



(51) International Patent Classification:

C12Q 1/68 (2018.01) G06F 19/22 (2011.01)

(21) International Application Number:

PCT/US2018/023191

(22) International Filing Date:

19 March 2018 (19.03.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/473,935 20 March 2017 (20.03.2017) US

(71) Applicant: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; 1111 Franklin Street, 5th Floor, Oakland, California 94607 (US).

(72) Inventor; and

(71) Applicant: ETHERIDGE, Alton [US/US]; c/o Pacific Northwest Research Institute, 720 Broadway, Seattle, Washington 98122 (US).

(72) Inventors: LAURENT, Louise; c/o UC San Diego, 9500 Gilman Drive, Mail Code: 0910, La Jolla, California 92093 (US). DE HOFF, Peter; c/o UC San Diego, 9500 Gilman Drive, Mail Code: 0910, La Jolla, California 92093 (US). ERLE, David; c/o UC San Francisco, 505 Parnassus Avenue, San Francisco, California 94143 (US). GODOY, Paula; c/o UC San Francisco, 505 Parnassus Avenue, San Francisco, California 94143 (US).

(74) Agent: KONSKI, Antoinette F. et al.; FOLEY & LARDNER LLP, 3000 K Street, N.W., Suite 600, Washington, District of Columbia 20007-5109 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))
- as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

(54) Title: VALIDATED SMALL RNA SPIKE-IN SET FOR EXRNA ANALYSIS

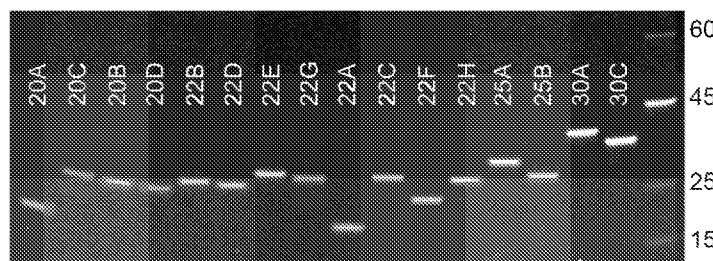


FIGURE 4

(57) Abstract: Aspects of this disclosure relate to artificial short RNA sequences that (1) have a base composition that mimics that of endogenous human miRNAs; (2) have broad sequence diversity; (3) cover a range of sequence lengths of between about 16 and 70 nucleotides; (4) do NOT share sequence identity with known endogenous sequences in human, animal, plant, fungus, bacterial, and/or viral genomes, and methods of use thereof.



VALIDATED SMALL RNA SPIKE-IN SET FOR exRNA ANALYSIS**CROSS REFERENCE TO RELATED APPLICATIONS**

[0001] This application claims priority under 35 U.S.C. 119(e) to U.S. Serial No. 62/473,935, filed March 20, 2017, the entirety of which is incorporated by reference herein.

STATEMENT REGARDING GOVERNMENT SUPPORT

[0002] This invention was made with government support under Grant No. U01HL126494-01 awarded by the National Center for Advancing Translational Sciences at the National Institutes of Health. The government has certain rights in the invention.

BACKGROUND

[0003] Extracellular RNA (exRNAs) in human biofluids are of great interest for many of the groups within the Extracellular RNA Communication Consortium (ERCC) as a largely untapped source of accurate prognostic and diagnostic biomarkers for the human disease states being investigated by the Consortium's research groups. However, accurate quantification of the exRNAs contained in these biofluids, an essential starting point for biomarker discovery, is confounded by the broad range of expression levels for different exRNAs and the variability introduced by the multiple "wet lab" steps involved in obtaining exRNA quantification data. Specifically, biases can be introduced during the RNA extraction and the RNA preparation steps used in downstream processing. Moreover, different biases can be encountered not only among different analytical methods (e.g. hybridization vs. qRT-PCR vs. next-generation sequencing), but also within a single method, depending on the chemistry used. This problem is compounded when attempting to compare expression patterns across experiments and/or labs, as robust "housekeeping" exRNAs for normalization have not been identified, and are not likely to exist.

[0004] Academic and commercially available small and long RNA standards are available (Exiqon (Denmark), Miltenyi Biotec (San Diego, CA)), but these existing standards are either biochemically incompatible with small RNA-seq pipelines or are not sufficiently adaptable to

address all of the challenges associated with next generation sequencing (NGS)-based exRNA quantification workflows.

[0005] The Exiqon small RNA spike-in set is the most comparable commercially available product that attempts to address some of the QC and normalization challenges associated with small RNA NGS discovery work. This small RNA-seq spike-in set consists of 52 ~20-21nt small RNAs, mixed at a range of concentrations, to be added at either the RNA isolation or at the library generation phase. Despite claims found in written materials provided by the manufacturer, a bioinformatic analysis of the Exiqon small RNA spike-in set on the standardized Genboree Small RNA-seq pipeline developed by the ERCC Data Management and Resource Repository (DMRR) shows that 16 of these spike-in RNAs directly match previously identified human miRNAs, and all of them match miRNAs found in other species.

[0006] The Milteyi Biotec spike-ins are intentionally identical to known human, mouse, rat and viral miRNA sequences found within miRbase, are appropriate for use within a microarray setting, and thus have limited use as spike-in controls in NGS small RNA discovery work. The External RNA Controls Consortium spike-ins³ and the Sequins² are restricted to long RNAs (>200nt) and thus are biochemically incompatible with most small RNA-seq NGS library preparation pipelines.

[0007] The spike-in set developed by the Tuschl lab consists of two equimolar pools of 10 synthetic 22nt small RNAs with no matches to the human or mouse genomes. One pool of 10 oligos is added during RNA isolation and the second pool is added at the beginning of library preparation. These pools allow for QC of both the RNA isolation and library preparation steps. However, the pools consist of relatively few oligos and the size is appropriate only for monitoring recovery of miRNA-sized fragments as opposed to other classes of small RNAs. In addition, it is unclear whether any oligos in the set have matches to sequences in databases beyond the human and mouse genome, which could make them problematic for use in libraries where exogenous RNAs are of interest.

SUMMARY

[0008] For use in exRNA analysis, it is essential that none of the spike-in standards match human miRNA sequences, but it is also important that they not match other sequences in the

human genome, as exRNAs include fragments of both coding and non-coding long RNAs. Furthermore, as this type of discovery work is starting to encompass exogenous RNAs from non-human organisms found in human biofluids, it is important that sequences from non-human organisms also be avoided.

[0009] Despite the limitations of prior methods, some aspects of the criteria used in the design and usage of these available standards can provide guidance in how to most effectively develop and utilize small RNA spike-in standards for NGS. Applicants propose that properly prepared small RNA spike-in standards can be used to both standardize exRNA expression across experiments, as well as to provide quality control (QC) metrics for the various procedural steps in sample preparation.

[0010] Aspects of this disclosure relate to artificial short RNA sequences that (1) have a base composition that mimics that of endogenous human miRNAs; (2) have broad sequence diversity; (3) cover a range of sequence lengths of between about 16 to 70 nucleotides; and/or (4) do not share sequence identity with known endogenous sequences in human, animal, plant, fungus, bacterial, and/or viral genomes and methods of identification thereof. In some embodiments, the artificial short RNA sequences have a nucleotide frequency, optionally GC content, of approximately 25% for each nucleotide, which mimics the nucleotide frequency of endogenous human miRNAs in miRBase. In some embodiments, the artificial short RNA sequences have a dinucleotide frequency that mimics the dinucleotide frequency of endogenous human miRNAs in miRBase. In some embodiments, the artificial short RNA sequences are selected from the group of artificial short RNA sequences provided in **Table 3** or **Figure 2**. In some embodiments, the artificial short RNA is detectably labeled, optionally 5' phosphorylated.

