾长度和分析聚(A)尾对基因表达调控的调节功能方面具有强大的作用。此外,直接 RNA 测序可以直接检测 RNA 分子的修饰,例如 m6A、7-甲基鸟苷(m7G)、伪尿嘧啶(Ψ)和 2'-O-甲基化(Nm),因为这些修饰会产生不同的电流波动。

[0005] 此外,无扩增文库制备方法可以避免扩增或逆转录(RT)偏差,特别适用于识别出外显子内含物(exitrons),这些外显子内含物很可能源自逆转录造成的伪迹。这些优势使直接 RNA 测序技术在特定亚型的 RNA 结构确定、tRNA 和 rRNA 鉴定以及 RNA 病毒、原核生物和寄生虫等生物体基因组测序方面具有特殊的优势。近期的研究还利用直接 RNA 测序技术在广泛的物种中进行转录组或表观转录组分析,包括人类、动物(线虫、昆虫)、植物、浮游动物和病原体等。直接 RNA 测序技术的另一个应用是通过使用碱基类似物标记新生 RNA,例如 5-乙炔基尿嘧啶和 4-硫尿嘧啶,随后进行 ONT 直接 RNA 测序来分析 RNA 代谢动力学。

[0006] 尽管 DRS 具有独特的优点,但它有两个主要的限制,包括缺乏多样本混样测序方法和需要大量 RNA,通常每次测序需要 500ng RNA。众所周知,ONT 提供用于多路复用 cDNA 文库的条形码协议,这些协议依赖于直接连接 DNA 适配器到 cDNA 序列上。在这种情况下,条形码和 cDNA 序列都可以在 DNA 模型下轻松地被识别。然而,在 DRS 的情况下,由于适配器是 DNA 寡核苷酸,而目标片段是 RNA 分子,因此 DNA 适配器不能在 RNA 模型或 DNA 模型下被正确识别。因为 DNA 的过孔速度(450 bp/sec)比 RNA(70 nt/sec)快 6 倍。为了克服这些限制,已经有团队开发了 DeePlexiCon 这一方法,将有限的 RNA 样品混入到一张芯片中,生成了多路复用(multiplex)直接 RNA 测序数据集,以此降低单个样品的测序成本。但由于条形码设计和算法的限制,该方法只能在一张芯片中多路复用 4 个样品。尽管已经有文献报道 DeePlexiCon 可以训练新模型,但如果使用不同的序列和多达 4 个以上的条形码,它将无法准确地解复用读取。因此,提供更多条形码以增加多路复用能力并实现高精度仍然是一个挑战。

发明内容

[0007] 第一方面,本发明提供了一种直接 RNA 纳米孔测序方法(简称为 DecodeR),其特征在于,所述方法用于对 n 种测试样本 RNA 多核苷酸进行测序,其中 $2 \leq n \leq 24$ 且 n 为整数,所述方法包括:

[0008] S1 针对每一种测试样本 RNA 多核苷酸,提供一种双链 DNA 多核苷酸来捕获一种待测 RNA 序列并形成第一杂交结构;所述双链 DNA 多核苷酸为逆转录接头的前向链和后向链组装成的双链结构,所述前向链从 5'至 3'方向包括条形码序列和第一连接序列,所述后向链

从 5'至 3'方向包括第二连接序列、条形码序列和 Poly(T)序列, 所述待测 RNA 序列包括所述测试样本 RNA 多核苷酸和 Poly(A)序列, 所述双链 DNA 多核苷酸通过所述 Poly(T)序列捕获所述待测 RNA 序列; 所述条形码序列对所述每一种测试样本 RNA 多核苷酸具有特异性, 所述条形码序列选自 SEQ ID NOs:1-24 中的任意一种;

[0009] S2 将所述第一杂交结构逆转录, 形成第二杂交结构;

[0010] S3 将测序接头与所述第二杂交结构连接, 形成第三杂交结构;

[0011] S4 对所述第三杂交结构进行纳米孔测序, 获得原始测序数据;

[0012] S5 提取所述原始测序数据对应的电流信号数据;

[0013] S6 将所述电流信号数据中所述条形码序列的电流信号数据进行分割, 并调用基于所述条形码序列训练得到的分类模型进行分析, 获得所述条形码序列的序列信息, 进而获得所述 n 种测试样本 RNA 多核苷酸的分类信息。

[0014] 在一些实施例中, 所述双链 DNA 多核苷酸经预退火生成。

[0015] 在一些实施例中, 所述预退火的方法包括将每个所述逆转录接头的前向链和后向链混合在缓冲液中, 按照以下程序进行: 95°C 变性 5 分钟, 65°C、50°C、37°C、22°C 分别退火 30 分钟, 4°C 保持。

[0016] 在一些实施例中, 所述第一连接序列包括 TAGTAGGTTC。

[0017] 在一些实施例中, 所述第二连接序列包括 GAGGCGAGCGGTCAATTTT。

[0018] 在一些实施例中, 所述 Poly(T)序列包括 10-30 个重复 T 序列。

[0019] 在一些实施例中, 所述 Poly(A)序列包括 10-30 个重复 A 序列。

[0020] 在一些实施例中, 所述 Poly(T)序列包括 10 个重复 T 序列。

[0021] 在一些实施例中, 所述 Poly(A)序列包括 10 个重复 A 序列。

[0022] 在一些实施例中, 所述分类模型的训练方法包括将训练样本数据中条形码序列的电流信号数据进行分割, 利用机器学习算法对分割后的条形码序列的电流信号数据进行训练学习, 从而确定所述分类模型, 其中所述训练样本数据包括一种或多种训练样本 RNA 多核苷酸和对应的条形码序列的电流信号数据, 所述条形码序列选自 SEQ ID NOs:1-24 中的任意一种。

[0023] 在一些实施例中, 所述分类模型为基于选自 SEQ ID NOs:1-24 中的任意 2 种、3 种、4 种、5 种、6 种、7 种、8 种、9 种、10 种、11 种、12 种、13 种、14 种、15 种、16 种、17 种、18 种、19 种、20 种、21 种、22 种、23 种或 24 种的所述条形码序列训练得到的分类模型。

[0024] 在一些实施例中, 所述训练样本 RNA 多核苷酸包括体外转录 RNA。

[0025] 在一些实施例中, 所述分割为将所述条形码序列的电流信号数据分割成 70-100 列矩阵。

[0026] 在一些实施例中, 所述分割为将所述条形码序列的电流信号数据分割成 100 列矩阵。

[0027] 在一些实施例中, 所述机器学习算法包括 K 最近邻、神经网络、朴素贝叶斯分类、回归树、AdaBOOST 和随机森林中的一种或多种。

[0028] 在一些实施例中, 所述机器学习算法包括随机森林。

[0029] 在一些实施例中, 所述测试样本 RNA 多核苷酸的来源包括病原体、植物、浮游动物、昆虫、哺乳动物和肿瘤中的一种或多种。

[0030] 在一些实施例中, 所述病原体包括 RNA 病毒、细菌、真菌和寄生虫中的一种或多种。

[0031] 在一些实施例中, 所述体外转录 RNA 包括表 2 所示的基因中的一种或多种。

[0032] 在一些实施例中, 所述分类的分类可能性的阈值>0.3。

[0033] 在一些实施例中, 所述 S5 进一步包括提取所述原始测序数据对应的原始电流信号数据, 并对所述原始电流信号数据进行预处理, 获得预处理后的电流信号数据, 所述预处理包括降噪处理、抛光处理和归一化处理中的一种或多种。

[0034] 在一些实施例中, 所述方法进一步包括将所述一种或多种训练样本 RNA 多核苷酸的

电流信号数据与对应的参考序列比对，以比对质量大于 60、比对至唯一参考序列为条件将不满足条件的读取序列过滤。

[0035] 第二方面，本发明提供了一种用于直接 RNA 测序的试剂盒，其特征在于，所述试剂盒包括如 SEQ ID NOs:1-24 所示的逆转录接头的条形码序列。

[0036] 第三方面，本发明提供了一种直接 RNA 纳米孔测序系统，其特征在于，包括：

[0037] 测序数据接收模块，用于接收和存储测序数据，所述测序数据包括双链多核苷酸的测序数据，所述双链多核苷酸包括第一链和第二链，所述第一链从 5'至 3'方向包括 RNA 多核苷酸、Poly(A)序列、条形码序列、第一连接序列和测序接头，所述第二链从 5'至 3'方向包括测序接头、第二连接序列、条形码序列、Poly(T)序列和所述 RNA 多核苷酸的逆转录序列；所述测序数据包括测试样本数据和训练样本数据；所述测试样本数据包括 n 种测试样本 RNA 多核苷酸对应的测序数据，其中 $2 \leq n \leq 24$ 且 n 为整数，所述条形码序列对每一种测试样本 RNA 多核苷酸具有特异性，所述条形码序列选自 SEQ ID NOs:1-24 中的任意一种；

[0038] 信号处理模块，用于将所述测序数据处理成电流信号数据；

[0039] 分类模块，用于调用基于所述测试样本数据中的所述条形码序列构建的分类模型，对所述测试样本数据的电流信号数据进行分类并输出所述 n 种测试样本 RNA 多核苷酸的分类结果。

[0040] 在一些实施例中，所述系统进一步包括模型训练模块，所述模型训练模块利用机器学习算法对所述训练样本数据的电流信号数据进行训练学习，从而确定所述分类模型。

