

DESCRIPTION

TITLE OF INVENTION: [MULTIPLE MIXED SAMPLE DIRECT RNA NANOPORE SEQUENCING METHOD AND SYSTEM]

PRIORITY APPLICATION

[0001] This application claims the priority to Chinese patent application No. 2023102647091, titled A Novel Method and System for Direct RNA Nanopore Sequencing, filed on 17 March 2023, the contents of which are incorporated herein by reference.

TECHNICAL FIELD

[0002] The present disclosure relates generally to the field of sequencing technologies, and more particularly to a multiplexed mixed-sample direct RNA nanopore sequencing method and system.

BACKGROUND

[0003] The development of third-generation sequencing technologies has fundamentally transformed the capability of long-read sequencing for genomes and transcriptomes. These technologies enable direct sequencing of DNA and RNA and their modifications without fragmentation or amplification. Oxford Nanopore Technologies (ONT) direct RNA sequencing (DRS) performs sequencing of RNA molecules directly through a series of protein nanopores embedded in a synthetic membrane.

[0004] As a motor protein gradually threads an individual RNA molecule through a nanopore, different nucleotides produce distinct current blockages, resulting in electrical-current fluctuations that can be recorded through base-calling algorithms to determine the RNA sequence. Direct RNA sequencing can produce long reads (up to 21 kb) and is designed to obtain full-length or near-full-length RNA transcripts or RNA genomes. Full-length sequencing enables powerful applications such as measuring RNA poly(A) tail length and analyzing the regulatory functions of poly(A) tails in gene expression. In addition, direct RNA sequencing can directly detect RNA modifications, such as m⁶A, 7-methylguanosine (m⁷G), pseudouridine (Ψ), and 2'-O-methylation (Nm), because these modifications generate distinct electrical-current signatures.

[0005] Furthermore, amplification-free library preparation avoids amplification or reverse-transcription (RT) bias, making DRS particularly suitable for identifying exons (exonic introns), which are often artifacts caused by reverse transcription. These advantages make DRS especially powerful for determining RNA structural subtypes, identifying tRNA and rRNA, and sequencing genomes of organisms such as RNA viruses, prokaryotes, and parasites. Recent studies have also applied DRS to transcriptome or epitranscriptome profiling across diverse species, including humans, animals (e.g., nematodes, insects), plants, zooplankton, and pathogens. Another application of DRS is analyzing RNA metabolic dynamics through incorporation of base analogs into nascent RNA, such as 5-ethynyl uridine or 4-thiouridine, followed by ONT direct RNA sequencing.

[0006] Despite its unique advantages, DRS has two major limitations: the lack of an effective multi-sample mixed-sequencing method, and the requirement for large RNA input, typically around 500 ng per sequencing run. ONT does provide barcoding protocols for multiplex cDNA libraries, in which DNA adapters are directly ligated onto cDNA, allowing both the barcode and the cDNA sequences to be recognized under the DNA signal model. However, in DRS, the adapters are DNA oligonucleotides while the targets are RNA molecules, making it impossible for DNA adapters to be accurately recognized under either the RNA model or DNA model. This is

because DNA translocates through the nanopore much faster (450 bp/sec) than RNA (70 nt/sec), by a factor of six. To overcome these limitations, a method called DeePlexiCon has been developed to multiplex a limited number of RNA samples on a single flow cell to generate a multiplexed direct RNA sequencing dataset, thereby reducing the per-sample sequencing cost. However, due to constraints in barcode design and its algorithm, this method can only multiplex four samples per flow cell. Although some publications report that DeePlexiCon can be retrained to produce new models, it still fails to accurately demultiplex reads when different sequences are used or when more than four barcodes are included. Therefore, providing additional barcodes to increase multiplexing capacity while maintaining high accuracy remains a significant challenge.

SUMMARY

[0007] In a first aspect, the present invention provides a direct RNA nanopore sequencing method (abbreviated as DecodeR). The method is used for sequencing n types of test-sample RNA polynucleotides, wherein $2 \leq n \leq 24$ and n is an integer. The method comprises:

[0008] S1: For each test-sample RNA polynucleotide, providing a double-stranded DNA polynucleotide to capture a target RNA sequence and to form a first hybrid structure; the double-stranded DNA polynucleotide is a double-stranded structure assembled from a forward strand and a reverse strand of a reverse transcription adapter. The forward strand comprises, in the 5'-to-3' direction, a barcode sequence and a first ligation sequence; the reverse strand comprises, in the 5'-to-3' direction, a second ligation sequence, the barcode sequence, and a poly(T) sequence. The target RNA sequence comprises the test-sample RNA polynucleotide and a poly(A) sequence. The double-stranded DNA polynucleotide captures the target RNA sequence through the poly(T) sequence. The barcode sequence is specific to each test-sample RNA polynucleotide and is selected from any one of SEQ ID NOs: 1–24.

[0009] S2: Reverse-transcribing the first hybrid structure to form a second hybrid structure.

[0010] S3: Ligating a sequencing adapter to the second hybrid structure to form a third hybrid structure.

[0011] S4: Performing nanopore sequencing on the third hybrid structure to obtain raw sequencing data.

[0012] S5: Extracting electrical-current signal data corresponding to the raw sequencing data.

[0013] S6: Segmenting electrical-current signal data corresponding to the barcode sequence from the electrical-current signal data, and analyzing the segmented electrical-current signal data using a classification model trained based on the barcode sequence to obtain sequence information of the barcode sequence, thereby obtaining classification information for the n types of test-sample RNA polynucleotides.

[0014] In some embodiments, the double-stranded DNA polynucleotide is generated by pre-annealing.

[0015] In some embodiments, the pre-annealing comprises mixing the forward and reverse strands of each reverse transcription adapter in a buffer solution, followed by the following program: denaturation at 95°C for 5 min; annealing at 65°C, 50°C, 37°C, and 22°C for 30 min each; storage at 4°C.

[0016] In some embodiments, the first ligation sequence comprises TAGTAGGTTC.

[0017] In some embodiments, the second ligation sequence comprises GAGGCGAGCGGTCAATTTT.

- [0018]** In some embodiments, the poly(T) sequence comprises 10–30 repeated T nucleotides.
- [0019]** In some embodiments, the poly(A) sequence comprises 10–30 repeated A nucleotides.
- [0020]** In some embodiments, the poly(T) sequence comprises 10 repeated T nucleotides.
- [0021]** In some embodiments, the poly(A) sequence comprises 10 repeated A nucleotides.
- [0022]** In some embodiments, the training method of the classification model comprises segmenting electrical-current signal data of the barcode sequence in training-sample data, and training the classification model using a machine-learning algorithm based on the segmented electrical-current signal data to determine the classification model; the training-sample data comprises one or more training-sample RNA polynucleotides and corresponding electrical-current signal data of the barcode sequence; and the barcode sequence is selected from any one of SEQ ID NOs: 1–24.
- [0023]** In some embodiments, the classification model is trained based on any 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, or 24 barcode sequences selected from SEQ ID NOs: 1–24.
- [0024]** In some embodiments, the training-sample RNA polynucleotide comprises an in vitro-transcribed RNA.
- [0025]** In some embodiments, the segmentation comprises segmenting electrical-current signal data of the barcode sequence into a 70–100 column matrix.
- [0026]** In some embodiments, the segmentation comprises segmenting electrical-current signal data of the barcode sequence into a 100-column matrix.
- [0027]** In some embodiments, the machine-learning algorithm comprises one or more of k-nearest neighbor, neural network, Naive Bayes classification, classification and regression tree, AdaBOOST, and random forest.
- [0028]** In some embodiments, the machine-learning algorithm comprises random forest.
- [0029]** In some embodiments, the source of the test-sample RNA polynucleotide comprises one or more selected from pathogens, plants, zooplankton, insects, mammals, and tumors.
- [0030]** In some embodiments, the pathogens comprise one or more selected from RNA viruses, bacteria, fungi, and parasites.
- [0031]** In some embodiments, the in vitro-transcribed RNA comprises one or more genes listed in Table 2.
- [0032]** In some embodiments, the threshold of classification probability is > 0.3 .
- [0033]** In some embodiments, Step S5 further comprises extracting raw electrical-current signal data corresponding to the raw sequencing data and preprocessing the raw electrical-current signal data to obtain preprocessed electrical-current signal data, the preprocessing comprising one or more selected from denoising, polishing, and normalization.
- [0034]** In some embodiments, the method further comprises aligning the electrical-current signal data of the one or more training-sample RNA polynucleotides to corresponding reference sequences, and filtering out read sequences that do not satisfy alignment quality > 60 and alignment to a unique reference sequence.