[0011] A number of aspects relate to the use of these RNAs as “spike-ins.” In some of these embodiments, the aim is 0.5-1% or 1-5% spike-in reads per library (i.e., for a 10M total read library, there are a total of 100,000 calibrator reads). In some embodiments, this is accomplished using a starting point of 5% spike-in with ongoing experiments to optimize this value. Related aspects relate to a pool of artificial short RNA sequences disclosed above, referred to herein as a “spike-in set” or a “set of spike-in RNAs.” In some embodiments, the set of spike-in RNAs comprises one or more of the artificial short RNA sequences, optionally

between about 10 to 100 sequences. In some embodiments, the set of spike-in RNAs all have the same length. In some embodiments, the set of spike-in RNAs have a range of lengths from 16 to 70 nucleotides.

[0012] Further aspects of the disclosure relate to the use of these sequences in the analysis of small extracellular RNAs (exRNA) – such as, but not limited to, the RNAs in exosomes and other extracellular vesicles – and/or to quantify the amount of exRNA in a sample or to normalize exRNA expression level data between different samples.

[0013] Some aspects relate to the addition of the artificial short RNA sequences to biofluids and/or RNA samples at different steps of experimental procedures or the normalization of endogenous RNA among different samples. Expression levels of these artificial short RNA sequences can be determined using qRT-PCR, hybridization or next-generation sequencing (NGS). In some embodiments, equimolar or ratiometric molar concentrations of the spike ins are used.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] **FIG. 1** depicts an exemplary 32nt oligo phylogenetic tree. Red Line: Groups (6). Blue Square: Chosen oligo.

[0015] **FIG. 2** lists spike-in sequences. These sequences were run through the exceRpt pipeline and had no alignments to known endogenous nor to known exogenous (non-human) sequences. RNA oligos were ordered from IDT with a 5' phosphate and HPLC purification. The RNA oligos were diluted to 1 μ M and pooled into an A and a B pool.

[0016] **FIG. 3** provides a characterization of libraries made with the spike-ins. For plasma+spike-in libraries, set A oligos were added to Qiazol at indicated amounts before adding to the sample to extract the RNA.

[0017] **FIG. 4** shows spike-in oligos on a 10% Acrylamide TBE-urea gel.

[0018] **FIG. 5** shows measurements of the total miRNA concentration in a female plasma pool based on known spike-in quantity added to the plasma pool during RNA isolation.

DETAILED DESCRIPTION

[0019] Unless otherwise defined, all terms (including technical and scientific terms) used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. It will be further understood that terms, such as those defined in commonly used dictionaries, should be interpreted as having a meaning that is consistent with their meaning in the context of the present application and relevant art and should not be interpreted in an idealized or overly formal sense unless expressly so defined herein. While not explicitly defined below, such terms should be interpreted according to their common meaning.

[0020] The terminology used in the description herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. All publications, patent applications, patents and other references mentioned herein are incorporated by reference in their entirety.

[0021] The practice of the present technology will employ, unless otherwise indicated, conventional techniques of tissue culture, immunology, molecular biology, microbiology, cell biology, and recombinant DNA, which are within the skill of the art.

[0022] Unless the context indicates otherwise, it is specifically intended that the various features of the invention described herein can be used in any combination. Moreover, the disclosure also contemplates that in some embodiments, any feature or combination of features set forth herein can be excluded or omitted. To illustrate, if the specification states that a complex comprises components A, B and C, it is specifically intended that any of A, B or C, or a combination thereof, can be omitted and disclaimed singularly or in any combination.

[0023] Unless explicitly indicated otherwise, all specified embodiments, features, and terms intend to include both the recited embodiment, feature, or term and biological equivalents thereof.

[0024] All numerical designations, e.g., pH, temperature, time, concentration, and molecular weight, including ranges, are approximations which are varied (+) or (-) by increments of 1.0 or 0.1, as appropriate, or alternatively by a variation of +/- 15 %, or

alternatively 10%, or alternatively 5%, or alternatively 2%. It is to be understood, although not always explicitly stated, that all numerical designations are preceded by the term “about”. It also is to be understood, although not always explicitly stated, that the reagents described herein are merely exemplary and that equivalents of such are known in the art.

Definitions

[0025] As used in the description of the invention and the appended claims, the singular forms “a,” “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise.

[0026] The term “about,” as used herein when referring to a measurable value such as an amount or concentration and the like, is meant to encompass variations of 20%, 10%, 5%, 1 %, 0.5%, or even 0.1 % of the specified amount.

[0027] The terms or “acceptable,” “effective,” or “sufficient” when used to describe the selection of any components, ranges, dose forms, etc. disclosed herein intend that said component, range, dose form, etc. is suitable for the disclosed purpose.

[0028] Also as used herein, “and/or” refers to and encompasses any and all possible combinations of one or more of the associated listed items, as well as the lack of combinations when interpreted in the alternative (“or”).

[0029] As used herein, the term “comprising” is intended to mean that the compositions and methods include the recited elements, but do not exclude others. As used herein, the transitional phrase “consisting essentially of” (and grammatical variants) is to be interpreted as encompassing the recited materials or steps “and those that do not materially affect the basic and novel characteristic(s)” of the recited embodiment. *See, In re Herz*, 537 F.2d 549, 551-52, 190 U.S.P.Q. 461, 463 (CCPA 1976) (emphasis in the original); *see also* MPEP § 2111.03. Thus, the term “consisting essentially of” as used herein should not be interpreted as equivalent to “comprising.” “Consisting of” shall mean excluding more than trace elements of other ingredients and substantial method steps for administering the compositions disclosed herein. Aspects defined by each of these transition terms are within the scope of the present disclosure.

[0030] The term “encode” as it is applied to nucleic acid sequences refers to a polynucleotide which is said to “encode” a polypeptide if, in its native state or when manipulated by methods well known to those skilled in the art, can be transcribed and/or translated to produce the mRNA for the polypeptide and/or a fragment thereof. The antisense strand is the complement of such a nucleic acid, and the encoding sequence can be deduced therefrom.

[0031] The terms “equivalent” or “biological equivalent” are used interchangeably when referring to a particular molecule, biological, or cellular material and intend those having minimal homology while still maintaining desired structure or functionality.

[0032] “Homology” or “identity” or “similarity” refers to sequence similarity between two peptides or between two nucleic acid molecules. Homology can be determined by comparing a position in each sequence which may be aligned for purposes of comparison. When a position in the compared sequence is occupied by the same base or amino acid, then the molecules are homologous at that position. A degree of homology between sequences is a function of the number of matching or homologous positions shared by the sequences. As used herein, referring to a sequence that “does not share identity” intends that a sequence (oligo) shares less than 100% identity with or alternatively has one or more mismatches when compared to another sequence across the length of the sequence (oligo).

[0033] The term “ortholog” is used in reference of another gene or protein and intends a homolog of said gene or protein that evolved from the same ancestral source. Orthologs may or may not retain the same function as the gene or protein to which they are orthologous.