[0041] 在一些实施例中，所述双链多核苷酸为 RNA/DNA 杂化双链结构。

[0042] 在一些实施例中，所述信号处理模块对所述测序数据进行预处理和分割。

[0043] 在一些实施例中，所述预处理包括降噪处理、抛光处理和归一化处理中的一种或多种。

[0044] 在一些实施例中，所述分割为将所述条形码序列的电流信号数据分割成 70-100 列矩阵。

[0045] 在一些实施例中，所述分割为将所述条形码序列的电流信号数据分割成 100 列矩阵。

[0046] 在一些实施例中，所述机器学习算法包括 K 最近邻、神经网络、朴素贝叶斯分类和回归树、AdaBOOST 和随机森林中的一种或多种。

[0047] 在一些实施例中，所述机器学习算法包括随机森林。

[0048] 在一些实施例中，所述系统进一步包括过滤模块，用于基于设置的分类可能性的阈值，将低于所述阈值的读取过滤掉。

[0049] 在一些实施例中，所述训练样本数据包括体外转录 RNA 的测序数据。

[0050] 在一些实施例中，所述测试样本 RNA 多核苷酸的来源包括病原体、植物、浮游动物、昆虫、哺乳动物和肿瘤中的一种或多种。

[0051] 在一些实施例中，所述病原体包括 RNA 病毒、细菌、真菌和寄生虫中的一种或多种。

[0052] 在一些实施例中，所述阈值 > 0.3 。

[0053] 在一些实施例中，所述体外转录 RNA 包括表 2 所示的基因中的一种或多种。

[0054] 在一些实施例中，所述训练样本数据中用作训练集和测试集的比例为 7:3。

[0055] 在一些实施例中，所述模型训练模块用于将所述训练样本数据的电流信号数据与对应的参考序列比对，以比对质量大于 60、比对至唯一参考序列为条件将不满足条件的读取序列过滤。

[0056] 有益效果

[0057] 现有技术中纳米孔直接 RNA 混样测序分类器 DeePlexiCon 能够将有限的 RNA 样本混入到一张芯片中，生成多路复用 (multiplex) 直接 RNA 测序数据集，以降低单个样本的测序成本。然而，DeePlexiCon 只能在同一芯片上拆分 4 个样本，具有混样样本通量低、拆分准确率低、实现准确拆分需要舍弃大量的测序数据等诸多限制，造成测序成本大大提高。

[0058] 本发明提供的方法和系统（也可以理解为分类器）（简称 DecodeR）提供了 24 种条

形码以在一张芯片中标记多种 RNA 样本，并通过对直接 RNA-seq 数据解复用 (demultiplex) 和机器学习分类，实现在一张芯片中高精度地多路复用更多样本且检测限低，克服了 RNA 样本量低和测序成本高的限制。本发明采用的将条形码电流分割成 70-100 列矩阵，能够降低电流之间的干扰，降低了噪音。经验证，DecodeR 的恢复率和准确率可以达到 100% 和 92.2%。将 DecodeR 应用于多种病原体 (RNA 病毒、细菌、真菌和疟原虫) 和胶质瘤样本，均得到了良好的验证结果。此外，本发明通过 DecodeR 还揭示了不同胶质瘤样本的转录组和突变的一些有意义的发现，这表明本发明具有较大的实际应用潜力。本发明在纳米孔直接 RNA 测序中实现在一张芯片中同时对多个生物样本测序，降低了生物样本起始投入量，降低了多样本测序成本，能够实现微量生物样本的测序，已在肿瘤患者样本中得到应用，并具有在其他临床样本中应用的潜力。

[0059] 综上所述，本发明提供了一种完整、成熟、高通量、低成本的直接 RNA 混样测序方法及系统，提供更多条形码以增加多路复用能力并实现对纳米孔 DRS 中的多种生物样本的高精度的测序和分析，有望在临床样本中实际应用。

[0060] 如本文所使用，“核酸”、“核酸分子”、“多核苷酸”和“寡核苷酸”可互换地使用，是指共价连接的核苷酸序列 (即用于 RNA 的核糖核苷酸和用于 DNA 的脱氧核糖核苷酸)，其中一个核苷酸的戊糖的 3' 位置通过磷酸二酯基团连接到下一个核苷酸的戊糖的 5' 位置。

“多核苷酸”包括单链和双链多核苷酸，例如各种 DNA、RNA 分子、或其任何杂交体或其衍生物或组合或片段。

[0061] 如本文所使用，“测试样本 RNA 多核苷酸”是指期望被测序的 RNA。所述 RNA 可以是编码和/或非编码 RNA，例如 mRNA、siRNA、rRNA、miRNA、tRNA、lncRNA、snoRNA、snRNA、exRNA、piRNA。

[0062] 如本文所使用，“互补”是指两条核酸序列能够根据碱基配对原则 (Watson-Crick 原则) 在彼此之间形成氢键，并由此形成双链结构。在本发明中，“互补”包括实质上互补和完全互补；其中完全互补是指一条核酸序列中的每一个碱基都能够与另一条核酸序列中的碱基配对，而不存在错配或缺口；实质上互补是指一条核酸序列中的大部分碱基都能够与另一条核酸序列中的碱基配对，其允许错配或缺口存在。通常，在允许核酸杂交、退火或扩增的条件下，互补 (例如实质上互补或完全互补) 的两条核酸序列将选择性地/特异性地发生杂交或退火，并形成双链结构。相应地，“不互补” (或“不匹配”) 是指两条核酸序列在允许核酸杂交、退火或扩增的条件下不能发生杂交或退火。

[0063] 如本文所使用，“杂交”和“退火”可互换地使用，是指互补的单链核酸分子形成双链核酸的过程。通常，完全互补或实质上互补的两条核酸序列可发生杂交或退火。

[0064] 如本文所使用，“后端” (或“下游”) 用于描述两条核酸序列的相对位置关系，并且具有本领域技术人员通常理解的含义。例如，“一条核酸序列在另一条核酸序列的后端”是指，当以 5' 端至 3' 端方向排列时，与后者相比，前者位于更靠后的位置 (即更接近 3' 端的位置)。相应地，“前端” (或“上游”) 具有与“后端”相反的含义。

[0065] 如本文所使用，“条形码”是一种特定的核苷酸序列，用于标记期望被测序的 RNA。在本发明中，条形码的序列是经设计的。换句话说，条形码的序列是已知的。在一些实施例中，所述条形码的序列包括 SEQ ID NOs:1-24 中的一种或多种。在一次测序过程中，每一种期望被测序的 RNA 对应的条形码序列是唯一且特定的 (也可以理解为具有特异性)。此外，条形码序列在本发明的测序结构的位置是固定的 (例如在 poly(A) 的下游)，因此可以通过 poly(A) 的连续 A 信号来定位读取序列中的条形码序列。本发明可以通过机器学习算法对基于不同条形码组合 (例如 SEQ ID NOs:1-24 中的中的任意 1 种、2 种、3 种、4 种、5 种、6 种、7 种、8 种、9 种、10 种、11 种、12 种、13 种、14 种、15 种、16 种、17 种、18 种、19 种、20 种、21 种、22 种、23 种或 24 种) 的训练样本数据进行训练学习，以构建分类模型。进而，在一次测序过程中 (例如一张芯片中)，基于期望被测序的 RNA 相连的条形码序列，可以选

择相应条形码的组合训练得到的分类模型来分析测序数据,进而实现对期望被测序的 RNA 的分类。在一些实施例中,在一次测序过程中,本发明提供的方法可以实现对 1 种、2 种、3 种、4 种、5 种、6 种、7 种、8 种、9 种、10 种、11 种、12 种、13 种、14 种、15 种、16 种、17 种、18 种、19 种、20 种、21 种、22 种、23 种或 24 种期望被测序的 RNA 的测序。

[0066] 如本文所使用,“读取”是指通过测序方法获得的核酸分子的任何部分或全部的测序信息。读取可以存储在存储介质中并被适当地处理以确定其是否与参考序列匹配或符合其他标准。可以从测序装置直接获得读取,或者从存储的关于样本的序列信息间接获得。

[0067] 如本文所使用,“参考序列”是指读取与其比对的任何已知序列。在一些实施例中,参考序列可以对应于生物体的基因组或转录组的全部或仅一部分。

[0068] 如本文所使用,“比对”是指将读取与参考序列进行比较的过程,由此确定参考序列是否含有读取序列。如果参考序列含有读取,则读取可以被定位到参考序列,或者被定位到参考序列中的特定位置。比对中读取的匹配可以是 100% 序列匹配或小于 100% (非完全匹配)。

[0069] 如本文所使用,“混样”是指含有不同基因组的核酸分子混合物的样本。

[0070] 如本文所使用,“库”或“文库”是指源自一种或多种核酸样本的核酸分子的集合。在一些实施例中,所述核酸分子包括可鉴别的序列标记(例如条形码)。在一些实施例中,可以将两个或多个文库合并以创建文库池。