[0035] In a second aspect, the present invention provides a kit for direct RNA sequencing, comprising the barcode sequences of reverse transcription adapters shown as SEQ ID NOs: 1–24.

[0036] In a third aspect, the present invention provides a direct RNA nanopore sequencing system, comprising:

[0037] A sequencing-data receiving module configured to receive and store sequencing data, the sequencing data comprising sequencing data of a double-stranded polynucleotide. The double-stranded polynucleotide comprises a first strand and a second strand. The first strand comprises, in the 5'-to-3' direction, an RNA polynucleotide, a poly(A) sequence, a barcode sequence, a first ligation sequence, and a sequencing adapter. The second strand comprises, in the 5'-to-3' direction, a sequencing adapter, a second ligation sequence, the barcode sequence, a poly(T) sequence, and a reverse-transcription sequence of the RNA polynucleotide. The sequencing data comprise test-sample data and training-sample data. The test-sample data comprise sequencing data corresponding to n test-sample RNA polynucleotides, wherein $2 \leq n \leq 24$ and n is an integer. The barcode sequence is specific to each test-sample RNA polynucleotide and is selected from any one of SEQ ID NOs: 1–24.

[0038] A signal-processing module configured to process the sequencing data into electrical-current signal data.

[0039] A classification module configured to invoke a classification model constructed based on the barcode sequences in the test-sample data, to classify electrical-current signal data of the test-sample data, and to output classification results of the n test-sample RNA polynucleotides.

[0040] In some embodiments, the system further comprises a model-training module configured to train electrical-current signal data of the training-sample data using a machine-learning algorithm to determine the classification model.

[0041] In some embodiments, the double-stranded polynucleotide is an RNA/DNA hybrid double-stranded structure.

[0042] In some embodiments, the signal-processing module performs preprocessing and segmentation on the sequencing data.

[0043] In some embodiments, the preprocessing comprises one or more selected from denoising, polishing, and normalization.

[0044] In some embodiments, the segmentation comprises segmenting electrical-current signal data of the barcode sequence into a 70–100 column matrix.

[0045] In some embodiments, the segmentation comprises segmenting electrical-current signal data of the barcode sequence into a 100-column matrix.

[0046] In some embodiments, the machine-learning algorithm comprises one or more of k-nearest neighbor, neural network, Naive Bayes classification, classification and regression tree, AdaBOOST, and random forest.

[0047] In some embodiments, the machine-learning algorithm comprises random forest.

[0048] In some embodiments, the system further comprises a filtering module configured to filter out reads having classification probabilities below a predetermined threshold.

[0049] In some embodiments, the training-sample data comprise sequencing data of an in vitro-transcribed RNA.

[0050] In some embodiments, the source of the test-sample RNA polynucleotide comprises one or more selected from pathogens, plants, zooplankton, insects, mammals, and tumors.

[0051] In some embodiments, the pathogens comprise one or more selected from RNA viruses, bacteria, fungi, and parasites.

[0052] In some embodiments, the predetermined threshold of classification probability is greater than 0.3.

[0053] In some embodiments, the in vitro-transcribed RNA comprises one or more genes listed in Table 2.

[0054] In some embodiments, the ratio of the training set to the testing set in the training-sample data is 7:3.

[0055] In some embodiments, the model-training module is configured to align electrical-current signal data of the training-sample data to corresponding reference sequences, and to filter out read sequences that do not satisfy alignment quality greater than 60 and alignment to a unique reference sequence.

[0056] Advantageous Effects

[0057] In the prior art, the nanopore direct RNA mixed-sample sequencing classifier DeePlexiCon can multiplex a limited number of RNA samples on a single flow cell to generate a multiplex direct RNA sequencing dataset, thereby reducing the sequencing cost per sample. However, DeePlexiCon can only deconvolute four samples on the same flow cell, and has multiple limitations, including low multiplexing throughput (number of mixed samples), low deconvolution accuracy, and the need to discard a large amount of sequencing data in order to achieve accurate deconvolution. These limitations greatly increase the overall sequencing cost.

[0058] The method and system (which may also be understood as a classifier) provided by the present invention (abbreviated as DecodeR) provide 24 barcode sequences to label multiple RNA samples on a single flow cell, and, by demultiplexing direct RNA-seq data and performing machine-learning-based classification, enable high-precision multiplexing of a larger number of samples with a low detection limit on a single flow cell, thereby overcoming the limitations of low RNA sample amount and high sequencing cost. In the present invention, segmenting the barcode electrical-current signals into a 70–100 column matrix can reduce interference between current signals and reduce noise. Validation has shown that the recovery rate and accuracy of DecodeR can reach 100% and 92.2%, respectively. DecodeR has been applied to various pathogens (RNA viruses, bacteria, fungi, and *Plasmodium*) and glioma samples, and has yielded good validation results. In addition, DecodeR has revealed several meaningful findings regarding the transcriptomes and mutations of different glioma samples, indicating that the present invention has substantial potential for practical application. In nanopore direct RNA sequencing, the present invention can achieve simultaneous sequencing of multiple biological samples on a single flow cell, reduce the initial input amount of biological samples, reduce the cost of multi-sample sequencing, and enable sequencing of trace biological samples. DecodeR has already been applied to tumor patient samples and has potential for application to other clinical samples.

[0059] In summary, the present invention provides a complete, robust, high-throughput, and low-cost direct RNA mixed-sample sequencing method and system. By providing more barcode sequences, the present invention increases multiplexing capacity and enables high-precision sequencing and analysis of multiple biological samples in nanopore direct RNA sequencing (DRS), and is expected to be practically applicable to clinical samples.

[0060] As used herein, the terms “nucleic acid,” “nucleic acid molecule,” “polynucleotide,” and “oligonucleotide” are used interchangeably and refer to a sequence of covalently linked

nucleotides (i.e., ribonucleotides for RNA and deoxyribonucleotides for DNA), in which the 3' position of the pentose of one nucleotide is linked via a phosphodiester group to the 5' position of the pentose of the next nucleotide. "Polynucleotide" includes single-stranded and double-stranded polynucleotides, for example various DNA and RNA molecules, any hybrids thereof, or derivatives, combinations, or fragments thereof.

[0061] As used herein, the term "test-sample RNA polynucleotide" refers to RNA that is intended to be sequenced. The RNA may be coding and/or non-coding RNA, for example mRNA, siRNA, rRNA, miRNA, tRNA, lncRNA, snoRNA, snRNA, exRNA, or piRNA.

[0062] As used herein, the term "complementary" refers to the ability of two nucleic acid sequences to form hydrogen bonds with each other according to base-pairing principles (Watson-Crick rules), thereby forming a double-stranded structure. In the present invention, "complementary" includes substantially complementary and fully complementary. "Fully complementary" means that every base in one nucleic acid sequence can pair with a base in the other nucleic acid sequence, without mismatches or gaps. "Substantially complementary" means that most bases in one nucleic acid sequence can pair with bases in the other nucleic acid sequence, while allowing mismatches or gaps. Typically, under conditions that permit nucleic acid hybridization, annealing, or amplification, two complementary (for example, substantially or fully complementary) nucleic acid sequences will selectively/specifically hybridize or anneal to each other and form a double-stranded structure. Conversely, "non-complementary" (or "mismatched") refers to two nucleic acid sequences that cannot hybridize or anneal under conditions that permit nucleic acid hybridization, annealing, or amplification.