[0034] As used herein, the term “expression” refers to the process by which polynucleotides are transcribed into mRNA and/or the process by which the transcribed mRNA is subsequently being translated into peptides, polypeptides, or proteins. If the polynucleotide is derived from genomic DNA, expression may include splicing of the mRNA in a eukaryotic cell. The expression level of a gene may be determined by measuring the amount of mRNA or protein in a cell or tissue sample; further, the expression level of multiple genes can be determined to establish an expression profile for a particular sample.

[0035] As used herein, the term “functional” may be used to modify any molecule, biological, or cellular material to intend that it accomplishes a particular, specified effect.

As used herein, the terms “nucleic acid sequence” and “polynucleotide” are used interchangeably to refer to a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides. Thus, this term includes, but is not limited to, single-, double-, or multi-stranded DNA or RNA, genomic DNA, cDNA, DNA-RNA hybrids, or a polymer comprising purine and pyrimidine bases or other natural, chemically or biochemically modified, non-natural, or derivatized nucleotide bases.

[0036] The terms “ribonucleic acid” or “RNA” are used interchangeably to refer to a nucleic acid comprising a ribonucleic acid backbone. Canonically, bases adenine (A), uracil (U), guanine (G), and cytosine (C) are found in RNA.

[0037] The term “protein”, “peptide” and “polypeptide” are used interchangeably and in their broadest sense to refer to a compound of two or more subunits of amino acids, amino acid analogs or peptidomimetics. The subunits may be linked by peptide bonds. In another aspect, the subunit may be linked by other bonds, e.g., ester, ether, etc. A protein or peptide must contain at least two amino acids and no limitation is placed on the maximum number of amino acids which may comprise a protein’s or peptide’s sequence. As used herein the term “amino acid” refers to either natural and/or unnatural or synthetic amino acids, including glycine and both the D and L optical isomers, amino acid analogs and peptidomimetics.

[0038] The term “spike-in” refers to a nucleic acid sequence, *e.g.* RNA, added to a sample as a control to assess performance of a nucleic acid quantification technology such as qPCR, next generation sequence (NGS), or a microarray. The term “microarray” refers to a collection of nucleic acids attached to a surface used to measure expression level of multiple genes simultaneously.

[0039] As used herein, the term “subject” is intended to mean any animal. In some embodiments, the subject may be a mammal; in further embodiments, the subject may be a human, mouse, or rat.

[0040] The term “tissue” is used herein to refer to tissue of a living or deceased organism or any tissue derived from or designed to mimic a living or deceased organism. The tissue may be healthy, diseased, and/or have genetic mutations. The biological tissue may include any single tissue (e.g., a collection of cells that may be interconnected) or a group of tissues making up an organ or part or region of the body of an organism. The tissue may comprise a

homogeneous cellular material or it may be a composite structure such as that found in regions of the body including the thorax which for instance can include lung tissue, skeletal tissue, and/or muscle tissue. Exemplary tissues include, but are not limited to those derived from liver, lung, thyroid, skin, pancreas, blood vessels, bladder, kidneys, brain, biliary tree, duodenum, abdominal aorta, iliac vein, heart and intestines, including any combination thereof.

[0041] As used herein, the term “biofluid” refers to a biological fluid, such as but not limited to those excreted (*e.g.* urine or sweat), secreted (*e.g.* breast milk or bile), circulating (*e.g.* blood, blood components such as plasma, or cerebrospinal fluid), and/or developed as a result of a pathological process in a subject (*e.g.* pus or other blister or cyst fluids). Non-limiting examples of such fluids found in a human subject include amniotic fluid, aqueous humour, vitreous humour, bile, blood, blood plasma, blood serum, breast milk, cerebrospinal fluid, cerumen (earwax), chyle, chime, endolymph, perilymph, exudates, feces, ejaculate (male or female), gastric acid, gastric juice, lymph, mucus, pericardial fluid, peritoneal fluid, pleural fluid, pus, rheum, saliva, sebum, serous fluid, semen, smegma, sputum, synovial fluid, sweat, tears, urine, vaginal secretion, vomit, intracellular fluid, extracellular fluid, intravascular fluid, interstitial fluid, lymphatic fluid, and transcellular fluid. It is appreciated that the fluid may be from a subject of any age, such as but not limited to an adult, child, infant, neonate, or fetus. Thus, the biofluid may optionally be one or more of amniotic fluid, cerebrospinal fluid, adult serum, adult plasma, cord blood plasma, cord blood serum, bronchoalveolar lavage fluid, saliva, sputum, and adult urine.

[0042] As used herein, “treating” or “treatment” of a disease in a subject refers to (1) preventing the symptoms or disease from occurring in a subject that is predisposed or does not yet display symptoms of the disease; (2) inhibiting the disease or arresting its development; or (3) ameliorating or causing regression of the disease or the symptoms of the disease. As understood in the art, “treatment” is an approach for obtaining beneficial or desired results, including clinical results. For the purposes of the present technology, beneficial or desired results can include one or more, but are not limited to, alleviation or amelioration of one or more symptoms, diminishment of extent of a condition (including a disease), stabilized (*i.e.*, not worsening) state of a condition (including disease), delay or slowing of condition (including disease), progression, amelioration or palliation of the

condition (including disease), states and remission (whether partial or total), whether detectable or undetectable.

[0043] As used herein, the term “vector” intends a recombinant vector that retains the ability to infect and transduce non-dividing and/or slowly-dividing cells and integrate into the target cell’s genome. The vector may be derived from or based on a wild-type virus. Aspects of this disclosure relate to an adeno-associated virus vector.

[0044] As used herein, the term “label” intends a directly or indirectly detectable compound or composition that is conjugated directly or indirectly to the composition to be detected, e.g., N-terminal histidine tags (N-His), magnetically active isotopes, e.g., ^{115}Sn , ^{117}Sn and ^{119}Sn , a non-radioactive isotopes such as ^{13}C and ^{15}N , polynucleotide or protein such as an antibody so as to generate a “labeled” composition. The term also includes sequences conjugated to the polynucleotide that will provide a signal upon expression of the inserted sequences, such as green fluorescent protein (GFP) and the like. While the term “label” generally intends compositions covalently attached to the composition to be detected, it specifically excludes naturally occurring nucleosides and amino acids that are known to fluoresce under certain conditions (*e.g.* temperature, pH, etc.) and generally any natural fluorescence that may be present in the composition to be detected. The label may be detectable by itself (*e.g.* radioisotope labels or fluorescent labels) or, in the case of an enzymatic label, may catalyze chemical alteration of a substrate compound or composition which is detectable. The labels can be suitable for small scale detection or more suitable for high-throughput screening. As such, suitable labels include, but are not limited to magnetically active isotopes, non-radioactive isotopes, radioisotopes, fluorochromes, chemiluminescent compounds, dyes, and proteins, including enzymes. The label may be simply detected or it may be quantified. A response that is simply detected generally comprises a response whose existence merely is confirmed, whereas a response that is quantified generally comprises a response having a quantifiable (*e.g.*, numerically reportable) value such as an intensity, polarization, and/or other property. In luminescence or fluorescence assays, the detectable response may be generated directly using a luminophore or fluorophore associated with an assay component actually involved in binding, or indirectly using a luminophore or fluorophore associated with another (*e.g.*, reporter or indicator) component. Examples of luminescent labels that produce signals include, but are not limited to bioluminescence and chemiluminescence. Detectable

luminescence response generally comprises a change in, or an occurrence of a luminescence signal. Suitable methods and luminophores for luminescently labeling assay components are known in the art and described for example in Haugland, Richard P. (1996) Handbook of Fluorescent Probes and Research Chemicals (6th ed). Examples of luminescent probes include, but are not limited to, aequorin and luciferases.