[0071] 如本文所使用,“复用”或“多路复用”是指一种或多种核酸样本或其所衍生的库分子在一个池、试管或反应中(例如在一张芯片中)的集合。

[0072] 如本文所使用,“样本”是指包含生物分子的样本,典型地源自生物学流体、细胞、组织、器官或生物体。“生物样本”是指包含生物分子的任何生物来源样本。在本发明中,所述生物分子包括核酸或核酸混合物,优选为 RNA。在一些实施例中,“样本”或“生物样本”可以是液体样本(例如体液,包括血液、血浆、血清、尿液、阴道液、水囊肿液、阴道冲洗液、胸膜液、腹水液、脑脊髓液、唾液、汗液、泪液、痰液、支气管肺泡灌洗液、乳头排出液、来自身体不同部位(例如,甲状腺、乳房)的抽吸液)或固体样本(例如组织样本、活检样本或从中衍生的组织培养物或细胞及其子代)。在一些实施例中,“样本”或“生物样本”可以包括富集特定类型分子的样本,例如核酸。在一些实施例中,“样本”或“生物样本”还可以包括临床样本,例如通过手术切除获得的组织、通过活检获得的组织、培养中的细胞、细胞上清液、细胞裂解物、组织样本、器官、骨髓、血液、血浆、血清等。在一些实施例中,“样本”或“生物样本”还可以包括肿瘤样本,例如来自癌症患者的细胞(例如癌细胞)的样本或从癌症患者的癌细胞(例如,细胞裂解物、细胞提取物)中获得的核酸样本。在一些实施例中,“样本”或“生物样本”中的核酸可以是游离的。在一些实施例中,“样本”或“生物样本”的来源可以是病原体,例如 RNA 病毒、细菌、真菌、寄生虫。在一些实施例中,“样本”或“生物样本”的来源可以是植物、浮游动物、昆虫、哺乳动物(例如人类、狗、猫、马、山羊、绵羊、牛、猪、大鼠、小鼠等)。在一些实施例中,“样本”或“生物样本”的来源还可以是肿瘤。“样本”或“生物样本”可以在从来源获得后直接使用,或者在对其进行预处理(例如,固定、嵌入介质、切片、清洗、富集)后使用。在一些实施例中,还需提取“样本”或“生物样本”中的核酸(例如 RNA 多核苷酸)。

[0073] 如本文所使用,“测试样本”是指包含期望被测序的核酸的样本或生物样本。在一些实施例中,所述核酸优选为 RNA 多核苷酸。在一些实施例中,所述测试样本来源于生物。在一些实施例中,所述测试样本还可以来源于环境,例如土壤、湖、河等。

附图说明

[0074] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍。在所有附图中,类似的元件或部分一般由类似的附图标记标识。附图中,各元件或部分并不一定按照实际的比例绘制。显而易见地,下面描述中的附图是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性

的前提下，还可以根据这些附图获得其它的附图。

[0075] 图 1 示出了 DecodeR 的示意图、示例性流程和算法比较；

[0076] 图 2 示出了本发明的 DecodeR 与现有技术的分类器的比较；

[0077] 图 3 示出了不同条形码的比较；

[0078] 图 4 示出了本发明的 DecodeR 在真实样本中的性能效果；

[0079] 图 5 示出了本发明的 DecodeR 检测来自病毒样本 RNA；

[0080] 图 6 示出了本发明的 DecodeR 对肿瘤样本的突变和 m6A 分析；

[0081] 图 7 示出了本发明提供的一种直接 RNA 纳米孔测序方法的流程图；

[0082] 图 8 示出了本发明提供的一种多条形码直接 RNA 纳米孔测序分类系统的模块示意图。

具体实施方式

[0083] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述。显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0084] 本文中，使用用于表示元件的诸如“模块”、“部件”或“单元”的后缀仅为了有利于本发明的说明，其本身没有特定的意义。因此，“模块”、“部件”或“单元”可以混合地使用。

[0085] 本文中，术语“上”、“下”、“内”、“外”“前”、“后”、“一端”、“另一端”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性。

[0086] 本文中“和/或”包括任何和所有一个或多个列出的相关项的组合。

[0087] 本文中“多个”意指两个或两个以上，即其包含两个、三个、四个、五个等。

[0088] 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0089] 如在本说明书中使用的，术语“大约”，典型地表示为所述值的 $\pm 5\%$ ，更典型的是所述值的 $\pm 4\%$ ，更典型的是所述值的 $\pm 3\%$ ，更典型的是所述值的 $\pm 2\%$ ，甚至更典型的是所述值的 $\pm 1\%$ ，甚至更典型的是所述值的 $\pm 0.5\%$ 。

[0090] 在本说明书中，某些实施方式可能以一种处于某个范围的格式公开。应该理解，这种“处于某个范围”的描述仅仅是为了方便和简洁，且不应该被解释为对所公开范围的僵化限制。因此，范围的描述应该被认为是已经具体地公开了所有可能的子范围以及在此范围内的独立数字值。例如，范围 1~6 的描述应该被看作已经具体地公开了子范围如从 1 到 3，从 1 到 4，从 1 到 5，从 2 到 4，从 2 到 6，从 3 到 6 等，以及此范围内的单独数字，例如 1，2，3，4，5 和 6。无论该范围的广度如何，均适用以上规则。

[0091] 实施例一

[0092] 1.1 条形码设计及合成

[0093] 如表 1 所示，本实施例纳入本发明设计的 40 个新的逆转录接头（Reverse Transcript Adapter, RTA）的条形码序列（RTA-01 至 RTA-40，长度在 20-28bp 之间），以及来自 ONT 官方的 RTA 的条形码序列（RTA-00）、DeePlexiCon 的 3 个 RTA 的条形码序列（RTA-41、RTA-42 和 RTA-43）和来自 Poreplex 的 4 个 RTA 的条形码序列（RTA-44、RTA-45、RTA-46 和 RTA-47），共计 48 个 RTA 的条形码序列。

[0094] 本发明所基于的测序结构如图 1a 所示。每一对 RTA 由前向链 (Forward strands) 和后向链 (Reverse strands) 组成。按照 5'端至 3'端的方向, 前向链序列包括条形码序列 (X1) 以及接在条形码后端的第一连接序列 (X2, 也可以称为第一不匹配多核苷酸序列), 后向链包括第二连接序列 (X2, 也可以称为第二不匹配多核苷酸序列)、条形码序列 (X1) 以及接在条形码后端的 Poly(T)序列 (例如 10-30 个重复 T 序列)。将 RTA 的前向链和后向链序列进行杂交并组装成双链结构。根据碱基互补配对原则, RTA 包括条形码区域 (X1)、不匹配多核苷酸区域 (X2) 以及突出末端的 Poly(T)序列。接着, 将组装好的双链结构捕获并杂交带有 Poly(A)序列的待测 RNA 序列, 经逆转录形成 RNA/DNA 杂化双链结构。然后, 再将 RNA/DNA 杂化双链结构与 ONT 测序接头 (X3) 连接, 最终形成本发明基于的完整的测序结构。在一些实施例中, 所述第一不匹配多核苷酸序列包括 TAGTAGGTTC (SEQ ID NO:49), 第二不匹配多核苷酸序列包括 GAGGCGAGCGGTCAATTTT (SEQ ID NO:50)。

[0095] 接下来, 本实施例使用 OligoAnalyzer™工具评估了本发明所形成的 DNA 寡核苷酸的 GC 含量、熔点和二级结构。本发明设计的条形码的长度被控制 20-28 nt, 避免了超长的条形码可能会干扰磁珠纯化步骤中去除过量的 RTA 的问题。

[0096] 表 1 自定义 RTA 序列

RTA ID	来源	条形码序列	SEQ ID NO:
RTA-00	DeePlexiCon-BC1(D-BC1)	GGCTTCTTCTTGCTCTTAGG	25
RTA-01	NewDesign01	CCTGGTAAGTGGGACACAAGACTC	1
RTA-02	NewDesign02	ACGTAACTTGGTTTGTTCCTGAA	2
RTA-03	NewDesign03	GAGAGGACAAAGGTTTCAACGCTT	3
RTA-04	NewDesign04	TCCATTCCCTCCGATAGATGAAAC	26
RTA-05	NewDesign05	CACACACCGACAACCTTCTT	4
RTA-06	NewDesign06	TAGGGACGATAGAATCCGAA	27
RTA-07	NewDesign07	AAGGTTACACACTGGACAAG	28
RTA-08	NewDesign08	AACCCTCGCTGTGCCTAGTT	5
RTA-09	NewDesign09	CGAGGAGGTTCACTGGGTAG	6
RTA-10	NewDesign10	CTAACCCATCATGCAGAAGC	7
RTA-11	NewDesign11	AAGAGTTGTCTGGTGTCTTAATGTG	29
RTA-12	NewDesign12	TTCGGATTCTATTCTCGTGTCTA	8
RTA-13	NewDesign13	TTCTCGCAAAGGCAGTAAGAAGTC	30
RTA-14	NewDesign14	AAGCGTCTTTGTCTGAAACCTCTC	9
RTA-15	NewDesign15	AGAACCATACTCCGACTTGTGTGA	10
RTA-16	NewDesign16	AGGTCTACCTAACACTCGCCACTG	31
RTA-17	NewDesign17	ACCCTCCAGAAGTACCTCTGAT	11
RTA-18	NewDesign18	TTGCGTCCCTGTTACGAGAACAT	32
RTA-19	NewDesign19	CATACCGACTACGCATTCTCAT	12
RTA-20	NewDesign20	CTATACGAAGCTGAGGGACTGC	13
RTA-21	NewDesign21	TAGTGGATGACCAAGGATAGCC	14
RTA-22	NewDesign22	TCAGTGAGGATCTACTTCGCCA	15
RTA-23	NewDesign23	CATACCTGGAACGTGGTACACCTGTA	16
RTA-24	NewDesign24	GCTAGGTCAATCGATCCTTCGGAAGT	17
RTA-25	NewDesign25	GGAGTGTCCAGAGAAGTACACG	33
RTA-26	NewDesign26	AGTAGAAGGTTCTTCCCACTC	34
RTA-27	NewDesign27	GATCACAGAGATGCCTTCAGTG	18
RTA-28	NewDesign28	AGATTACAGACCAAGTCTCATGCAAAG	35