[0063] As used herein, the terms "hybridization" and "annealing" are used interchangeably and refer to the process by which complementary single-stranded nucleic acid molecules form double-stranded nucleic acid. Typically, two fully complementary or substantially complementary nucleic acid sequences can undergo hybridization or annealing.

[0064] As used herein, the term "downstream" (or "3'-ward") is used to describe the relative positional relationship between two nucleic acid sequences, and has the meaning commonly understood by those skilled in the art. For example, "one nucleic acid sequence is downstream of another nucleic acid sequence" means that, when arranged in the 5'-to-3' direction, the former is located at a position further toward the 3' end than the latter. Conversely, "upstream" (or "5'-ward") has the opposite meaning to "downstream."

[0065] As used herein, the term "barcode" refers to a specific nucleotide sequence used to label RNA to be sequenced. In the present invention, the barcode sequence is designed, i.e., the barcode sequence is known. In some embodiments, the barcode sequence comprises one or more selected from SEQ ID NOs: 1–24. In a single sequencing run, the barcode sequence corresponding to each RNA to be sequenced is unique and specific (i.e., has specificity). In addition, the position of the barcode sequence in the sequencing construct of the present invention is fixed (for example, downstream of the poly(A) sequence), and therefore the barcode sequence in a read can be located based on the continuous "A" signal of the poly(A) sequence. In the present invention, a classification model is constructed by training on training-sample data generated from different barcode combinations (for example, any 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, or 24 barcode sequences selected from SEQ ID NOs: 1–24) using a machine-learning algorithm. Accordingly, in a single sequencing run (for example, on a single flow cell), based on the barcode sequence linked to the RNA to be sequenced, an appropriate classification model trained with the corresponding barcode combination can be selected to analyze the sequencing data, thereby achieving classification of the RNA to be sequenced. In some embodiments, in a single sequencing run, the method provided by the present invention can achieve sequencing of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, or 24 RNAs to be sequenced.

[0066] As used herein, the term “read” refers to any part or all of the sequencing information of a nucleic acid molecule obtained by a sequencing method. Reads can be stored on a storage medium and appropriately processed to determine whether they match a reference sequence or meet other criteria. Reads can be obtained directly from a sequencing device, or indirectly from stored sequence information of a sample.

[0067] As used herein, the term “reference sequence” refers to any known sequence to which reads are aligned. In some embodiments, the reference sequence may correspond to all or only part of a genome or transcriptome of an organism.

[0068] As used herein, the term “alignment” refers to the process of comparing a read with a reference sequence to determine whether the reference sequence contains the read sequence. If the reference sequence contains the read, the read can be mapped to the reference sequence or to a specific position within the reference sequence. The match between the read and the reference sequence in an alignment may be a 100% sequence match or less than 100% (i.e., not a perfect match).

[0069] As used herein, the term “mixed sample” refers to a sample comprising a mixture of nucleic acid molecules derived from different genomes.

[0070] As used herein, the term “library” refers to a collection of nucleic acid molecules derived from one or more nucleic acid samples. In some embodiments, the nucleic acid molecules comprise distinguishable sequence tags (for example, barcodes). In some embodiments, two or more libraries can be combined to create a library pool.

[0071] As used herein, the term “multiplex” or “multiplexing” refers to the combination of one or more nucleic acid samples or nucleic acid library molecules derived therefrom in a single pool, tube, or reaction (for example, on a single flow cell).

[0072] As used herein, the term “sample” refers to a sample containing biomolecules, typically derived from biological fluids, cells, tissues, organs, or organisms. The term “biological sample” refers to any sample of biological origin that contains biomolecules. In the present invention, the biomolecules comprise nucleic acids or nucleic acid mixtures, preferably RNA. In some embodiments, the “sample” or “biological sample” may be a liquid sample (for example, body fluids, including blood, plasma, serum, urine, vaginal fluid, ovarian cyst fluid, vaginal lavage fluid, pleural fluid, ascites, cerebrospinal fluid, saliva, sweat, tears, sputum, bronchoalveolar lavage fluid, nipple aspirate, and aspirates from different body sites such as thyroid or breast), or a solid sample (for example, tissue samples, biopsy samples, or tissue culture or cells and progeny thereof). In some embodiments, the “sample” or “biological sample” may comprise a sample enriched for a particular type of molecule, for example nucleic acids. In some embodiments, the “sample” or “biological sample” may also comprise clinical samples, for example tissues obtained by surgical resection, tissues obtained by biopsy, cells in culture, cell supernatants, cell lysates, tissue samples, organs, bone marrow, blood, plasma, serum, and the like. In some embodiments, the “sample” or “biological sample” may further comprise tumor samples, for example samples of cells (such as cancer cells) from cancer patients, or nucleic acid samples obtained from cancer cells of cancer patients (for example, cell lysates, cell extracts). In some embodiments, nucleic acids in the “sample” or “biological sample” may be cell-free. In some embodiments, the source of the “sample” or “biological sample” may be pathogens, for example RNA viruses, bacteria, fungi, or parasites. In some embodiments, the source of the “sample” or “biological sample” may be plants, zooplankton, insects, or mammals (for example humans, dogs, cats, horses, goats, sheep, cattle, pigs, rats, mice, and the like). In some embodiments, the source of the “sample” or “biological sample” may also be tumors. The “sample” or “biological sample” may be used directly after being obtained from its source, or may be used after pretreatment (for example fixation, embedding in a medium, sectioning, washing, or enrichment). In some embodiments, nucleic acids (for example RNA polynucleotides) in the “sample” or “biological sample” are further extracted.

[0073] As used herein, the term “test sample” refers to a sample or biological sample containing nucleic acids that are intended to be sequenced. In some embodiments, the nucleic acids are preferably RNA polynucleotides. In some embodiments, the test sample is derived from a biological source. In some embodiments, the test sample may also be derived from the environment, for example soil, lakes, rivers, and the like.

BRIEF DESCRIPTION OF THE DRAWINGS

[0074] In order to more clearly illustrate the technical solutions of the embodiments of the present invention or of the prior art, the drawings required for use in describing the embodiments or the prior art are briefly introduced below. In all of the drawings, similar elements or parts are generally denoted by similar reference numerals. The elements or parts shown in the drawings are not necessarily drawn to scale. It will be apparent that the drawings described below illustrate only some embodiments of the present invention, and that, for those of ordinary skill in the art, other drawings may also be derived from these drawings without inventive effort.

[0075] **FIG. 1** illustrates a schematic representation of DecodeR, an exemplary workflow, and an algorithm comparison.

[0076] **FIG. 2** illustrates a comparison between DecodeR of the present invention and classifiers of the prior art.

[0077] **FIG. 3** illustrates a comparison among different barcode sequences.

[0078] **FIG. 4** illustrates the performance of DecodeR of the present invention in real biological samples.

[0079] **FIG. 5** illustrates the detection of viral-sample RNA using DecodeR of the present invention.

[0080] **FIG. 6** illustrates mutation analysis and m⁶A analysis of tumor samples using DecodeR of the present invention.

[0081] **FIG. 7** illustrates a flowchart of a direct RNA nanopore sequencing method provided by the present invention.

[0082] **FIG. 8** illustrates a schematic diagram of modules of a multi-barcode direct RNA nanopore sequencing classification system provided by the present invention.