Modes of Carrying out the Disclosure

[0045] Embodiments according to the present disclosure will be described more fully hereinafter. Aspects of the disclosure may, however, be embodied in different forms and should not be construed as limited to the embodiments set forth herein. Rather, these embodiments are provided so that this disclosure will be thorough and complete, and will fully convey the scope of the invention to those skilled in the art. The terminology used in the description herein is for the purpose of describing particular embodiments only and is not intended to be limiting.

Generation of Artificial Short RNA Sequences

[0046] The sequence space represented by endogenous small RNAs (miRNA, piRNA, tRNA fragments, Y RNA fragments, etc.) is largely limited to <75 nt. This increases the probability of a randomly generated small RNA standard either being mis-called as an endogenously expressed small RNA, or that the base composition of the synthetic sequences may diverge excessively from the known small RNA sequence space to the point that it reduces the user's ability to make reliable inferences on relative small RNA target expression.

[0047] To address this, aspects of this disclosure relate to artificial short RNA sequences that (1) have a base composition that mimics that of endogenous human miRNAs; (2) have broad sequence diversity; (3) cover a range of sequence lengths of between about 16 to 70 nucleotides; and/or (4) do not share sequence identity with known endogenous sequences in human, animal, plant, fungus, bacterial, and/or viral genomes and methods of identification thereof. Non-limiting exemplary artificial short RNA sequences include those provided in **Table 3** or **Figure 2**. In some embodiments, the sequences further comprise degenerate sequences thereof. Said degenerate sequences meet the criteria (1) to (4) noted above and vary from the artificial short RNA sequence, optionally by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, or 16 nucleotides. The invariant sequence of the degenerate sequence comprises

the portion of the base composition that mimics that of endogenous miRNA from the artificial short RNA sequence and may be contiguous or non-contiguous.

Mimicking Endogenous Human miRNA

[0048] Techniques of designing artificial short RNA sequences that have a base composition that mimics that of endogenous human miRNAs are generally based on analytics from miRNA databases. miRBase is the art-recognized standard database that contains high-quality human small RNAs, many of which are grouped into families with similar seed sequences. However, it is appreciated that any suitable miRNA database may be used for the purposes disclosed herein. To generate artificial short RNA sequences that mimic endogenous human miRNA, exemplary miRNAs from each family can be sampled to gain a representative population distribution of the di-nucleotide frequencies for naturally occurring human miRNAs. Using these miRNAs as a template for random small RNA generation can yield populations with similar di-nucleotide frequencies as those found within the biologically relevant source, and reduce the chances of highly divergent and less useful small RNAs for inclusion into the spike-in sets. This matched di-nucleotide frequency distribution can extend to synthetic spike-in oligos of any size, permitting detailed quality control validation of various aspects of RNA extraction and/or the size selection portion of the NGS library generation. This process has the potential to reduce the occurrence of “jackpotting” where a single spike-in oligo sequences particularly well and reduces signal of the other oligo standards. This “jackpotting” phenomenon has been previously observed on Ion Torrent (Thermo Scientific) based instruments, and on Illumina-based instruments.

[0049] A novel sequence can be generated in such a way as to sample the frequency of any target database of sequences. This database of sequences can be of an entire species, sequences from multiple species, or subsets of sequences of either. Specifically, but not limited to, whole or truncated versions of mRNAs, miRNAs, piRNAs, tRNAs, rRNAs, YRNAs, DNAs, peptides, amino acids, or other sequences. The base pair frequency can be of single nucleotide, di-nucleotide, tri-nucleotide, or any number of nucleotides or amino acids as desired by the user. The resultant sequence can be screened for any number of sequence motifs of any length. These motifs consist of consecutive strings of nucleotides or amino acids whose identities are provided within the IUPAC codes for nucleotides or amino

acids, which additionally represents subsets of specific nucleotides or amino acids at each position within a sequence. These sequence motifs can be homopolymers of two nucleotides, three nucleotides, four nucleotides, or any number of nucleotides or amino acids. These sequence motifs can be patterns of two, three, four, or any number of sequences. These motifs can include repeats of motifs of two, three, four, or any number.

[0050] In one embodiment, the overrepresentation of related sequences within the database can be reduced through a pre-clustering of the sequences within the database into groups or families. This clustering can be performed de novo through a number of sequence similarity discovery algorithms, such as, but not limited to, those embodied by the BLAST, BLAT, MUSCLE, ClustalW programs, or by the Smith-Waterman algorithm, or by the Needleman-Wunsch algorithm. In certain cases, this clustering into families is available through the metadata associated with each sequence in the database, such as the naming system within miRBase. In the case of miRNAs, reducing sequence overrepresentation is important to ensure increased sequence diversity within a selected oligo pool and reduces the impact of miRNA family expansion that has occurred over evolutionary timescales within eukaryotic species. In all cases, the database of sequences is catenated into a single string of nucleotide or amino acid sequences with a series of nonsense text markers (ie “N” or ”X”) placed in between each sequence. The user designates the length of the stretch of nucleotides or amino acids for which they desire in their generated sequence to possess the same frequency representation as that found within their database of sequences (ie the user chooses a single nucleotide, or two nucleotides, or three nucleotides, or any number of nucleotides or amino acids). The user also selects the final length for their desired randomly-generated oligo. The length of that oligo is evenly divisible by the length of the stretch of nucleotides or amino acids. For instance, a selected length of three nucleotides (tri-nucleotide) will have a final randomly samples oligo of length 3, 6, 9, 12, nucleotides or any number of nucleotides evenly divisible by three.

[0051] A random number is generated, and that position is chosen within the catenated string of nucleotides or amino acids and a number of nucleotides or amino acids after that random position, of the length of nucleotides or amino acids chosen by the user, minus 1, is selected. If this set of nucleotides or amino acids contains the nonsense text marker (ie “N”

or "X") then the entire selected sequence is discarded, and the process is restarted with a new random number. If the selected sequence does not contain the nonsense marker, it is screened against the previously chosen motifs (repeats, homopolymers, etc) and discarded if it matches any of those sequences and the process repeats with a new randomly generated number. If the chosen oligo does not fail on the prior two filtering steps, it is selected for catenation. This process is repeated with a newly generated random number and the oligos that pass the filters are catenated to the previous oligo. This newly generated oligo is then filtered against the undesired motifs, and if those are found, the previously catenated oligos are removed. This process is repeated until the target number of oligos of the specified length that pass the filters are generated.

[0052] In an exemplary embodiment, the artificial short RNA sequences have a GC content of approximately 25% for each nucleotide, which mimics the nucleotide frequency of endogenous human miRNAs in miRBase.

Having a Broad Sequence Diversity

[0053] To assess "sequence diversity" a variety of algorithms may be used for alignment and subsequent diversity scoring. For example, alignment may be performed using any suitable alignment program, including but not limited to a MUSCLE algorithm. Sequence diversity can likewise be analyzed using a suitable model such as the Maximum Composite Likelihood method with Substitutions to include Transitions+Transversions in MEGA 7, assuming a uniform pattern among lineages. It is appreciated that sequences having a diversity metric calculated by this method above 4.00 have suitably "broad" sequence diversity. An ordinary skilled artisan can appreciate that equivalents to this 4.00 threshold are contemplated should an alternative, equivalent model be used and would appreciate how to convert the threshold value in view of the model and corresponding assumptions.