RTA-29	NewDesign29	TGGAAGATGAGACATCCTGATCTACG	19
RTA-30	NewDesign30	TCCTACTCACGACAGGTGGCATGAA	20
RTA-31	NewDesign31	CAGGTTACTCCTCTACGTGAGTCTGA	36
RTA-32	NewDesign32	CATGTTCAACGGCAAGGCTTCTATGG	37
RTA-33	NewDesign33	CACCCACACTTACGCTTCAGGACGTA	21
RTA-34	NewDesign34	TTAGCTGAAGTTCCTGGGTCTTGAAC	38
RTA-35	NewDesign35	ATGCTTGTTACATCACAGAACCCTGGAC	22
RTA-36	NewDesign36	AACAACCGAACCTTACAGTGAATCAGAA	39
RTA-37	NewDesign37	CGGATGAACATAGGACAGTTAGCGATTC	40
RTA-38	NewDesign38	TCTTCTACTACCGACAGATCCGAAGCAG	41
RTA-39	NewDesign39	TTGTGAGTGGAACAGAGATACAGGACC	42
RTA-40	NewDesign40	CCTGTCTGGAAGAAGAACTTTGGACTT	43
RTA-41	DeePlexiCon-BC2(D-BC2)	GTGATTCTCGTCTTTCTGCG	44
RTA-42	DeePlexiCon-BC3(D-BC3)	GTACTTTTCTCTTTGCGCGG	45
RTA-43	DeePlexiCon-BC4(D-BC4)	GGTCTTCGCTCGGTCTTATT	46
RTA-44	Poreplex-BC1(P-BC1)	CCTCCCCTAAAAACGAGCCGCATTTGCG	47
RTA-45	Poreplex-BC2(P-BC2)	CCTCGTCGGTTCTAGGCATCGCGTATGC	23
RTA-46	Poreplex-BC3(P-BC3)	CCTCCCCTTTTACACGCACTAACCAGG	48
RTA-47	Poreplex-BC4(P-BC4)	CCTCCTTCAGAAGAGGGTCGCTTCTACC	24

[0097] 1.2 体外转录 (IVT) RNA 的制备

[0098] 为了测试条形码并进行模型训练,本实施例生成了 52 个体外转录(in-vitro transcribed, IVT) RNA, 其序列来自 27 个小鼠基因、22 个人类基因和 3 个 SARS-CoV-2 基因(表 2)。针对这些基因序列,使用 Primer3web 工具 (Version4.1.0) 设计了前向引物和后向引物,并在前向引物中添加了 T7 RNA 聚合酶 (T7 RNA Polymerase) 启动子序列,在反向引物中添加了 15 个重复 T 序列 (oligo(dT)15) 序列。

[0099] 使用逆转录酶,将人类和小鼠总 RNA 以及 SARS-CoV-2 基因组 RNA 逆转录为第一链 cDNA。使用上述设计的引物,对不同的 DNA 片段进行 PCR 扩增,使用琼脂糖凝胶电泳确定 PCR 产物的质量。使用 1 μ g 纯化的 PCR 产物作为模板,通过 T7 RNA 聚合酶生成 IVTRNA,并使用 RNA Clean & Concentrator-5 按照制造商的建议纯化。使用 Qsep100 确定 IVTRNA 的长度,并使用 NanoDrop 2000 测量每个 IVTRNA 产物的浓度和纯度。

[0100] 表 2 体外转录 RNA 序列信息

基因名称	RNA 名称	长度
hActB	NM_001101.5 Homo sapiens actin beta (ACTB), mRNA 499 1011	513
hB2M	NM_004048.4 Homo sapiens beta-2-microglobulin (B2M), mRNA 57 636	580
hCYC1	NM_001916.5 Homo sapiens cytochrome c1 (CYC1), mRNA; nuclear gene for mitochondrial product 615 1118	504
hDDX5	NM_004396.5 Homo sapiens DEAD-box helicase 5 (DDX5), transcript variant 2, mRNA 1004 1632	629
hDDX5-2	NM_004396.5 Homo sapiens DEAD-box helicase 5 (DDX5), transcript variant 2, mRNA 434 1023	590
hEEF1A1	NM_001402.6 Homo sapiens eukaryotic translation elongation factor 1 alpha 1 (EEF1A1), mRNA 612 1127	515
hEEF1A1-2	NM_001402.6 Homo sapiens eukaryotic translation elongation factor 1 alpha 1 (EEF1A1), mRNA 2671 3103	433
hGAPDH	NM_001357943.2 Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH),	177

	transcript variant 7, mRNA 476 652	
hGUSB	NM_000181.3 Homo sapiens glucuronidase beta (GUSB), transcript variant 1, mRNA 1281 1928	648
hH3F3A	NM_002107.7 Homo sapiens H3.3 histone A (H3-3A), transcript variant 1, mRNA 3 506	504
hHMBS	NM_000190.4 Homo sapiens hydroxymethylbilane synthase (HMBS), transcript variant 1, mRNA 560 1059	500
hHUWE1	NM_031407.7 Homo sapiens HECT, UBA and WWE domain containing E3 ubiquitin protein ligase 1 (HUWE1), mRNA 3740 4426	687
hPOLR2A	NM_000937.3 Homo sapiens polymerase (RNA) II (DNA directed) polypeptide A, 220kDa (POLR2A), mRNA 551 1257	707
hPOLR2A-4	NM_000937.3 Homo sapiens polymerase (RNA) II (DNA directed) polypeptide A, 220kDa (POLR2A), mRNA 2534 3166	633
hPRT1	NM_000194.3 Homo sapiens hypoxanthine phosphoribosyltransferase 1 (HPRT1), mRNA 104 667	564
hRPL13-3	NM_000977.3 Homo sapiens ribosomal protein L13 (RPL13), transcript variant 1, mRNA 3174 3668	495
hRPL13A	NM_012423.4 Homo sapiens ribosomal protein L13a (RPL13A), transcript variant 1, mRNA 464 930	467
hRPL29	NM_000992.3 Homo sapiens ribosomal protein L29 (RPL29), mRNA 163 721	559
hRPS23	NM_001025.4 Homo sapiens ribosomal protein S23 (RPS23), mRNA 1711 2285	575
hSDHA	NM_004168.2 Homo sapiens succinate dehydrogenase complex, subunit A, flavoprotein (Fp) (SDHA), mRNA 231 857	627
hTBP	NM_003194.5 Homo sapiens TATA-box binding protein (TBP), transcript variant 1, mRNA 144 668	525
hTBP-2	NM_003194.5 Homo sapiens TATA-box binding protein (TBP), transcript variant 1, mRNA 1018 1748	731
hYWHAZ-2	NM_001135702.1 Homo sapiens tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta (YWHAZ), transcript variant 6, mRNA 2090 2506	417
mActB	XM_030254057.1 PREDICTED: Mus musculus actin, beta (Actb), transcript variant X1, mRNA 1399 2095	697
mB2M	NM_009735.3 Mus musculus beta-2 microglobulin (B2m), mRNA 156 829	673
mEef2	NM_007907.2 Mus musculus eukaryotic translation elongation factor 2 (Eef2), mRNA 533 1083	551
mEef2-2	NM_007907.2 Mus musculus eukaryotic translation elongation factor 2 (Eef2), mRNA 1864 2546	683
mEef2-3	NM_007907.2 Mus musculus eukaryotic translation elongation factor 2 (Eef2), mRNA 1401 1883	483
mGAPDH	XM_017321385.2 PREDICTED: Mus musculus glyceraldehyde-3-phosphate dehydrogenase (Gapdh), transcript variant X1, mRNA 826 1373	548
mGUSB-2	NM_010368.2 Mus musculus glucuronidase, beta (Gusb), transcript variant 1, mRNA 1450 1851	402
mGUSB-3	NM_010368.2 Mus musculus glucuronidase, beta (Gusb), transcript variant 1, mRNA 105 529	425
mHPRT-1	NM_013556.2 Mus musculus hypoxanthine guanine phosphoribosyl transferase (HpRT), mRNA 323 826	504
mPPia	NM_008907.2 Mus musculus peptidylprolyl isomerase A (Ppia), mRNA 49 623	574