DETAILED DESCRIPTION

[0083] To make the objectives, technical solutions, and advantages of the embodiments of the present invention clearer, the technical solutions of the embodiments of the present invention will be described in detail below with reference to the accompanying drawings. It is apparent that the embodiments described are only part of the embodiments of the present invention, rather than all embodiments. Based on the embodiments of the present invention, all other embodiments obtained by those of ordinary skill in the art without inventive effort shall fall within the scope of protection of the present invention.

[0084] In the present specification, the suffixes “module,” “component,” or “unit,” when used to describe elements, are merely for facilitating the description of the present invention and do not carry any specific limiting meaning. Therefore, the terms “module,” “component,” or “unit” may be used interchangeably.

[0085] In the present specification, the terms “upper,” “lower,” “inner,” “outer,” “front,” “rear,”

“one end,” “the other end,” and the like are used to indicate positional or spatial relationships based on the orientation or positional relationships shown in the drawings. These descriptions are intended merely for convenience and clarity of illustration, rather than indicating or implying that the referenced devices or elements must have a particular orientation or must be constructed or operated in a particular orientation. Therefore, such terms should not be construed as limiting the present invention. In addition, the terms “first” and “second” are used merely for descriptive purposes and should not be understood as indicating relative importance.

[0086] As used herein, the term “and/or” includes any and all combinations of one or more of the listed items.

[0087] As used herein, the term “plurality” means two or more, including two, three, four, five, and so on.

[0088] It should be noted that, in the present specification, the terms “comprise”, “include”, or any variations thereof are intended to cover non-exclusive inclusion. Thus, a process, method, article, or apparatus that comprises a series of elements not only includes those elements, but may also include other elements not expressly listed, or may include inherent elements of such process, method, article, or apparatus. Without additional limitations, an element defined by the phrase “comprising a ...” does not exclude the presence of other identical elements in the process, method, article, or apparatus that comprises the stated element.

[0089] As used in the present specification, the term “about” typically refers to $\pm 5\%$ of the stated value, more typically $\pm 4\%$, more typically $\pm 3\%$, more typically $\pm 2\%$, even more typically $\pm 1\%$, and even more typically $\pm 0.5\%$ of the stated value.

[0090] In the present specification, certain embodiments may be disclosed in a “range” format. It should be understood that such “range” descriptions are provided merely for convenience and conciseness, and should not be interpreted as a rigid limitation of the disclosed range. Accordingly, the description of a range should be considered as specifically disclosing all possible sub-ranges and individual numerical values within that range. For example, the disclosure of a range from 1 to 6 should be considered as specifically disclosing sub-ranges such as 1 to 3, 1 to 4, 1 to 5, 2 to 4, 2 to 6, and 3 to 6, as well as individual values within the range, such as 1, 2, 3, 4, 5, and 6. The above rule applies regardless of the breadth of the disclosed range.

[0091] EMBODIMENT 1

[0092] 1.1 Barcode Design and Synthesis

[0093] As shown in Table 1, this embodiment includes 40 newly designed barcode sequences of reverse transcription adapters (RTAs) of the present invention (RTA-01 to RTA-40, each 20–28 bp in length), as well as the barcode sequence of the official ONT RTA (RTA-00), three barcode sequences from DeePlexiCon RTAs (RTA-41, RTA-42, and RTA-43), and four barcode sequences from Poreplex RTAs (RTA-44, RTA-45, RTA-46, and RTA-47), for a total of 48 RTA barcode sequences.

[0094] The sequencing construct on which the present invention is based is shown in FIG. 1a. Each RTA pair consists of a forward strand and a reverse strand. In the 5'-to-3' direction, the forward strand comprises a barcode sequence (X1) followed by a first ligation sequence (X2, also referred to as a first mismatched polynucleotide sequence). The reverse strand comprises a second ligation sequence (X2, also referred to as a second mismatched polynucleotide sequence), a barcode sequence (X1), and a poly(T) sequence (e.g., 10–30 repeated T nucleotides) downstream of the barcode sequence. The forward and reverse strands of the RTAs are hybridized and assembled into a double-stranded structure. According to the principle of base complementarity, each RTA includes a barcode region (X1), a mismatched polynucleotide region

(X2), and an overhanging poly(T) tail. Next, the assembled double-stranded structure captures and hybridizes with a target RNA sequence containing a poly(A) tail, and reverse transcription produces an RNA/DNA hybrid duplex. The resulting RNA/DNA hybrid duplex is then ligated to the ONT sequencing adapter (X3), thereby forming the complete sequencing construct employed by the present invention. In some embodiments, the first mismatched polynucleotide sequence comprises TAGTAGGTTC (SEQ ID NO:49), and the second mismatched polynucleotide sequence comprises GAGGCGAGCGGTCAATTTT (SEQ ID NO:50).

[0095] Subsequently, this embodiment uses the OligoAnalyzer™ tool to evaluate the GC content, melting temperature, and secondary structure of the DNA oligonucleotides designed in the present invention. The length of the designed barcode sequences is controlled within 20–28 nt to avoid potential issues during bead-based purification, where excessively long barcodes may interfere with removing excess RTAs.

[0096] Table 1 Customized RTA Sequences

RTA ID	Source	Barcode Sequence	SEQ ID NO:
RTA-00	DeePlexiCon-BC1(D-BC1)	GGCTTCTTCTTGCTCTTAGG	25
RTA-01	NewDesign01	CCTGGTAACTGGGACACAAGACTC	1
RTA-02	NewDesign02	ACGTAAC TTGGTTTGTTCCTGAA	2
RTA-03	NewDesign03	GAGAGGACAAAGGTTTCAACGCTT	3
RTA-04	NewDesign04	TCCATTCCCTCCGATAGATGAAAC	26
RTA-05	NewDesign05	CACACACCGACAAC TTTCTT	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	TTCGGATTCTATTCCTCGTGTCTA	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	TTGCGTCCTGTTACGAGAACAT	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	AGATTTCAGACCAAGTCTCATGCAAAG	35
RTA-29	NewDesign29	TGGAAGATGAGACATCCTGATCTACG	19
RTA-30	NewDesign30	TCACTACTCACGACAGGTGGCATGAA	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 Preparation of In Vitro-Transcribed (IVT) RNA

[0098] To evaluate the barcode sequences and to perform model training, this embodiment generated 52 in vitro-transcribed (IVT) RNAs. The RNA sequences were derived from 27 mouse genes, 22 human genes, and 3 SARS-CoV-2 genes (Table 2). For these gene sequences, forward and reverse primers were designed using the Primer3web tool (Version 4.1.0). A T7 RNA polymerase promoter sequence was added to each forward primer, and an oligo(dT)15 sequence (15 repeated T nucleotides) was added to each reverse primer.

[0099] Human and mouse total RNA, as well as SARS-CoV-2 genomic RNA, were reverse-transcribed into first-strand cDNA using a reverse transcriptase. PCR amplification of the various DNA fragments was then performed using the primers described above, and agarose gel electrophoresis was used to assess the quality of the PCR products. One microgram of purified PCR product was used as a template for IVT RNA synthesis by T7 RNA polymerase, followed by purification using the RNA Clean & Concentrator-5 kit according to the manufacturer's instructions. The length of each IVT RNA was determined using the Qsep100 system, and the concentration and purity of each IVT RNA product were measured using a NanoDrop 2000.

[0100] Table 2 Information on In Vitro-Transcribed RNA Sequences

Gene	RNA Name	Length
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), transcript variant 7, mRNA 476 652	177
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 Workflow for Mixed-Sample Direct RNA Sequencing Library Preparation

[0102] Assembly of customized RTAs. For each RTA, the forward strand (1.54 μ M) and the reverse strand (1.4 μ M) were mixed in buffer (10 mM Tris-HCl, pH 7.5, 50 mM KCl) (Table 3). In this embodiment, an excess amount of forward strand was used to ensure complete consumption of the reverse strand, thereby preventing residual reverse strand from reducing the yield of the RNA-RTA ligation product. The customized RTAs were assembled using the following PCR program: denaturation at 95°C for 5 min; annealing at 65°C, 50°C, 37°C, and 22°C for 30 min each; followed by holding at 4°C (Table 4). Finally, 4% agarose gel electrophoresis was used to confirm whether RTA assembly was successful. The assembled products serve as RTAs carrying

different barcode sequences for use in the mixed-sample direct RNA sequencing workflow of the present invention.