[0054] To assure broad sequence diversity within a pool of short artificial RNA sequences, phylogenetic trees can be generated to group sequences of the same length into families. Sequences can then be chosen based on the generation of phylogenetic trees. For example, sequences can be chosen based on their position in a phylogenetic tree in such a way to both maximize inter-sequence diversity and representation of the total diversity within the pool of

sequences. The program generating the tree, *e.g.* MEGA7, can be configured so that approximately the same number of related sequences are categorized into each subtree. To generate a pool having broad sequence diversity, the sequence with the highest degree of similarity to the other sequences within its subtree can be selected from each subtree. Not to be bound by theory, this sequence will likely be the basal sequence in the subtree.

Having a Sequence lengths of between about 16 to 70 nucleotides

[0055] It is appreciated that the short artificial RNA sequences disclosed herein may vary in length between 16 to 70 nucleotides, *i.e.*, having at least 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, but not more than 70 nucleotides; and/or at most 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, but not less than 16 nucleotides; and/or having 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, or 70 nucleotides.

Not Sharing Sequence Identity with known Endogenous Sequences

[0056] It is appreciated that methods of comparing sequences to determine identity are well known in the art. Thus, an ordinary skilled artisan is aware of a variety of algorithms and databases through which to determine whether a sequence has identity with a plurality of endogenous sequences.

[0057] For example, the Extracellular RNA Communication Consortium (ERCC) has developed a comprehensive and standardized small RNA seq mapping pipeline available through Genboree (www.genboree.org), called the exceRpt pipeline. This pipeline can be utilized to screen randomly generated synthetic putative small RNA sequences to remove those that match the human genome (endogenous) or known non-human (exogenous) sequences. This greatly reduces the chance that a spike-in sequence will be miscalled as a relevant small RNA within a sample. By processing the sequences through Genboree exceRpt small RNA mapping pipeline against endogenous (human) and exogenous reference libraries

at a stringency of 0 allowed mismatches, rather than using the default of 1 allowed mismatched, sequences that do not share sequence identity to the sequences in the data base can be identified.

[0058] Similar comparisons can be run against comparable datasets of small RNA sequences, such as those from the Erle or Galas groups and/or those available on one of the other databases known in the art. Thus, generated oligo sequences can be curated to avoid mapping to human, animal, plant, fungus, bacterial, and/or viral genomes based on the selected data set. For example, the Erle and Galas group datasets provide those exRNAs found in biofluids, such as amniotic fluid, cerebrospinal fluid, adult serum, adult plasma, cord blood plasma, bronchoalveolar lavage fluid, saliva, sputum, and adult urine.

Spike-In Sets

[0059] A number of aspects relate to the use of these RNAs as “spike-ins.” In some of these embodiments, the aim is 0.5-1% or 1-5% spike-in reads per library (i.e., for a 10M total read library, there are a total of 100,000 spike-in reads). In some embodiments, this is accomplished using a starting point of 5% spike-in with ongoing experiments to optimize this value. Related aspects relate to a pool of artificial short RNA sequences that (1) have a base composition that mimics that of endogenous human miRNAs; (2) have broad sequence diversity; (3) cover a range of sequence lengths of between about 16 to 70 nucleotides; and/or (4) do not share sequence identity with known endogenous sequences in human, animal, plant, fungus, bacterial, and/or viral genomes – *i.e.* those artificial short RNA sequences disclosed herein above. The pool of these sequences is referred to herein as a “spike-in set” or a “set of spike-in RNAs.”

[0060] As disclosed above, pools may be generated by the same methods. In some embodiments, the pools may comprise between about 2 and 100 unique artificial short RNA sequences, optionally at least 50 unique artificial short RNA sequences selected according to the metrics disclosed herein above. In some embodiments, the set of spike-in RNAs comprises one or more of the artificial short RNA sequences, optionally at least 5 sequences, at least 10 sequences, at least 15 sequences, at least 20 sequence, at least 25 sequence, at least 30 sequence, at least 40 sequences, at least 45, at least 50 sequences, at least 55 sequences, at

least 60 sequences, at least 65 sequence, and least 70 sequences, at least 75 sequences, at least 80 sequences, at least 85 sequences, at least 90 sequences, at least 95 sequences, or at least 100 sequences. In some embodiments, the pools may comprise a unique artificial short RNA sequence and degenerate sequences thereof. Said degenerate sequences meet the criteria (1) to (4) noted above and vary from the artificial short RNA sequence, optionally by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, or 16 nucleotides. The invariant sequence of the degenerate sequence comprises the portion of the base composition that mimics that of endogenous miRNA from the artificial short RNA sequence and may be contiguous or non-contiguous.

[0061] In some embodiments, the set of spike-in RNAs all have the same length. Accordingly, in some embodiments all of the sequences have 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, or 70 nucleotides. In some embodiments, the length is dependent on the target biofluid or desired RNA class. For example, for plasma miRNA, the target range would be 20 to 24 nucleotides.

[0062] Not to be bound by theory, Applicants suspect that the proposed spike-in oligo sets will be the most comprehensive available for human small RNA-seq experiments and have high value both within the U01 laboratories, as well as across the larger scientific community. Bioinformatic exclusion of known human and non-human sequences within the synthetic pools reduces the chance of misidentifying reads from a spike-in standard as relevant small RNAs, and will be more stringent than currently available oligo spike-in sets. The dinucleotide selection procedure in the synthetic oligo generation process can better match the nucleotide biases seen in miRbase families of small RNAs, and increase the chance that the spike-in oligos behaves in a similar manner on the NGS sequencing platform. Applicants note that although the disclosed spike-in set is being designed with NGS considerations in mind, they may be used in other small RNA analysis methods.

Methods of Use

[0063] Further aspects of the disclosure relate to the use of these sequences or pools in the analysis of small extracellular RNAs (exRNA) – such as, but not limited to, the RNAs in

exosomes and other extracellular vesicles – and/or quantify the amount of exRNA in a sample or to normalize exRNA data between different samples.

[0064] Some aspects relate to the addition of the artificial short RNA sequences to biofluids and/or RNA samples at different steps of experimental procedures or the normalization of endogenous RNA among different samples. Thus, in some embodiments, the artificial short RNA is detectably labeled, optionally 5' phosphorylated. Levels of these artificial short RNA sequences can be determined using qRT-PCR, microarray or next-generation sequencing (NGS). In some embodiments, equimolar or ratiometric molar concentrations are used.

[0065] The collection of RNA oligo spike-ins disclosed herein contains a sufficient number of oligos (~100) such that distinct spike-in sets, with a useful range of sizes and/or concentrations, can be added at *both* the RNA isolation *and* at the NGS library generation phases. Furthermore, the proposed spike-in oligos have a wide range of lengths (16 to 70 nt), allowing for both the detection of size bias in the RNA isolation steps, and providing the ability to gauge the effectiveness of the size selection processes during library preparation. Additionally, this larger size range increases the flexibility of the spike-ins to include all relevant small RNA species (i.e., miRNA, piRNA, snoRNA, snRNA, Y RNA fragments, etc.) whereas the currently available spike-ins (Exiqon) are restricted to only the miRNA class of small RNAs.