mPpib	NM_011149.2 Mus musculus peptidylprolyl isomerase B (Ppib), mRNA 160 759	600
mRer1-2	NM_026395.2 Mus musculus retention in endoplasmic reticulum sorting receptor 1 (Rer1), transcript variant 1, mRNA 165 738	574
mRn18S	NR_046233.2 Mus musculus 45S pre-ribosomal RNA (Rn45s), ribosomal RNA 4173 4783	610
mRn18S-2	NR_003278.3 Mus musculus 18S ribosomal RNA (Rn18s), ribosomal RNA 939 1516	578
mRplp0	NM_007475.5 Mus musculus ribosomal protein, large, P0 (Rplp0), mRNA 692 1100	409
mSDHA-2	BC011301.1 Mus musculus succinate dehydrogenase complex, subunit A, flavoprotein (Fp), mRNA 1697 2365	669
mSptbn	NM_175836.2 Mus musculus spectrin beta, non-erythrocytic 1 (Sptbn1), transcript variant 1, mRNA 1368 2115	748
mSptbn1-2	NM_175836.2 Mus musculus spectrin beta, non-erythrocytic 1 (Sptbn1), transcript variant 1, mRNA 5773 6481	709
mTgfb1	NM_011577.2 Mus musculus transforming growth factor, beta 1 (Tgfb1), mRNA 1212 1801	590
mTGFB1-2	NM_011577.2 Mus musculus transforming growth factor, beta 1 (Tgfb1), mRNA 711 1231	521
mTubala	XR_002381333.2 PREDICTED: Mus pahari tubulin alpha-1 chain-like (LOC110339099), misc_RNA 709 1180	471
mWrnip	NM_030215.3 Mus musculus Werner helicase interacting protein 1 (Wrnip1), mRNA 1408 2152	745
mWrnip1-2	NM_030215.3 Mus musculus Werner helicase interacting protein 1 (Wrnip1), mRNA 903 1408	506
mYwhaz	NM_001253807.1 Mus musculus tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide (Ywhaz), transcript variant 4, mRNA 593 1273	681
mYwhaz-2	NM_001253807.1 Mus musculus tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide (Ywhaz), transcript variant 4, mRNA 1254 1953	700
Nsp1	NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome 346 851	506
ORF1ab	NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome 13231 13825	595
mSptbn1-4	NM_175836.2 Mus musculus spectrin beta, non-erythrocytic 1 (Sptbn1), transcript variant 1, mRNA 3289 3783	495

[0101] 1.3 混样直接 RNA 测序建库流程

[0102] 自定义 RTA 的组装。将每个 RTA 的前向链 (1.54 μ M) 和后向链 (1.4 μ M) 分别混合在缓冲液中 (10mM Tris-HCl, pH 7.5, 50mM KCl) (表 3)。本实施例采用过量的前向链的原因在于, 过量的前向链可以确保后向链的完全消耗, 避免了后向链的存在可能会降低 RNA-RTA 的连接产物收率的问题。自定义 RTA 的组装遵循如下 PCR 程序进行: 95°C 变性 5 分钟, 65°C、50°C、37°C、22°C 分别退火 30 分钟, 4°C 保持 (表 4)。最后, 通过 4% 琼脂糖凝胶电泳确认 RTA 是否完成组装。该组装产物能作为带有不同条形码的 RTA, 用于本发明的混样直接 RNA 测序方案。

[0103] 表 3 自定义 RTA 的组装

试剂	储液浓度	终浓度	用量
前向链 (For)	10uM	1.54uM	2.31uL
后向链 (Rev)	10uM	1.40uM	2.1uL
Tris-HCL	50mM	10mM	3 μ L
KCL	2M	50mM	0.375 μ L
H2O	-	-	7.215 μ L

[0104] 表 4 PCR 程序设计

步骤	温度	时间	循环数
预变性	95°C	5min	1
	65°C	30min	
	50°C	30min	
退火	37°C	30min	1
	37°C	30min	
	22°C	30min	
保存	4°C	∞	1

[0105] 混样直接 RNA 测序文库制备。使用 NEBNext 快速连接模块 (NEB, E6056), 将每个带有 Poly(A)序列的 RNA 样本单独与预退火的自定义 RTA 进行杂交和连接, 并且记录其一一对应的关系。并使用 SuperScriptIV 逆转录酶 (ThermoFisher, 18090200) 将其逆转录, 形成 RNA/DNA 稳定双链并破坏 RNA 的二级结构。使用 1.8XRNA Clean XPbeads (Beckman, A63987) 对产物进行纯化。然后, 将每个反应的逆转录产物按相同的摩尔质量进行混合, 并在室温下将 RNA 适配体混合物 (RMX) 连接到 RNA/DNA 杂交体的 3'端。连接后, 将反应混合物使用 1XRNA Clean XPbeads 进行纯化 (Beckman, A63987), 用缓冲液 (WSB) 洗涤两次, 并用洗脱缓冲液 (EB) 洗脱产物。接下来, 将遵从制造商的说明书 (ONT, SQK-RNA002) 构建上机混合液, 根据 MinION 或 PromethION 测序仪适配的文库体积进行调整。接下来, 将上机混合液加载至质检合格的 ONT 测序芯片中, 设置 MinKNOW 软件中的测序参数, 测序持续 48h 以上以获取足够的测序数据。

[0106] 1.4 DecodeR 机器学习模型的训练

[0107] 首先, 使用了 48 种 RTA 连接 IVTRNA 的各种组合进行了 5 次独立的实验, 获得了 5 次 DRS 原始电信号数据集, 在碱基调用 (Base-calling) 和与参考序列比对后, 生成了总共 6,276,168 个有效 Reads (范围从 623K 到 2.30M)。质量控制后, 由于生成的读取 (Reads) 数量较少, 24 个 RTA 条形码被过滤。

[0108] 接下来, 将 IVTRNA 下机的原始信号数据 (FAST5) 使用 Guppy (v3.1.5) 在高精度模型下进行碱基调用。使用 minimap2 (v2.17-r941) 将碱基调用 Reads (FASTQ) 与 IVTRNA 参考序列比对。接下来, 以比对质量 (MAPQ) 大于 60, 比对至唯一参考序列为条件过滤比对 Reads, 并根据它们比对的连接序列将其分组。使用 R 包 rhdf5 (<https://github.com/grimbough/rhdf5>) 从 FAST5 文件中提取原始信号。还可以使用 R 包 smother (<https://CRAN.R-project.org/package=smother>) 使得原始电流信号更平滑, 以减少原始数据的噪声。

[0109] 鉴于 Poly(A)电流信号是连续的碱基 A 信号, 本实施例首先抛光 (polish) 和归一化 (normalize) 电流信号, 并从 RNA 3'端的信号的均值和方差中确定 Poly(A)的位置。于是, 获得了 ONT 测序接头 (也可以理解为“测序接头”、“适配体”、“适配器”、“RNA 适配体混合物”等)、RTA 条形码、poly(A)和体外转录 RNA 序列产生的一系列电流信号 (图 1b)。由于每个读取的测序接头序列是相同的, 因此在分析时没有删除 ONT 测序接头的信号。然后, 使用 changepoint 包 (<https://CRAN.R-project.org/package=changepoint>) 中的 cpt.meanvar 函数, 根据均值和方差对 Poly(A)前面的 RTA 条形码的电流信号进行分割, 每个条形码信号被分成 70-100 段, 优选为 100 段 (图 1c)。

[0110] 以分割成 100 列矩阵为例, 将所有条形码生成的 100 列矩阵最终用于模型训练。R 包 caret (<https://CRAN.R-project.org/package=caret>) 的实施是为了简化基于随机森林分类模型生成预测模型的过程。为避免过拟合, 本实施例采用重复 K 折交叉验证 (10 折, 重复 10 次) 的方法进行重采样。为了提高模型训练的计算效率, 采用了 R 包中 doParallel 包支持的并行处理框架。由于每个条形码的序列读取数不同, 因此对于每个分类器, 本实施例首先使用 caret 包中的 downSample 函数随机抽取一个数据集, 以便所有分类组与少数类具有相同的频率。然后, 将这些 Reads 的 70%用于训练集, 其余所有 Reads 用作测试集。本实施例训练了 2、4、

6、8、10、12、14、16、18、20、24 种条形码组合的模型。应当理解，还可以训练其他种类条形码组合（例如 3 种、5 种、11 种、19 种、22 种等）来产生更多的模型。

[0111] 综上所述，如图 7 所示，本发明提供的多重混样直接 RNA 纳米孔测序方法包括：首先，针对每一种测试样本 RNA 多核苷酸，将 RTA 的前向链和后向链序列进行杂交并组装成双链结构（即双链 DNA 多核苷酸）。将组装好的双链结构捕获并杂交带有 Poly(A) 序列的待测 RNA 序列，形成第一杂交结构（S1）。第一杂交结构经逆转录形成 RNA/DNA 杂化双链结构（即第二杂交结构）（S2）。再将 RNA/DNA 杂化双链结构与 ONT 测序接头连接，形成第三杂交结构（S3）。将第三杂交结构从带有马达蛋白的 3' 端序列开始测序，获得原始测序数据（S4），进而获得包括 ONT 测序接头、条形码、Poly(A) 和测试样本 RNA 多核苷酸产生的一系列原始电流信号并可选地对其进行处理（例如提取、降噪、归一化）（S5）。然后，对条形码的电流信号数据进行分割，选择基于特定数量条形码的训练好的分类模型对电流信号数据进行预测，可选地选择恰当的过滤阈值（ >0.3 ），最终获得每条电信号的所属条形码和分类概率（S6）。