[0103] Table 3 Assembly of Customized RTAs

Reagent	Stock Concentration	Final Concentration	Volume
Forward strand (For)	10uM	1.54uM	2.31uL
Reverse strand (Rev)	10uM	1.40uM	2.1uL
Tris-HCL	50mM	10mM	3μL
KCL	2M	50mM	0.375μL
H2O	-	-	7.215μL

[0104] Table 4 PCR Program for Assembly of Customized RTAs

Step	Temperature	Time	Cycles
Pre-denaturation	95°C	5min	1
	65°C	30min	
Annealing	50°C	30min	1
	37°C	30min	
	22°C	30min	
Storage	4°C	∞	1

[0105] Mixed-sample Direct RNA Sequencing Library Preparation. Using the NEBNext Quick Ligation Module (NEB, E6056), each RNA sample containing a poly(A) tail was individually hybridized and ligated to a pre-annealed customized RTA, and the one-to-one correspondence between the RNA sample and the RTA was recorded. SuperScript IV reverse transcriptase (ThermoFisher, 18090200) was then used to reverse-transcribe the ligated product to form a stable RNA/DNA duplex and to remove secondary structure from the RNA. The product was purified using 1.8× RNA Clean XP beads (Beckman, A63987). Subsequently, equal molar amounts of the reverse-transcription products from each reaction were pooled, and the RNA adapter mixture (RMX) was ligated to the 3' end of the RNA/DNA hybrid at room temperature. After ligation, the reaction mixture was purified using 1× RNA Clean XP beads (Beckman, A63987), washed twice with wash buffer (WSB), and eluted with elution buffer (EB). Next, the sequencing loading mixture was prepared following the manufacturer's instructions (ONT, SQK-RNA002) and adjusted according to the library volume required for the MinION or PromethION sequencer. The loading mixture was then loaded onto a qualified ONT flow cell, sequencing parameters were set in the MinKNOW software, and sequencing was performed for more than 48 hours to obtain sufficient data.

[0106] 1.4 Training of the DecodeR Machine-learning Model

[0107] First, five independent experiments were performed using various combinations of 48 RTAs ligated to IVT RNA, generating five sets of raw DRS electrical-signal data. After base-calling and alignment to reference sequences, a total of 6,276,168 valid reads were generated (ranging from 623K to 2.30M). After quality control, 24 RTAs were filtered out due to low read counts.

[0108] Next, raw signal data (FAST5) from the IVT RNA sequencing runs were base-called using Guppy (v3.1.5) under the high-accuracy model. The base-called reads (FASTQ) were aligned to IVT RNA reference sequences using minimap2 (v2.17-r941). Reads were filtered using criteria of mapping quality (MAPQ) > 60 and unique mapping to a reference sequence, and the reads were grouped according to their aligned ligation sequences. Raw electrical signals were extracted from FAST5 files using the R package *rhdf5* (<https://github.com/grimbough/rhdf5>). The R package *smoother* (<https://CRAN.R->

project.org/package=smoother) may also be used to smooth raw electrical signals to reduce noise.

[0109] Given that poly(A) electrical signals correspond to continuous A-base signals, this embodiment polished and normalized the electrical signals and determined the position of the poly(A) tail based on the mean and variance of the signal near the RNA 3' end. Accordingly, a series of electrical signals corresponding to the ONT sequencing adapter (also referred to as "sequencing adapter," "adapter," "RNA adapter mixture," etc.), RTA barcode, poly(A), and IVT RNA sequence were obtained (FIG. 1b). Because the adapter sequence is identical in all reads, the ONT adapter signal was not removed during analysis. The *cpt.meanvar* function in the *changepoint* package (<https://CRAN.R-project.org/package=changepoint>) was then used to segment the electrical-current signal of the RTA barcode, based on changes in mean and variance. Each barcode signal was segmented into 70–100 segments, preferably 100 segments (FIG. 1c).

[0110] Taking segmentation into a 100-column matrix as an example, all 100-column matrices generated from barcode signals were used for model training. The R package *caret* (<https://CRAN.R-project.org/package=caret>) was used to streamline training of prediction models based on the random forest classification algorithm. To prevent overfitting, this embodiment used repeated K-fold cross-validation (10-fold, repeated 10 times) for resampling. To improve computational efficiency, parallel processing supported by the *doParallel* package was used. Because the number of sequence reads differed among barcodes, for each classifier, this embodiment first used the *downSample* function in *caret* to randomly sample a dataset so that the frequencies of all classes matched that of the minority class. Then, 70% of these reads were used as the training set, and the remaining reads served as the testing set. Models were trained using combinations of 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, and 24 barcodes. It should be understood that models may also be trained using other barcode combinations (e.g., 3, 5, 11, 19, 22, etc.) to generate additional models.

[0111] In summary, as shown in FIG. 7, the multiplexed direct RNA nanopore sequencing method of the present invention comprises: first, for each test-sample RNA polynucleotide, hybridizing and assembling the forward and reverse strands of an RTA into a double-stranded structure (i.e., a double-stranded DNA polynucleotide). The assembled duplex captures and hybridizes with a target RNA containing a poly(A) sequence to form a first hybrid structure (S1). The first hybrid structure is reverse-transcribed to form an RNA/DNA hybrid duplex (i.e., a second hybrid structure) (S2). The RNA/DNA hybrid duplex is ligated to the ONT sequencing adapter to form a third hybrid structure (S3). Sequencing of the third hybrid structure begins from the 3'-end motor protein region to obtain raw sequencing data (S4), which include a series of electrical signals originating from the ONT adapter, barcode, poly(A), and test-sample RNA polynucleotide. These signals may optionally be extracted, denoised, or normalized (S5). The electrical-current signal corresponding to the barcode is then segmented, and a trained classification model based on a specific number of barcodes is used to classify the signal. An appropriate filtering threshold (> 0.3) may optionally be applied. Finally, the barcode and classification probability of each electrical signal are obtained (S6).

[0112] Based on the above method, the present invention further provides a multi-barcode direct RNA nanopore sequencing system 100.

[0113] An exemplary structure of the sequencing system of the present invention is

described below. In some embodiments, as shown in FIG. 8, the sequencing system 100 may comprise:

[0114] A sequencing-data receiving module 102 configured to receive and store sequencing data. The sequencing data comprise sequencing data of a double-stranded polynucleotide comprising a first strand and a second strand. The first strand comprises, in the 5'-to-3' direction, an RNA polynucleotide, a poly(A) sequence, a barcode sequence, a first ligation sequence, and a sequencing adapter. The second strand comprises, in the 5'-to-3' direction, a sequencing adapter, a second ligation sequence, the barcode sequence, a poly(T) sequence, and a reverse-transcription sequence of the RNA polynucleotide. The sequencing data include test-sample data and training-sample data.

[0115] In some embodiments, the test-sample data comprise sequencing data corresponding to n test-sample RNA polynucleotides, wherein $2 \leq n \leq 24$ and n is an integer. The barcode sequence is specific to each test-sample RNA polynucleotide and is selected from any of SEQ ID NOs: 1–24.

[0116] In some embodiments, the source of the test-sample RNA polynucleotide comprises one or more of pathogens, plants, zooplankton, insects, mammals, and tumors. In some embodiments, the pathogens comprise one or more of RNA viruses, bacteria, fungi, and parasites.