[0066] Potential uses for the spike-ins disclosed herein include but are not limited to (1) evaluating RNA yield through RNA isolation, (2) evaluating the amount of input RNA into library prep using sequence data (*e.g.* variation % calibrator in sequencing reads), (3) evaluating library quality (*e.g.*, overt failures, limitation of detection/complexity), (4) where multiple sets of calibrators are used, detecting cross-contamination or sample misidentification in batches of libraries, (5) normalizing libraries into a sequencing run, (6) normalizing data (*e.g.* taking in to account RNA length), (7) evaluating performance of size selection, and (8) evaluating library preparation steps (*e.g.* adapter ligation). Some embodiments, *e.g.* library quality evaluation for limitation of detection/complexity, a range of concentrations of spike-in RNAs may be used.

[0067] It is appreciated that certain sequences in a “spike-in set” may be considered inappropriate for use in a particular method due to an effect known as “jackpotting.” For example, for the purposes on Next Generation Sequencing (NGS), a sequence may be considered to be “jackpotting” (i.e. representing a higher or lower fraction of the small RNA sequencing reads than would be expected based on the equimolar pooling of the component sequences) when, in an equimolar spike in pool consisting only of spike in sequence, that sequence represented over 4% of the resultant reads in any of the tested NGS library preparation techniques. In such an experiment, a sequence can be considered to have failed when it receives <0.02% of the available reads. For example, the spike in set in **Table 3** was tested with a representative and broad range of NGS library preparation techniques and chemistries, and is, thus, expected to perform in a similar fashion to other kits not specifically tested here, as well as future NGS library preparation chemistries. This method can, in turn, be optimized for other uses with other analytical methods (e.g., qRT-PCR, microarray, etc.) by performing the filtering step using the planned analytical method and adapting the criteria for removal accordingly.

[0068] Using the known concentration of spike-ins added during RNA isolation, the total molar amount and concentration of miRNAs (or other types of RNA) in a sample can be estimated. For example, a biofluid may be divided into aliquots of different volumes and spiked with a range of concentrations of a set of spike-in oligonucleotides; For example, in plasma, the biofluid may be aliquoted into volumes ranging from about 100µL to about 450µL and the corresponding spike-in set may be introduced in a range of concentrations from about 1×10^{-17} to about 1×10^{-18} moles per 100µL. An ordinary skilled artisan would appreciate that the range is adjusted the match the expected specific biofluid content. Extracellular RNA can then be isolated or analyzed using methods including but not limited to qRT-PCR, microarray, NanoString, or targeted small RNA sequencing. The read count of spike-in RNAs is then compared to the total miRNA read count in each library. Using the known concentration of spike-ins added during RNA isolation, along with the number of spike-in oligonucleotides in the pool and the volume of plasma used for RNA isolation, the absolute concentration of miRNA in the plasma can be calculated.

Examples

[0069] The following examples are non-limiting and illustrative of procedures, which can be used in various instances in carrying the disclosure into effect. Additionally, all reference disclosed herein are incorporated by reference in their entirety.

Example 1 – Design and synthesis of a set of small RNA spike-in synthetic oligos for use in NGS analysis and other small RNA analytical methods.

[0070] Multiple oligo sizes across a relevant range (16, 18, 20, 22, 24, 28, 32, 36, 40, 50, & 70nt) were selected to examine size selection biases at various steps in the different processes. To mimic biological relevance of the oligos, to the extent possible, the dinucleotide frequencies of the synthetic oligos were matched those found in miRBase, a high quality human miRNA database. Specifically, these small RNAs were extracted from the database and grouped by family. A random selection of a proportional number of sequences from each family was used to form a single synthetic concatemer sequence that represented the dinucleotide frequency of the population (**Table 1**).

Table 1: miRNAs Utilized as Template for Random Short Oligo Generation

Name	Sequence	Name	Sequence
hsa-let-7a-3p	CUAUACAAUCUACUGUCUUUC	hsa-miR-340-3p	UCCGUCUCAGUUACUUUAUAGC
hsa-let-7d-3p	CUAUACGACCUGCGCCUUUCU	hsa-miR-342-3p	UCUCACACAGAAAUCGCACCCGU
hsa-let-7f-5p	UGAGGUAGUAGAUUGUAUAGUU	hsa-miR-345-3p	GCCCUGAACGAGGGGUCUGGAG
hsa-miR-103a-2-5p	AGCUUCUUUACAGUGCUGCCUUG	hsa-miR-34b-5p	UAGGCAGUGUCAUUAGCUGAUUG
hsa-miR-105-5p	UCAA AUGCUCAGACUCCUGUGGU	hsa-miR-3605-3p	CCUCCGUGUUACCUGUCCUCUAG
hsa-miR-106b-5p	UAAAGUGCUGACAGUGCAGAU	hsa-miR-3613-5p	UGUUGUACUUUUUUUUUUUGUUC
hsa-miR-10b-5p	UACCCUGUAGAACC GAAUUUGUG	hsa-miR-3614-3p	UAGCCUUCAGAUCCUUGGUGUUUU
hsa-miR-1249-3p	ACGCCC UCCCCCCC UUCUUCA	hsa-miR-361-5p	UUAUCAGAAUCUCCAGGGGUAC
hsa-miR-126-3p	UCGUACCGUGAGUAAUAAUGCG	hsa-miR-362-5p	AAUCCUUGGAACCUAGGUGUGAGU
hsa-miR-1271-3p	AGUGCCUGCUAUGUGCCAGGCA	hsa-miR-363-5p	CGGGUGGAUCACGAUGCAAUUU
hsa-miR-1278	UAGUACUGUGCAUAUCAUCUAU	hsa-miR-365a-3p	UAAUGCCCCUAAAAAUCCUUAU
hsa-miR-128-2-5p	GGGGGCCGAUACACUGUACGAGA	hsa-miR-370-3p	GCCUGCUGGGGUGGAACCUUGGU
hsa-miR-1285-3p	UCUGGGCAACAAAGUGAGACCU	hsa-miR-374a-3p	CUUAUCAGAUUGUAUUGUAAUU
hsa-miR-1285-5p	GAUCUCACUUUGUUGCCAGG	hsa-miR-376a-2-5p	GGUAGAUUUUCCUUCUAUGGU
hsa-miR-1287-5p	UGCUGGAUCAGUGGUUCGAGUC	hsa-miR-378a-5p	CUCCUGACUCCAGGUCCUGUGU
hsa-miR-129-2-3p	AAGCCCUUACCCCAAAAAGCAU	hsa-miR-382-3p	AAUCAUUCACGGACAACACUU
hsa-miR-1296-3p	GAGUGGGGCUUCGACCCUAACC	hsa-miR-382-5p	GAAGUUGUUCGUGGUGGAUUCG
hsa-miR-1304-3p	UCUCACUGUAGCCUCGAACCCC	hsa-miR-423-3p	AGCUCGGUCUGAGGCCCCUCAGU
hsa-miR-1306-5p	CCACCUC CCCUGCAAACGUCCA	hsa-miR-424-5p	CAGCAGCAAUUAUGUUUUUGAA
hsa-miR-1307-5p	UCGACCGGACCUCGACCGGCU	hsa-miR-425-3p	AUCGGGAUGUCGUGUCCGCC