[0112] 基于上述方法，本发明还提供了一种多条形码直接 RNA 纳米孔测序系统 100。

[0113] 下面说明本发明的测序系统的示例性结构，在一些实施例中，如图 8 所示，所述测序系统 100 可以包括：

[0114] 测序数据接收模块 102，用于接收和存储测序数据，所述测序数据包括双链多核苷酸的测序数据，所述双链多核苷酸包括第一链和第二链，所述第一链从 5' 至 3' 方向包括 RNA 多核苷酸、Poly(A) 序列、条形码序列、第一连接序列和测序接头，所述第二链从 5' 至 3' 方向包括测序接头、第二连接序列、条形码序列、Poly(T) 序列和所述 RNA 多核苷酸的逆转录序列；所述测序数据包括测试样本数据和训练样本数据。

[0115] 在一些实施例中，所述测试样本数据包括 n 种测试样本 RNA 多核苷酸对应的测序数据，其中 $2 \leq n \leq 24$ 且 n 为整数，所述条形码序列对每一种测试样本 RNA 多核苷酸具有特异性，所述条形码序列选自 SEQ ID NOs:1-24 中的任意一种。

[0116] 在一些实施例中，所述测试样本 RNA 多核苷酸的来源包括病原体、植物、浮游动物、昆虫、哺乳动物和肿瘤中的一种或多种。在一些实施例中，所述病原体包括 RNA 病毒、细菌、真菌和寄生虫中的一种或多种。

[0117] 信号处理模块 104，用于将所述测序数据处理成电流信号数据。

[0118] 在一些实施例中，所述信号处理模块 104 对所述测序数据进行预处理和分割。

[0119] 在一些实施例中，所述预处理包括降噪处理、抛光处理和归一化处理中的一种或多种。

[0120] 在一些实施例中，所述分割为将所述条形码序列的电流信号数据分割成 70-100 列矩阵。

[0121] 在一些实施例中，所述分割为将所述条形码序列的电流信号数据分割成 100 列矩阵。

[0122] 分类模块 106，用于调用基于所述测试样本数据中的所述条形码序列构建的分类模型，对所述测试样本数据的电流信号数据进行分类并输出所述 n 种测试样本 RNA 多核苷酸的分类结果。

[0123] 在一些实施例中，所述测序系统 100 进一步包括模型训练模块 108，所述模型训练模块 108 利用机器学习算法对所述训练样本数据的电流信号数据进行训练学习，从而确定所述分类模型。

[0124] 在一些实施例中，所述机器学习算法包括 K 最近邻、神经网络、朴素贝叶斯分类、回归树、AdaBOOST 和随机森林中的一种或多种。

[0125] 在一些实施例中，所述机器学习算法包括随机森林。

[0126] 在一些实施例中，所述训练样本数据包括体外转录 RNA 的测序数据。

[0127] 在一些实施例中，所述训练样本数据中用作训练集和测试集的比例为 7:3。

[0128] 在一些实施例中，所述体外转录 RNA 包括表 2 所示的基因中的一种或多种。

[0129] 在一些实施例中，所述模型训练模块 108 用于将所述训练样本数据的电流信号数据与对应的参考序列比对，以比对质量大于 60、比对至唯一参考序列为条件将不满足条件的读取序列过滤。

[0130] 在一些实施例中，所述测序系统 100 进一步包括过滤模块 110，用于基于设置的分类可能性的阈值，将低于所述阈值的读取过滤掉。

[0131] 在一些实施例中，所述阈值>0.3。

[0132] 在一些实施例中，所述过滤模块 110 和所述分类模块 106 可以整合为一个分析模块 112。

[0133] 1.5 DecodeR 的条形码识别分类流程

[0134] 在一些实施例中，DecodeR 的具体工作流程如下：

[0135] 根据建库实验的条形码组合类型，选择不同的经过训练的模型，输入至 model 参数；

[0136] 选择 MinKNOW 软件保存并过滤后的原始电信号数据，一般在 fast5_pass 路径下，输入至 fast5 参数；

[0137] 选择 0-1 范围内值作为阈值，对分类的标准进行过滤，输入至 cutoff 参数；

[0138] 选择运行的 CPU 数目，不超过运算机器最大核心数，输入至 NT 参数。

[0139] 在一些实施例中，DecodeR 的默认使用方法如下：DecodeR (fast5, model, NT = 1, cutoff = 0, include.lowest = FALSE)，输出值将记录每条 ReadID 以及其所属的条形码信息和分类的可能性 (probability, 0-1)，如表 5 所示。

[0140] 表 5 DecodeR 输出结果示例

ReadID	条形码	可能性
read_0b3cabf5-44b4-4438-86df-4dfa672000e1	RTA-33	1
read_0d11e3c0-00ca-4531-a7fc-55fc75d3bb1f	RTA-33	1
read_10cefe5c-6441-4c80-ba98-dbd50c4d6cf3	RTA-33	1
read_12b81f83-f74f-4c45-b6b5-7e7168546ddb	RTA-33	1
read_160d5209-db4c-463c-9e1c-813e0d8e8737	RTA-35	0.994
read_180def1e-e4cb-4f8b-85de-100e7fd584f9	RTA-33	0.958

[0141] 根据分类可能性过滤掉分类不太可靠的 ReadID，并以 ReadID 为条件分类原始 Fast5 信号和 Fastq 序列。

[0142] 实施例二

[0143] 2.1 DecodeR 算法与其他算法的比较

[0144] 使用验证集评估了 6 种不同分类算法的准确性，包括 K 最近邻 (K-Nearest Neighbour, KNN)、神经网络 (Neural Network, NNET)、朴素贝叶斯 (Naive Bayes, NB) 分类、回归树 (Classification And Regression Tree, CART)、AdaBOOST 和随机森林 (Random Forest, RF) (表 6)。尽管 6 种方法都实现了接收操作特征曲线下的面积 (Area Under The Receiver Operating Characteristic Curve, AUROC) 高于 0.92 (图 1d)，并且相关的精确召回曲线下的面积 (Area Under The Precision-Recall Curve, AUPRC) 大于 0.83 (图 1e)，但随机森林算法优于其他方法，在 6 种方法中实现了最高的 AUROC (0.9961) 和最高的 AUPRC (0.9906) (图 1d、图 1e)。此外，随机森林算法的精确度 (accuracy)、灵敏度 (sensitivity)、特异性 (specificity)、准确性 (precision)、召回率 (recall) 和 F1 分数 (F1 score) 均显著高于其他五种方法 (图 1f, 表 6)。因此，本发明选择基于随机森林算法建立测序方法 (简称为 DecodeR) (图 1g)，用于进一步的训练、测试和应用，分别针对 RNA 病毒、细菌、真菌、寄生虫和肿瘤样本的 RNA 进行了性能评估和实施例分析 (图 1h)。

[0145] 表 6 多种算法的比较

模型	精确度	灵敏度	特异性	准确性	召回率	F1 分数
RF	0.955675	0.955675	0.985225	0.955793	0.955675	0.955695

NB	0.8353	0.8353	0.9451	0.838332	0.8353	0.83612
NNET	0.83465	0.83465	0.944883	0.834675	0.83465	0.834468
KNN	0.828975	0.828975	0.942992	0.842567	0.828975	0.829393
CART	0.92825	0.92825	0.976083	0.928364	0.92825	0.928274
AdaBOOST	0.933375	0.933375	0.977792	0.933412	0.933375	0.933383

[0146] 2.2 DecodeR 与其他条形码解复用方法的比较

[0147] 目前,已经有两种直接 RNA 测序 (DRS) 条形码解复用 (demultiplex) 计算方法,分别为基于残差神经网络分类器算法的图像识别的 DeePlexiCon (<https://github.com/Psy-Fer/deeplexicon>) 和尚未正式发表的 Poreplex (<https://github.com/hyeshik/poreplex>)。

[0148] 为了比较 DecodeR 与其他方法的性能,本实施例评估了每种方法的拆分精度和运行时间。首先,分别对 DeePlexiCon 和 Poreplex 的 8 种条形码连接 IVTRNA 进行了 5 次独立的实验,获得了 5 次 DRS 原始电信号数据集 (表 7)。接着,使用 DecodeR 对其中 3 次数据 (训练集) 进行训练模型,并使用该模型预测另外 2 次数据 (验证集)。类似地,使用 DeePlexiCon 和 Poreplex 对相同的验证集进行条形码拆分的预测。接收操作特征 (Receiver Operating Characteristic, ROC) 分析,曲线下面积 (Area Under Curve, AUC) 指标和精度召回 (recall) 曲线由 R 包 multiROC (<https://CRAN.R-project.org/package=multiROC>) 产生。CPU 时间由 GNUtime 命令评估,并将用户时间和系统时间相加。

[0149] 表 7 解复用方法比较数据集

条形码	RTA ID	IVTRNA	批次	唯一比对的 Reads	应用
D-BC1	RTA-00(ONT)	hGAPDH	1	16349	验证
D-BC2	RTA-41	mActB	1	3634	验证
D-BC3	RTA-42	ORF1ab	1	432000	验证
D-BC4	RTA-43	NSP1	1	133636	验证
D-BC1	RTA-00(ONT)	hGAPDH	2	57410	训练
D-BC2	RTA-41	mActB	2	2901	训练
D-BC3	RTA-42	ORF1ab	2	64469	训练
D-BC4	RTA-43	NSP1	2	38038	训练
P-BC1	RTA-44	mEef2	2	99137	训练
P-BC2	RTA-45	mGapdH	2	74356	训练
P-BC3	RTA-46	mRplp0	2	21677	训练
P-BC4	RTA-47	mTubala	2	58597	训练
D-BC1	RTA-00(ONT)	hGAPDH	3	2554	训练
D-BC2	RTA-41	mActB/hRPL13-3	3	1479	训练
D-BC3	RTA-42	ORF1ab	3	3947	训练
D-BC4	RTA-43	NSP1	3	2777	训练
P-BC1	RTA-44	mEef2	3	25004	训练
P-BC2	RTA-45	mGapdH	3	88221	训练
P-BC3	RTA-46	mRplp0	3	21655	训练
P-BC4	RTA-47	mTubala	3	35216	训练