[0117] A signal-processing module 104 configured to process the sequencing data into electrical-current signal data.

[0118] In some embodiments, the signal-processing module 104 performs preprocessing and segmentation on the sequencing data.

[0119] In some embodiments, the preprocessing comprises one or more of denoising, polishing, and normalization.

[0120] In some embodiments, the segmentation comprises segmenting electrical-current signal data of the barcode sequence into a 70–100 column matrix.

[0121] In some embodiments, the segmentation comprises segmenting electrical-current signal data of the barcode sequence into a 100-column matrix.

[0122] A classification module 106 configured to invoke a classification model constructed based on the barcode sequences in the test-sample data, classify the electrical-current signal data of the test-sample data, and output classification results of the n test-sample RNA polynucleotides.

[0123] In some embodiments, the sequencing system 100 further comprises a model-training module 108 configured to train electrical-current signal data of the training-sample data using a machine-learning algorithm to determine the classification model.

[0124] In some embodiments, the machine-learning algorithm comprises one or more of k-nearest neighbor, neural network, Naive Bayes classification, classification and regression tree, AdaBOOST, and random forest.

[0125] In some embodiments, the machine-learning algorithm comprises random forest.

[0126] In some embodiments, the training-sample data comprise sequencing data of an in vitro-transcribed RNA.

[0127] In some embodiments, the ratio of training set to testing set in the training-sample data is 7:3.

[0128] In some embodiments, the in vitro-transcribed RNA comprises one or more genes listed in Table 2.

[0129] In some embodiments, the model-training module 108 is configured to align electrical-current signal data of the training-sample data to corresponding reference sequences, and to filter out read sequences that do not satisfy alignment quality > 60 and alignment to a unique reference sequence.

[0130] In some embodiments, the sequencing system 100 further comprises a filtering module 110 configured to filter out reads having classification probabilities lower than a predetermined threshold of classification probability.

[0131] In some embodiments, the predetermined threshold of classification probability is greater than 0.3.

[0132] In some embodiments, the filtering module 110 and the classification module 106 may be integrated into a single analysis module 112.

[0133] 1.5 Barcode-Identification and Classification Workflow of DecodeR

[0134] In some embodiments, the specific workflow of DecodeR is as follows:

[0135] Based on the barcode-combination type used in the library-construction experiment, selecting an appropriate trained classification model and assigning it to the *model* parameter.

[0136] Selecting the raw electrical-signal data saved and filtered by the MinKNOW software—typically located in the *fast5_pass* directory—and assigning it to the *fast5* parameter.

[0137] Selecting a value within the range of 0–1 as a threshold for filtering the classification criteria, and assigning it to the *cutoff* parameter.

[0138] Selecting the number of CPU cores to be used (not exceeding the maximum number of cores available on the computational machine) and assigning it to the *NT* parameter.

[0139] In some embodiments, the default usage of DecodeR is as follows: DecodeR(*fast5*, *model*, *NT* = 1, *cutoff* = 0, *include.lowest* = FALSE). The output records each read ID, its corresponding barcode sequence, and the classification probability (probability value between 0 and 1), as shown in Table 5.

[0140] Table 5 Example of DecodeR Output

ReadID	Barcode	Classification Probability
read_0b3cabf5-44b4-4438-86df-4dfa672000e1	RTA-33	1
read_0d11e3c0-00ca-4531-a7fc-55fc75d3bb1f	RTA-33	1
read_10cefe5e-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] Based on the classification probability, reads with unreliable classification results are filtered out, and the original FAST5 electrical-signal data and FASTQ sequence data are classified according to the read ID.

[0142] EMBODIMENT 2

[0143] 2.1 Comparison of the DecodeR Algorithm with Other Algorithms

[0144] The accuracy of six different classification algorithms was evaluated using validation datasets, including k-nearest neighbour (KNN), neural network (NNET), Naïve Bayes (NB) classification, classification and regression tree (CART), AdaBOOST, and random forest (RF) (Table 6). Although all six methods achieved an area under the receiver operating characteristic curve (AUROC) above 0.92 (FIG. 1d) and an area under the precision–recall curve (AUPRC) above 0.83 (FIG. 1e), the random forest algorithm outperformed all other methods, achieving the highest AUROC (0.9961) and the highest AUPRC (0.9906) among the six methods (FIG. 1d, FIG. 1e). Additionally, the random forest algorithm showed significantly higher accuracy, sensitivity, specificity, precision, recall, and F1 score than the other five algorithms (FIG. 1f, Table 6). Therefore, the present invention selects the random forest–based classification model as the core of the sequencing method (DecodeR) (FIG. 1g), which is subsequently used for further training, testing, and application. Performance assessments and embodiment analyses were conducted for RNA from RNA viruses, bacteria, fungi, parasites, and tumor samples (FIG. 1h).

[0145] Table 6 Comparison of Various Algorithms

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
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 Comparison of DecodeR with Other Barcode Demultiplexing Methods

[0147] At present, two computational methods exist for barcode demultiplexing in direct RNA sequencing (DRS). These include DeePlexiCon (<https://github.com/Psy-Fer/deeplexicon>), which performs barcode classification through image-recognition based on a residual neural network classifier, and Poreplex (<https://github.com/hyeshik/poreplex>), which has not yet been formally published.

[0148] To compare the performance of DecodeR with these existing methods, this embodiment evaluated the demultiplexing accuracy and computation time of each method. First, five independent experiments were performed for eight barcodes ligated to IVT RNA using DeePlexiCon and Poreplex, generating five DRS raw electrical-signal datasets (Table 7). DecodeR was then trained using three datasets (training set), and the resulting model was used to predict barcode assignments for the remaining two datasets (validation set). Similarly, DeePlexiCon and Poreplex were used to perform barcode-demultiplexing predictions on the same validation datasets. Receiver operating characteristic (ROC) analysis, area under the curve

(AUC) metrics, and precision–recall curves were generated using the R package multiROC (<https://CRAN.R-project.org/package=multiROC>). CPU time was evaluated using the GNU time command, with user time and system time summed.

[0149] Table 7 Datasets for Comparing Demultiplexing Methods

Barcode	RTA ID	IVTRNA	Batch	Uniquely Aligned Reads	Application
D-BC1	RTA-00(ONT)	hGAPDH	1	16349	Validation
D-BC2	RTA-41	mActB	1	3634	Validation
D-BC3	RTA-42	ORF1ab	1	432000	Validation
D-BC4	RTA-43	NSP1	1	133636	Validation
D-BC1	RTA-00(ONT)	hGAPDH	2	57410	Training
D-BC2	RTA-41	mActB	2	2901	Training
D-BC3	RTA-42	ORF1ab	2	64469	Training
D-BC4	RTA-43	NSP1	2	38038	Training
P-BC1	RTA-44	mEef2	2	99137	Training
P-BC2	RTA-45	mGapdH	2	74356	Training
P-BC3	RTA-46	mRplp0	2	21677	Training
P-BC4	RTA-47	mTubala	2	58597	Training
D-BC1	RTA-00(ONT)	hGAPDH	3	2554	Training
D-BC2	RTA-41	mActB/hRPL13-3	3	1479	Training
D-BC3	RTA-42	ORF1ab	3	3947	Training
D-BC4	RTA-43	NSP1	3	2777	Training
P-BC1	RTA-44	mEef2	3	25004	Training
P-BC2	RTA-45	mGapdH	3	88221	Training
P-BC3	RTA-46	mRplp0	3	21655	Training
P-BC4	RTA-47	mTubala	3	35216	Training
D-BC1	RTA-00(ONT)	hGAPDH	4	12833	Training
D-BC2	RTA-41	hRPL13-3	4	5119	Training
D-BC3	RTA-42	ORF1ab	4	18938	Training
P-BC1	RTA-44	mTubala	4	11389	Training
P-BC2	RTA-45	mGapdH	4	32967	Training
P-BC3	RTA-46	mRplp0	4	8157	Training
P-BC4	RTA-47	mEef2	4	48263	Training
P-BC1	RTA-44	hGAPDH	5	167924	Validation
P-BC2	RTA-45	mActB	5	1563	Validation
P-BC3	RTA-46	ORF1ab	5	429767	Validation
P-BC4	RTA-47	NSP1	5	306156	Validation