hsa-miR-130b-5p	ACUCUUUCCUGUUGCACUAC	hsa-miR-431-5p	UGUCUUGCAGGCCGUCAUGCA
hsa-miR-132-3p	UACAGUCUACAGCCAUGGUCG	hsa-miR-433-5p	UACGGUGAGCCUGUCAUUUUC
hsa-miR-133a-5p	AGCUGGUAAAAUGGAACCAAU	hsa-miR-450b-3p	UUGGGAUCAUUUUGCAUCCAUA
hsa-miR-134-5p	UGUGACUGGUUGACCAGAGGGG	hsa-miR-4524a-5p	AUAGCAGCAUGAACCUGUCUCA
hsa-miR-135b-5p	UAUGGCUUUUCAUUCUAUGUGA	hsa-miR-454-5p	ACCCUAUCAAAUUAUGUCUCUGC
hsa-miR-136-3p	CAUCAUCGUCUCAAUAGAGUCU	hsa-miR-4707-5p	GCCCCGGCGCGGGCGGGUUCUGG
hsa-miR-138-1-3p	GCUACUUCACAACACCAGGGCC	hsa-miR-4755-5p	UUUCCCUUCAGAGCCUGGCCUUU
hsa-miR-138-5p	AGCUGGUGUUGUGAAUCAGGCCG	hsa-miR-4787-5p	GCGGGGUGGCGGCGGCAUCCC
hsa-miR-139-5p	UCUACAGUGCACGUGUCUCCAGU	hsa-miR-483-3p	UCACUCCUCUCCUCCCGUCUU
hsa-miR-1-3p	UGGAAUGUAAAGAAGUAUGUAU	hsa-miR-485-3p	GUCAUACACGGCUCUCCUCUCU
hsa-miR-140-5p	CAGUGGUUUUACCCUAUGGUAG	hsa-miR-488-5p	CCCAGAUAAUGGCACUCUCA
hsa-miR-141-5p	CAUCUUCAGUACAGUGUUGGA	hsa-miR-490-3p	CAACCUGGAGGACUCCAUGCUG
hsa-miR-143-5p	GGUGCAGUGCUGCAUCUCUGGU	hsa-miR-491-5p	AGUGGGGAACCCUCCAUGAGG
hsa-miR-144-3p	UACAGUAUAGAUGAUGUACU	hsa-miR-493-3p	UGAAGGUCUACUGUGGCCAGG
hsa-miR-146a-5p	UGAGAACUGAAUCCAUGGGUU	hsa-miR-495-3p	AAACAAACAUUGGUGCACUUCU
hsa-miR-148a-3p	UCAGUGCACUACAGAACUUUGU	hsa-miR-497-3p	CAAACCACACUGUGGUGUAGA
hsa-miR-149-5p	UCUGGCUCGGUGUCUUCACUCCC	hsa-miR-5001-3p	UUCUGCCUCUGUCCAGGUCCUU
hsa-miR-150-5p	UCUCCCAACCCUUGUACCAGUG	hsa-miR-500a-3p	AUGCACCUGGGCAAGGAUUCUG
hsa-miR-151a-5p	UCGAGGAGCUCACAGUCUAGU	hsa-miR-5010-3p	UUUUGUGUCUCCCAUUCCCCAG
hsa-miR-155-3p	CUCCUACAUUUAGCAUUAACA	hsa-miR-503-3p	GGGGUAAUUGUUCCGCUGCCAGG
hsa-miR-15a-5p	UAGCAGCACAUAAUGGUUUGUG	hsa-miR-504-3p	GGGAGUGCAGGGCAGGGUUUC
hsa-miR-1-5p	ACAUACUUCUUUAUAGCCCAU	hsa-miR-505-3p	CGUCAACACUUGCUGGUUCCU
hsa-miR-16-5p	UAGCAGCACGUAAAUUUGGCG	hsa-miR-509-3p	UGAUUGGUACGUCUGGGUAG
hsa-miR-181a-3p	ACCAUCGACCGUUGAUUGUACC	hsa-miR-514b-3p	AUUGACACCUCUGUGAGUGGA
hsa-miR-181d-3p	CCACCGGGGGAUGAAUGUCAC	hsa-miR-5196-5p	AGGGAAGGGGACGAGGGUUGGG
hsa-miR-182-3p	UGGUUCUAGACUUGCCAACUA	hsa-miR-532-3p	CCUCCACACCCAAGGCUUGCA
hsa-miR-183-3p	GUGAAUUACCGAAGGGCCAUA	hsa-miR-542-5p	UCGGGGAUCAUCAUGUCACGAGA
hsa-miR-185-5p	UGGAGAGAAAGGCAGUUCUGA	hsa-miR-545-5p	UCAGUAAAUGUUUAUAGAUGA
hsa-miR-186-3p	GCCCAAAGGUGAAUUUUUUGGG	hsa-miR-548d-5p	AAAAGUAAUUGUGGUUUUUGCC
hsa-miR-1908-5p	CGGCGGGGACGGCGAUUGGUC	hsa-miR-548o-3p	CCAAAACUGCAGUUAUUUUUGC
hsa-miR-190a-5p	UGAUUAUGUUUGAUUAUUAGGU	hsa-miR-551b-3p	GCGACCAUACUUGGUUUCAG
hsa-miR-1910-3p	GAGGCAGAAGCAGGAUGACA	hsa-miR-556-5p	GAUGAGCUCAUUGUAAUAUGAG
hsa-miR-191-3p	GCUGCGCUUGGAUUUCGUCCCC	hsa-miR-561-3p	CAAAGUUUAAGAUCUUGAAGU
hsa-miR-192-3p	CUGCCAAUCCAUAAGGUCACAG	hsa-miR-570-5p	AAAGGUAAUUGCAGUUUUUCCC
hsa-miR-193a-3p	AACUGGCCUACAAAGUCCAGU	hsa-miR-574-3p	CACGCUCAUGCACACACCCACA
hsa-miR-194-3p	CCAGUGGGGCGUCUGUUAUCUG	hsa-miR-576-3p	AAGAUGUGGAAAAAUUGGAAUC
hsa-miR-194-5p	UGUAACAGCAACUCCAUGUGGA	hsa-miR-577	UAGAUAAAUAUUGGUACCUG
hsa-miR-196a-3p	CGGCAACAAGAAACUGCCUGAG	hsa-miR-582-3p	UAACUGGUUGAACAAACUGAACC
hsa-miR-196b-3p	UCGACAGCACGACACUGCCUUC	hsa-miR-584-5p	UUAUGGUUUGCCUGGGACUGAG
hsa-miR-197-5p	CGGGUAGAGAGGGCAGUGGGAGG	hsa-miR-589-5p	UGAGAACCACGUCUGCUCUGAG
hsa-miR-199b-3p	ACAGUAGUCUGCACAUUGGUUA	hsa-miR-590-3p	UAAUUUAUUGUAUAAGCUAGU
hsa-miR-19b-2-5p	AGUUUUGCAGGUUUGCAUUUCA	hsa-miR-615-5p	GGGGGUCCCCGGUGCUCGGAUC
hsa-miR-202-3p	AGAGGUUAUAGGGCAUGGGAA	hsa-miR-616-5p	ACUCAAACCCUUCAGUGACUU
hsa-miR-203a-5p	AGUGGUUCUUAACAGUUCAACAGUU	hsa-miR-624-5p	UAGUACCAGUACCUUGUGUUCA
hsa-miR-208b-5p	AAGCUUUUUGCUCGAAUUAUGU	hsa-miR-625-5p	AGGGGGAAAGUUCUAUAGUCC