D-BC1	RTA-00(ONT)	hGAPDH	4	12833	训练
D-BC2	RTA-41	hRPL13-3	4	5119	训练
D-BC3	RTA-42	ORF1ab	4	18938	训练
P-BC1	RTA-44	mTubala	4	11389	训练
P-BC2	RTA-45	mGapdH	4	32967	训练
P-BC3	RTA-46	mRplp0	4	8157	训练
P-BC4	RTA-47	mEcf2	4	48263	训练
P-BC1	RTA-44	hGAPDH	5	167924	验证
P-BC2	RTA-45	mActB	5	1563	验证
P-BC3	RTA-46	ORF1ab	5	429767	验证
P-BC4	RTA-47	NSP1	5	306156	验证

[0150] 与 DeePlexiCon、Poreplex 两种方法相比，DecodeR 在解复中使用 Poreplex 预设的 4 种条形码中的分类准确度范围为 95.82 至 97.83% (96.98 ± 0.84)，而 Poreplex 的分类准确度范围为 83.4 到 93.18% (89.4 ± 4.21)（图 2a，unknown 表示未鉴定），而且准确性、召回率和 F1-分数均高于 Poreplex 方法。同时，DecodeR 在解复用 DeePlexiCon 条形码中的分类准确率范围为 90.57 至 94.43% (92.29 ± 1.60)，相比之下，DeePlexiCon 的分类准确率范围为 51.33 至 85.22% (68.39 ± 15.16)（图 2b，unknown 表示未鉴定）。此外，与四个条形码的多路分解相比，DecodeR 实现了显著更高的灵敏度、召回率和 F1 得分（图 2b，表 8）。DecodeR 获得了比 DeePlexiCon 的 0.946 更高的准确度 AUC 值 0.997，与 0.888 相比更高的精度召回 AUC 值 0.993，以及显著更高的分类精确度（Accuracy）和恢复率（图 2c、图 2d 和图 2e，表 8）。最后，为了评估解复用的效率，还计算了 DecodeR 和 DeePlexiCon 在处理相同数量（1-150K）的 DRS 读取时的 CPU 时间，发现 DecodeR 的运行速度比 DeePlexiCon 快 10 倍（图 2f）。

[0151] 表 8 DecdoeR 与其他解复用方法比较

方法	测试精确度	灵敏度	特异性	准确性	召回率	F1 分数
DecodeR（使用 DeePlexiCon 的 RTA）	0.9639	0.8048	0.9865	0.8226	0.8048	0.7888
DeePlexiCon	0.7907	0.7051	0.9330	0.5061	0.7051	0.5212
DecodeR（使用 Poreplex 的 RTA）	0.9696	0.9699	0.9903	0.9514	0.9699	0.9598
Poreplex	0.8986	0.8940	0.9988	0.8708	0.8940	0.8642

[0152] 2.3 DecodeR 的性能效果

[0153] 为了评估 DecodeR 在解复用（demultiplex）直接 RNA 测序方面的性能，使用 IVT-RNA 测试了它在解复用 2、4、6、8、10 和多达 24 个条形码时灵敏度、特异性、召回率等指标，解复用的全局精确度从 98.4% 下降到 92.2%（图 3c），而特异性的所有 AUC 值都大于 0.99（图 3a、图 3b，表 9）。精确度和恢复率虽然随着条形码数量从 2 个增加到 24 个有所下降，但始终保持在较高的水平（图 3c、图 3d）。

[0154] 表 9 不同条形码下 DecodeR 的性能表现

预测的条形码数	训练精确度	测试精确度	灵敏度	特异性	准确性	召回率	F1 分数
2	0.9842	0.9842	0.9850	0.9831	0.9866	0.9850	0.9858
4	0.9708	0.9712	0.9712	0.9707	0.9917	0.9877	0.9902
6	0.9644	0.9647	0.9649	0.9662	0.9690	0.9641	0.9940
8	0.9623	0.9624	0.9628	0.9723	0.9541	0.9589	0.9617

10	0.9553	0.9558	0.9559	0.9701	0.9482	0.9488	0.9523
12	0.9514	0.9524	0.9526	0.9394	0.9471	0.9420	0.9412
14	0.9461	0.9475	0.9472	0.9344	0.9409	0.9462	0.9463
16	0.9392	0.9418	0.9409	0.9303	0.9319	0.9295	0.9365
18	0.9112	0.9129	0.9132	0.8705	0.9007	0.9014	0.9114
20	0.9047	0.9085	0.9070	0.8713	0.8973	0.9068	0.8994
22	0.9018	0.9056	0.9030	0.8656	0.8949	0.8834	0.8945
24	0.9198	0.9217	0.9220	0.8935	0.9415	0.9140	0.9100

[0155] 鉴于 DecodeR 在多路分解 IVT-RNA 的 DRS 数据方面的理想性能,接下来将 DecodeR 应用于从病原体分离的真实的 RNA 样本,包括 2 种 RNA 病毒: Seneca Valley 病毒 (SVV) 和猪繁殖与呼吸综合征病毒 (PRRSV); 2 种细菌: 大肠杆菌 (E.coli) 和肠炎链球菌 (S.enteritidis); 1 种真菌: 酿酒酵母 (S.cerevisiae); 1 种寄生虫: 伯氏疟原虫 (P.berghei)。对不同的 RNA 样本进行了三次 DecodeR 建库实验和直接 RNA 测序,并生成了三批数据。ROC 分析显示,所有三次重复实验具有高 AUC 值的特异性 (0.999、0.998、0.976),灵敏度分别为 99.2%、93.3%、90.6%,特异性分别为 99.4%、98.4%、98.4%和最大精确度分别为 97.6%、95.9%、92.5% (图 4a,表 10)。DecodeR 在所有三批数据中也表现出高 AUC 值的精确召回率 (0.998、0.994、0.937) (图 4b)。在阈值 (cutoff) 和精确度 (accuracy) 方面,我们发现 cutoff 值越高,精确度越高,而召回率越低 (图 4c)。样本不平衡问题导致这些模型的精度值和 F1 分数不同,严重的样本不平衡问题与低精度和 F1 分数相关 (表 10)。

[0156] 表 10 真实样本下 DecodeR 的性能表现

批次	精确度	灵敏度(平均值)	特异性(平均值)	准确性(平均值)	召回率(平均值)	F1 分数(平均值)
1	0.9762	0.9920	0.9940	0.3011	0.9920	0.4447
2	0.9588	0.9334	0.9837	0.6494	0.9334	0.6772
3	0.9253	0.9063	0.9841	0.6134	0.9063	0.7473

[0157] 实施例三

[0158] 3.1 DecodeR 构建 RNA 病毒参考基因组

[0159] 为了评估 DecodeR 是否可以促进病毒基因组的构建,对三次 DecodeR 建库实验的 RNA 病毒数据进行了参考基因组比对和比对 Reads 长度的统计,发现 DecodeR 测得的 Reads 具有几乎完全覆盖 SVV 和 PRRSV 的参考基因组的能力 (图 5a、图 5b),这使得 RNA 病毒基因组的组装变得快速和容易。具体而言,724 条 Reads (9.68%,实验 1) 和 111 条 Reads (1.42%,实验 2) 接近完整的 SVV 基因组 (>7,000nt),42 条 Reads (2.26%) 接近完整的 PRRSV 基因组 (>15,000nt) (图 5b)。接下来分析了 SVV 基因组中的单碱基突变或单核苷酸多态性 (SNP),并在实验 1 和实验 2 中分别发现了 414 和 423 个突变,其中 411 个是 2 次实验中的共有突变 (图 5e),并且这些突变的突变比率具有高度可重复性 (图 5c)。举例来说,本实施例检测到了 SVV 病毒中 5700 位点 C 到 T 的突变,在两次实验中确实存在,且突变频率分别为 0.98 和 0.99 (图 5d)。DecodeR 也具有检测高频缺失 (deletion) 的能力,SVV 的缺失 (例如 4022-4024 位点的 3nt 缺失) 在两次实验中均被准确识别 (图 5d)。总体而言,DecodeR 可用于以低成本效益的方式检测病原体 RNA,并有可能应用于鉴定新型或突变型的 RNA 病毒。

[0160] 3.2 DecodeR 鉴定的肿瘤样本的突变和 m6A 修饰

[0161] 为了举例说明 DecodeR 在癌症中的潜在应用,对弥漫性中线胶质瘤-H3K27M 突变体 (DMG) 和胶质母细胞瘤 (GBM) 共 4 个样本进行了 DecodeR 混样建库实验和 RNA 直接测