[0150] Compared with the DeePlexiCon and Poreplex methods, DecodeR achieved a classification accuracy ranging from 95.82% to 97.83% (96.98 ± 0.84) when demultiplexing barcodes using the four Poreplex-preset barcodes, whereas Poreplex achieved a classification accuracy ranging from 83.40% to 93.18% (89.40 ± 4.21) (FIG. 2a; *unknown* indicates unclassified). DecodeR also demonstrated higher precision, recall, and F1 scores than Poreplex. Similarly, when demultiplexing the DeePlexiCon barcodes, DecodeR achieved a classification accuracy ranging from 90.57% to 94.43% (92.29 ± 1.60), whereas DeePlexiCon achieved 51.33% to 85.22% (68.39 ± 15.16) (FIG. 2b; *unknown* indicates unclassified). Furthermore, compared with the four-barcode multiplexing of DeePlexiCon, DecodeR achieved markedly higher sensitivity, recall, and F1 scores (FIG. 2b; Table 8). DecodeR produced a higher accuracy AUC value (0.997, compared with DeePlexiCon's 0.946), a higher precision–recall AUC value (0.993, compared with 0.888), and significantly higher classification accuracy and recovery rate (FIG. 2c, FIG. 2d, and FIG. 2e; Table 8). Finally, to evaluate demultiplexing efficiency, CPU time was calculated for DecodeR and DeePlexiCon when processing the same number (1–150K) of DRS reads. DecodeR was found to run 10-fold faster than DeePlexiCon (FIG. 2f).

[0151] Table 8 Comparison of DecodeR with Other Demultiplexing Methods

Method	Test Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
DecodeR (Use RTA of DeePlexiCon)	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 (Use RTA of Poreplex)	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 Performance of DecodeR

[0153] To evaluate the performance of DecodeR in demultiplexing direct RNA sequencing data, IVT RNA was used to assess its sensitivity, specificity, recall, and other metrics when demultiplexing 2, 4, 6, 8, 10, and up to 24 barcodes. The overall demultiplexing accuracy decreased from 98.4% to 92.2% as the number of barcodes increased (FIG. 3c), whereas all AUC values for specificity remained greater than 0.99 (FIG. 3a, FIG. 3b, Table 9). Although both precision and recovery rate declined as the number of barcodes increased from 2 to 24, they consistently remained at high levels (FIG. 3c, FIG. 3d).

[0154] Table 9 Performance of DecodeR under Different Barcode Conditions

Predicted Barcodes Numbers	Training Accuracy	Test Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
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] Given the excellent performance of DecodeR in demultiplexing DRS data from IVT RNA, DecodeR was next applied to real RNA samples isolated from pathogens, including two RNA viruses, Seneca Valley virus (SVV) and porcine reproductive and respiratory syndrome virus (PRRSV); two bacteria, *Escherichia coli* (E. coli) and *Salmonella enteritidis* (S. enteritidis); one fungus, *Saccharomyces cerevisiae*; and one parasite, *Plasmodium berghei*. Three independent DecodeR library-construction experiments and direct RNA sequencing runs were performed for each RNA sample, generating three batches of data. Receiver operating characteristic (ROC) analysis showed that all three replicate experiments exhibited high specificity AUC values (0.999, 0.998, 0.976), sensitivities of 99.2%, 93.3%, and 90.6%, specificities of 99.4%, 98.4%, and 98.4%, and maximum accuracies of 97.6%, 95.9%, and 92.5% (FIG. 4a, Table 10). DecodeR also demonstrated high precision, recall AUC values across all three batches (0.998, 0.994, 0.937) (FIG. 4b). With respect to the threshold (cutoff) and accuracy, we observed that higher cutoff values result in higher accuracy but lower recall (FIG. 4c). Sample-imbalance issues contributed to differences in precision and F1 scores across the models, and severe sample imbalance was associated with reduced precision and lower F1 scores (Table 10).

[0156] Table 10 Performance of DecodeR on Real Biological Samples

Batch	Accuracy	Sensitivity (Mean)	Specificity (Mean)	Precision (Mean)	Recall (Mean)	F1 Score (Mean)
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] EMBODIMENT 3

[0158] 3.1 Construction of RNA Virus Reference Genomes Using DecodeR

[0159] To evaluate whether DecodeR can facilitate the construction of viral genomes, RNA virus data obtained from three DecodeR library-construction experiments were aligned to reference genomes and analyzed for read-length distribution. DecodeR reads demonstrated the ability to almost completely cover the reference genomes of SVV and PRRSV (FIG. 5a, FIG. 5b), enabling rapid and straightforward assembly of RNA virus genomes. Specifically, **724 reads (9.68%, experiment 1)** and **111 reads (1.42%, experiment 2)** were near full-length SVV genomes (>7,000 nt), and **42 reads (2.26%)** were near full-length PRRSV genomes (>15,000 nt) (FIG. 5b). Next, single-nucleotide variants (SNVs) in the SVV genome were analyzed. A total of **414** and **423** SNVs were detected in experiment 1 and experiment 2, respectively, of which **411** were shared between both experiments (FIG. 5e), and the mutation ratios were highly reproducible (FIG. 5c). For example, the C-to-T substitution at position 5700 of SVV was detected in both experiments, with mutation frequencies of **0.98** and **0.99**, respectively (FIG. 5d). DecodeR also demonstrated the ability to detect high-frequency deletions—for instance, the 3-nt deletion at positions 4022–4024 of SVV was accurately identified in both

experiments (FIG. 5d). Overall, DecodeR enables cost-effective detection of pathogen RNA and has potential applications in identifying novel or mutated RNA viruses.

[0160] 3.2 Mutations and m6A Modifications Identified in Tumor Samples Using DecodeR

[0161] To illustrate potential applications of DecodeR in cancer research, DecodeR mixed-sample library construction and direct RNA sequencing were performed on four tumor samples, including diffuse midline glioma with H3K27M mutation (DMG) and glioblastoma (GBM). This yielded profiles of whole-transcriptome mutations, isoform identification, and RNA m⁶A modifications (FIG. 6a). In mutation-detection analysis, 17,587 SNVs were identified across the DMG and GBM samples, and 97–98% of them were validated by direct RNA sequencing (DRS) (FIG. 6b). Compared with next-generation sequencing (NGS) and Sanger sequencing, DecodeR exhibited similar accuracy in tumor-mutation detection. Among SNVs validated by Sanger sequencing, DecodeR and NGS showed highly correlated mutation frequencies ($R = 0.93$, Pearson correlation) (FIG. 6c). For example, the chr6:29913430:C>T mutation in the HLA-A gene exhibited high mutation frequencies in three DMG samples—0.87, 0.46, and 0.84 for DMG1, DMG2, and DMG3, respectively—and was absent in GBM. The same mutation was detected by NGS and Sanger sequencing with frequencies similar to DecodeR (FIG. 6d). Conversely, the chr16:2519933:C>A mutation in the ATP6V0C gene showed high mutation frequency in GBM but was nearly absent in all three DMG samples, and was also validated by NGS and Sanger sequencing (FIG. 6e). DecodeR also identified conserved deletion regions across tumor types, for example, a 5-nt deletion (chr9:12972603–12972607) in the SNORD137 gene consistently appeared in all tumor samples (FIG. 6f).

[0162] In m⁶A-modification analysis, DecodeR detected 2,158 high-confidence m⁶A sites across the DMG and GBM samples, of which 39% were supported by both DRS and the public m⁶A Atlas database, 60% were validated solely by DRS, and only 1% were not validated (FIG. 6g). DecodeR also showed strong consistency with DRS in gene-expression and m⁶A-detection analyses, with correlation values of 0.91 and 0.92, respectively (FIG. 6h). Overall, DecodeR enables reliable detection of tumor mutations and m⁶A RNA modifications, and the results are validated by Sanger sequencing, next-generation sequencing, and show high concordance with direct RNA sequencing.

[0163] From the above descriptions of the embodiments, those skilled in the art can clearly understand that the methods of the embodiments of the present invention may be implemented through software combined with necessary general-purpose hardware platforms, and may also be implemented purely in hardware, although the former is preferable in many cases. Based on this understanding, the technical solutions of the present invention—or the inventive contributions thereof—may be embodied in the form of a software product. The computer software product is stored in a storage medium (such as ROM, RAM, magnetic disk, or optical disk) and includes instructions for enabling a computer terminal (such as a mobile phone, computer, server, or network device) to execute the methods described in the embodiments of the present invention.

[0164] The above embodiments are described with reference to the accompanying drawings. However, the present invention is not limited to the specific embodiments described above. These embodiments are merely illustrative and not restrictive. Those skilled in the art, in light of the teachings of the present invention, may make various modifications or variations without departing from the spirit and scope of the invention

and the protection scope defined by the claims. All such modifications and variations fall within the protection scope of the present invention.

CLAIMS

Claim 1. A direct RNA nanopore sequencing method, wherein the method is used for sequencing n types of test-sample RNA polynucleotides, wherein $2 \leq n \leq 24$ and n is an integer; wherein the method comprises:

S1: for each test-sample RNA polynucleotide, providing a double-stranded DNA polynucleotide to capture a target RNA sequence and to form a first hybrid structure; the double-stranded DNA polynucleotide is a double-stranded structure assembled from a forward strand and a reverse strand of a reverse transcription adapter; the forward strand comprises, in the 5'-to-3' direction, a barcode sequence and a first ligation sequence; the reverse strand comprises, in the 5'-to-3' direction, a second ligation sequence, the barcode sequence, and a poly(T) sequence; the target RNA sequence comprises the test-sample RNA polynucleotide and a poly(A) sequence; the double-stranded DNA polynucleotide captures the target RNA sequence through the poly(T) sequence; the barcode sequence is specific to each test-sample RNA polynucleotide, and the barcode sequence is selected from any one of SEQ ID NOs: 1-24;

S2: reverse-transcribing the first hybrid structure to form a second hybrid structure;

S3: ligating a sequencing adapter to the second hybrid structure to form a third hybrid structure;

S4: performing nanopore sequencing on the third hybrid structure to obtain raw sequencing data;

S5: extracting electrical-current signal data corresponding to the raw sequencing data;

S6: segmenting electrical-current signal data of the barcode sequence from the electrical-current signal data, and analyzing the segmented electrical-current signal data using a classification model trained based on the barcode sequence to obtain sequence information of the barcode sequence, thereby obtaining classification information of the n test-sample RNA polynucleotides.

Claim 2. The method according to claim 1, wherein a training method of the classification model comprises segmenting electrical-current signal data of the barcode sequence in training-sample data, and training the classification model using a machine-learning algorithm based on the segmented electrical-current signal data of the barcode sequence to determine the classification model;

wherein the training-sample data comprises one or more training-sample RNA polynucleotides and corresponding electrical-current signal data of the barcode sequence; and wherein the barcode sequence is selected from any one of SEQ ID NOs: 1-24.

Claim 3. The method according to claim 1 or 2, wherein further comprising segmenting electrical-current signal data of the barcode sequence into a 70–100 column matrix.

Claim 4. The method according to claim 2, wherein the training-sample RNA polynucleotide comprises an in vitro-transcribed RNA.

Claim 5. The method according to claim 2, wherein the machine-learning algorithm comprises one or more selected from the group consisting of k-nearest neighbor, neural network, Naive Bayes classification, classification and regression tree, AdaBOOST, and random forest.

Claim 6. The method according to claim 1, wherein the source of the test-sample RNA polynucleotide comprises one or more selected from the group consisting of pathogens, plants, zooplankton, insects, mammals, and tumors.

Claim 7. The method according to claim 1, wherein the classification model applies a threshold of classification probability > 0.3 .

Claim 8. The method according to claim 1 or 2, wherein further comprising segmenting electrical-current signal data of the barcode sequence into a 100 column matrix.

Claim 9. The method according to claim 1, wherein Step S5 further comprises extracting raw electrical-current signal data corresponding to the raw sequencing data and preprocessing the raw electrical-current signal data to obtain preprocessed electrical-current signal data through a preprocessing method;

wherein the preprocessing method comprises one or more selected from the group consisting of denoising, polishing, and normalization.

Claim 10. The method according to claim 2, wherein the method further comprises aligning electrical-current signal data of the one or more training-sample RNA polynucleotides to corresponding reference sequences, and filtering out read sequences that do not satisfy conditions of an alignment quality greater than 60 and alignment to a unique reference sequence.

Claim 11. 11. A system for direct RNA sequencing, wherein comprising:
a sequencing-data receiving module, configured to receive and store sequencing data, the sequencing data comprising sequencing data of a double-stranded polynucleotide, the double-stranded polynucleotide comprising a first strand and a second strand;
the first strand comprising, in a 5'-to-3' direction, an RNA polynucleotide, a poly(A) sequence, a barcode sequence, a first ligation sequence, and a sequencing adapter;
the second strand comprising, in a 5'-to-3' direction, a sequencing adapter, a second ligation sequence, the barcode sequence, a poly(T) sequence, and a reverse-transcription sequence of the RNA polynucleotide;
the sequencing data comprising test-sample data and training-sample data;
the test-sample data comprising sequencing data corresponding to n test-sample RNA polynucleotides, wherein $2 \leq n \leq 24$ and n is an integer;
the barcode sequence being specific to each test-sample RNA polynucleotide, and the barcode sequence being selected from any one of SEQ ID NOs: 1-24;
a signal-processing module, configured to process the sequencing data into electrical-current signal data; and
a classification module, configured to invoke a classification model constructed based on the barcode sequence in the test-sample data, to classify electrical-current signal data of the test-sample data, and to output classification results of the n test-sample RNA polynucleotides.

Claim 12. The system according to claim 11, wherein the system further comprises a model-training module, the model-training module being configured to train electrical-current signal data of the training-sample data using a machine-learning algorithm to determine the classification model.

Claim 13. The system according to claim 11, wherein the signal-processing module performs a preprocessing method and a segmentation method on the sequencing data.

Claim 14. The system according to claim 13, wherein the preprocessing method comprises one or more selected from the group consisting of denoising, polishing, and normalization.

Claim 15. The system according to claim 13, wherein the segmentation method comprises segmenting electrical-current signal data of the barcode sequence into a 70–100 column matrix.

Claim 16. The system according to claim 12, wherein the machine-learning algorithm comprises one or more selected from the group consisting of k-nearest neighbor, neural network, Naive Bayes classification, classification and regression tree, AdaBOOST, and random forest.

Claim 17. The system according to claim 11, wherein the system further comprises a filtering module, configured to filter out reads having classification probabilities lower than a predetermined threshold of classification probability.

Claim 18. The system according to claim 11, wherein the training-sample data comprises sequencing data of an in vitro-transcribed RNA.

Claim 19. The system according to claim 11, wherein a source of the test-sample RNA polynucleotide comprises one or more selected from the group consisting of pathogens, plants, zooplankton, insects, mammals, and tumors.

Claim 20. The system according to claim 17, wherein the predetermined threshold of classification probability is greater than 0.3.