序, 并获得了全转录组的突变、isoform 鉴定和 RNAm⁶A 修饰的概况 (图 6a)。在突变检测分析中, 共有 17,587 个单碱基突变在 DMG 和 GBM 不同样本中被 DecodeR 检测到, 并且其中 97-98% 的突变被直接 RNA 测序 (DRS) 得到验证 (图 6b)。与二代测序 (NGS) 和 Sanger 测序相比, DecodeR 在检测肿瘤突变方面表现出相似的准确性。在通过 Sanger 测序验证的突变位点中, DecodeR 和 NGS 之间检测到的共有突变的突变频率存在高度相关性 ($R=0.93$, Pearson 相关性, 图 6c)。举例来说, HLA-A 基因上的 chr6:29913430:C>T 突变位点在三个 DMG 样本中表现出较高的突变频率, 在 DMG1、DMG2 和 DMG3 中分别为 0.87、0.46 和 0.84, 而 GBM 样本中的突变频率为 0, 同时在 NGS 和 sanger 测序中也检测到突变位点, 突变频率与 DecodeR 相似 (图 6d)。相反的是, ATP6V0C 基因上的 chr16:2519933:C>A 突变位点在 GBM 中具有较高的突变频率, 而在 DMG 三个样本中几乎不突变, 同样得到了 NGS 和 sanger 测序的验证 (图 6e)。DecodeR 还鉴定了不同肿瘤类型中的保守缺失区域, 例如 SNORD137 基因上 chr9:12972603-12972607 位点具有 5nt 的缺失, 在所有肿瘤样本中保守出现 (图 6f)。

[0162] 在 m6A 修饰分析中, DecodeR 在 DMG 和 GBM 不同样本中共检测到 2,158 个高度置信的 m6A 位点, 其中 39% 得到 DRS 和 m6A Atlas 公共数据库的支持, 60% 仅通过 DRS 验证, 只有 1% 未验证 (图 6g)。DecodeR 在基因表达和 m6A 检测方面也显示出与 DRS 的高度一致性, 相关性分别为 0.91 和 0.92 (图 6h)。总体而言, DecodeR 可用于肿瘤的疾病样本的突变和 m6A RNA 修饰的检测, 而且检出结果真实可靠, 被一代测序、二代测序验证, 并且与直接 RNA 测序一致性较高。

[0163] 通过以上的实施方式的描述, 本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现, 当然也可以通过硬件, 但很多情况下前者是更佳的实施方式。基于这样的理解, 本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来, 该计算机软件产品存储在一个存储介质 (如 ROM/RAM、磁碟、光盘) 中, 包括若干指令用以使得一台计算机终端 (可以是手机, 计算机, 服务器, 或者网络设备等) 执行本发明各个实施例所述的方法。

[0164] 上面结合附图对本发明的实施例进行了描述, 但是本发明并不局限于上述的具体实施方式, 上述的具体实施方式仅仅是示意性的, 而不是限制性的, 本领域的普通技术人员在本发明的启示下, 在不脱离本发明宗旨和权利要求所保护的范围情况下, 还可做出很多形式, 这些均属于本发明的保护之内。

权利要求

1. 一种直接 RNA 纳米孔测序方法, 其特征在于, 所述方法用于对 n 种测试样本 RNA 多核苷酸进行测序, 其中 $2 \leq n \leq 24$ 且 n 为整数, 所述方法包括:
S1 针对每一种测试样本 RNA 多核苷酸, 提供一种双链 DNA 多核苷酸来捕获一种待测 RNA 序列并形成第一杂交结构; 所述双链 DNA 多核苷酸为逆转录接头的前向链和后向链组装成的双链结构, 所述前向链从 5' 至 3' 方向包括条形码序列和第一连接序列, 所述后向链从 5' 至 3' 方向包括第二连接序列、条形码序列和 Poly(T) 序列, 所述待测 RNA 序列包括所述测试样本 RNA 多核苷酸和 Poly(A) 序列, 所述双链 DNA 多核苷酸通过所述 Poly(T) 序列捕获所述待测 RNA 序列; 所述条形码序列对所述每一种测试样本 RNA 多核苷酸具有特异性, 所述条形码序列选自 SEQ ID NOs:1-24 中的任意一种;
S2 将所述第一杂交结构逆转录, 形成第二杂交结构;
S3 将测序接头与所述第二杂交结构连接, 形成第三杂交结构;
S4 对所述第三杂交结构进行纳米孔测序, 获得原始测序数据;
S5 提取所述原始测序数据对应的电流信号数据;
S6 将所述电流信号数据中所述条形码序列的电流信号数据进行分割, 并调用基于所述条形码序列训练得到的分类模型进行分析, 获得所述条形码序列的序列信息, 进而获得所述 n 种测试样本 RNA 多核苷酸的分类信息。
2. 如权利要求 1 所述的方法, 其特征在于, 所述分类模型的训练方法包括将训练样本数据中条形码序列的电流信号数据进行分割, 利用机器学习算法对分割后的条形码序列的电流信号数据进行训练学习, 从而确定所述分类模型, 其中所述训练样本数据包括一种或多种训练样本 RNA 多核苷酸和对应的条形码序列的电流信号数据, 所述条形码序列选自 SEQ ID NOs:1-24 中的任意一种。
3. 如权利要求 1 或 2 所述的方法, 其特征在于, 所述分割为将所述条形码序列的电流信号数据分割成 70-100 列矩阵。
4. 如权利要求 2 所述的方法, 其特征在于, 所述训练样本 RNA 多核苷酸包括体外转录 RNA。
5. 如权利要求 2 所述的方法, 其特征在于, 所述机器学习算法包括 K 最近邻、神经网络、朴素贝叶斯分类、回归树、AdaBOOST 和随机森林中的一种或多种。
6. 如权利要求 1 所述的方法, 其特征在于, 所述测试样本 RNA 多核苷酸的来源包括病原体、植物、浮游动物、昆虫、哺乳动物和肿瘤中的一种或多种。
7. 如权利要求 1 所述的方法, 其特征在于, 所述分类的分类可能性的阈值 > 0.3 。
8. 如权利要求 1 或 2 所述的方法, 其特征在于, 所述分割为将所述条形码序列的电流信号数据分割成 100 列矩阵。
9. 如权利要求 1 所述的方法, 其特征在于, 所述 S5 进一步包括提取所述原始测序数据对应的原始电流信号数据, 并对所述原始电流信号数据进行预处理, 获得预处理后的电流信号数据, 所述预处理包括降噪处理、抛光处理和归一化处理中的一种或多种。
10. 如权利要求 2 所述的方法, 其特征在于, 所述方法进一步包括将所述一种或多种训练样本 RNA 多核苷酸的电流信号数据与对应的参考序列比对, 以比对质量大于 60、比对至唯一参考序列为条件将不满足条件的读取序列过滤。
11. 一种直接 RNA 测序系统, 其特征在于, 包括:
测序数据接收模块, 用于接收和存储测序数据, 所述测序数据包括双链多核苷酸

的测序数据，所述双链多核苷酸包括第一链和第二链，所述第一链从 5' 至 3' 方向包括 RNA 多核苷酸、Poly(A) 序列、条形码序列、第一连接序列和测序接头，所述第二链从 5' 至 3' 方向包括测序接头、第二连接序列、条形码序列、Poly(T) 序列和所述 RNA 多核苷酸的逆转录序列；所述测序数据包括测试样本数据和训练样本数据；所述测试样本数据包括 n 种测试样本 RNA 多核苷酸对应的测序数据，其中 $2 \leq n \leq 24$ 且 n 为整数，所述条形码序列对每一种测试样本 RNA 多核苷酸具有特异性，所述条形码序列选自 SEQ ID NOs:1-24 中的任意一种；信号处理模块，用于将所述测序数据处理成电流信号数据；分类模块，用于调用基于所述测试样本数据中的所述条形码序列构建的分类模型，对所述测试样本数据的电流信号数据进行分类并输出所述 n 种测试样本 RNA 多核苷酸的分类结果。

12. 如权利要求 11 所述的系统，其特征在于，所述系统还包括模型训练模块，所述模型训练模块利用机器学习算法对所述训练样本数据的电流信号数据进行训练学习，从而确定所述分类模型。
13. 如权利要求 11 所述的系统，其特征在于，所述信号处理模块对所述测序数据进行预处理和分割。
14. 如权利要求 13 所述的系统，其特征在于，所述预处理包括降噪处理、抛光处理和归一化处理中的一种或多种。
15. 如权利要求 13 所述的系统，其特征在于，所述分割为将所述条形码的电流信号数据分割成 70-100 列矩阵。
16. 如权利要求 12 所述的系统，其特征在于，所述机器学习算法包括 K 最近邻、神经网络、朴素贝叶斯分类和回归树、AdaBOOST 和随机森林中的一种或多种。
17. 如权利要求 11 所述的系统，其特征在于，所述系统进一步包括过滤模块，用于基于设置的分类可能性的阈值，将低于所述阈值的读取过滤掉。
18. 如权利要求 11 所述的系统，其特征在于，所述训练样本数据包括体外转录 RNA 的测序数据。
19. 如权利要求 11 所述的系统，其特征在于，所述测试样本 RNA 多核苷酸的来源包括病原体、植物、浮游动物、昆虫、哺乳动物和肿瘤中的一种或多种。
20. 如权利要求 17 所述的系统，其特征在于，所述阈值 >0.3 。