hsa-miR-210-5p	AGCCCCUGCCCACCGCACACUG	hsa-miR-627-5p	GUGAGUCUCUAAGAAAAGAGGA
hsa-miR-211-3p	GCAGGGACAGCAAAGGGGUGC	hsa-miR-628-5p	AUGCUGACAUUUUACUAGAGG
hsa-miR-218-1-3p	AUGGUUCCGUCAAGCACCAUGG	hsa-miR-629-3p	GUUCUCCCAACGUAAGCCCAGC
hsa-miR-219a-2-3p	AGAAUUGUGGCUGGACAUCUGU	hsa-miR-642a-3p	AGACACAUUUGGAGAGGGAACC
hsa-miR-222-5p	CUCAGUAGCCAGUGUAGAUCU	hsa-miR-6503-5p	AGGUCUGCAUUCAAAUCCCCAGA
hsa-miR-223-5p	CGUGUAUUUGACAAGCUGAGUU	hsa-miR-6511a-3p	CCUCACCAUCCCUUCUGCCUGC
hsa-miR-22-3p	AAGCUGCCAGUUGAAGAACUGU	hsa-miR-6511a-5p	CAGGCAGAAGUGGGGCUGACAGG
hsa-miR-23a-5p	GGGGUUCUGGGGAUGGGAUUU	hsa-miR-651-5p	UUUAGGAUAAGCUUGACUUUUG
hsa-miR-26a-2-3p	CCUAUUCUUGAUUACUUGUUUC	hsa-miR-652-3p	AAUGGCGCCACUAGGGUUGUG
hsa-miR-27b-3p	UUCACAGUGGCUAAGUUCUGC	hsa-miR-654-3p	UAUGUCUGCUGACCAUCACCUU
hsa-miR-296-5p	AGGGCCCCCCCCUCAUCCUGU	hsa-miR-664b-3p	UUCAUUUGCCUCCAGCCUACA
hsa-miR-299-5p	UGGUUUACCGUCCCAUAACAU	hsa-miR-671-3p	UCCGGUUCUCAGGGCUCCACC
hsa-miR-29b-1-5p	GCUGGUUUCAUAUGGUGGUUAGA	hsa-miR-675-3p	CUGUAUGCCCUCACCGCUCA
hsa-miR-302a-3p	UAAGUGCUUCCAUGUUUUGGUGA	hsa-miR-708-3p	CAACUAGACUGUGAGCUUCUAG
hsa-miR-3065-5p	UCAACAAAUCACUGAUGCUGGA	hsa-miR-7-1-3p	CAACAAAUCACAGUCUGCCAU
hsa-miR-3074-5p	GUUCCUGCUGAACUGAGCCAG	hsa-miR-744-5p	UGCGGGGCUAGGGCUAACAGCA
hsa-miR-30a-3p	CUUUCAGUCGGAUGUUUGCAGC	hsa-miR-758-3p	UUUGUGACCUGGUCCACUAACC
hsa-miR-30c-5p	UGUAAACAUCUACACUCUCAGC	hsa-miR-7-5p	UGGAAGACUAGUGAUUUUGUUGU
hsa-miR-3130-3p	GCUGCACCGGAGACUGGGUAA	hsa-miR-766-3p	ACUCCAGCCCCACAGCCUCAGC
hsa-miR-3140-5p	ACCUGAAUUACCAAAAGCUUU	hsa-miR-767-3p	UCUGCUCAUACCCCAUGGUUUUCU
hsa-miR-3158-3p	AAGGGCUUCCUCUCUGCAGGAC	hsa-miR-769-3p	CUGGGAUCUCCGGGGUCUUGGUU
hsa-miR-31-5p	AGGCAAGAUGCUGGCAUAGCU	hsa-miR-873-5p	GCAGGAACUUGUGAGUCUCCU
hsa-miR-32-5p	UAUUGCACAUAUACUAAAGUUGCA	hsa-miR-874-3p	CUGCCCUGGCCGAGGGACCGA
hsa-miR-328-3p	CUGGCCUCUCUGCCCUUCCGU	hsa-miR-876-5p	UGGAUUUCUUUGUGAAUCACCA
hsa-miR-330-5p	UCUCUGGGCCUGUGUCUUAGGC	hsa-miR-885-3p	AGGCAGCGGGGUGUAGUGGAUA
hsa-miR-331-3p	GCCCCUGGGCCUAUCCUAGAA	hsa-miR-887-5p	CUUGGGAGCCCUGUUAGACUC
hsa-miR-335-5p	UCAAGAGCAAUAACGAAAAAUGU	hsa-miR-92b-5p	AGGGACGGGACGCGGUGCAGUG
hsa-miR-337-3p	CUCCUAUAUGAUGCCUUUCUUC	hsa-miR-941	CACCCGGCUGUGUGCACAUGUGC
hsa-miR-338-3p	UCCAGCAUCAGUGAUUUUGUUG	hsa-miR-942-5p	UCUUCUCUGUUUUGGCCAUGUG
hsa-miR-339-3p	UGAGCGCCUCGACGACAGAGCCG	hsa-miR-96-5p	UUUGGCACUAGCACAUUUUGCU
hsa-miR-33b-3p	CAGUGCCUCGGCAGUGCAGCCC		

[0071] Nonsense sequences were placed in between each small RNA sequence in the concatemer as to remove potential dinucleotide bias that would occur with a direct end-to-end fusion of individual sequences. Between 50 and 118,000 oligos of each size were randomly generated (Table 2) from this sequence to be bioinformatically examined.

Table 2: RNA Oligo Spike-In Generation Process

Length	# Generated	# Passed Filters	# Chosen	
			Pool 1	Pool 2
16	117802	185	3	3
18	100	11	3	3

20	50	29	10	10
22	50	40	10	10
24	50	46	10	10
28	50	45	10	10
32	50	49	3	3
36	50	47	3	3
40	50	44	2	2
50	50	50	2	2
70	50	46	2	2

[0072] These oligos were processed through the Genboree exceRpt small RNA mapping pipeline against endogenous (human) and exogenous reference libraries at a stringency of 0 allowed mismatches, rather than the default of 1 allowed mismatch. The oligos that were found to have no hits using this pipeline were then used as a reference library against which human biofluid exRNA small RNA-seq datasets from the Erle and Galas groups were mapped to ensure that there was no mapping of reads from actual datasets to these sequences occurred. The biofluids represented in these datasets were: amniotic fluid; cerebrospinal fluid, adult serum, adult plasma, cord blood plasma, bronchoalveolar lavage fluid, saliva, sputum, and adult urine.

[0073] Applicants then selected the spike-in oligonucleotides for synthesis from the collection of oligos that passed these mapping filters. To increase the probability of sampling the sequencing diversity within this collection, the collection was divided according to oligo length and each length category was aligned using MUSCLE, and phylogenetic trees were generated using MEGA7. Within MEGA7, the optimal substitution matrix was determined to be Jukes Cantor, which was then used to create maximum likelihood trees with a bootstrap value of 100. Regions of the alignment that lacked complete coverage across the sequences were considered uninformative and these data were eliminated in branch determination. The resultant trees (example tree in **Figure 1**) were inspected visually and a random oligo was chosen from each apparent subgroup of the tree until the target number of oligos for each size range was achieved (**Table 3**).

Table 3: List of Oligos

Oligo Name	Sequence
16_102089	TTGTGCCCTAATCCGT

16_14298	ACTCACGGCCGTCTTA
16_43588	CTAATGCGTTCGGGCC
16_64719	CCGCCTATAGCGTACT
16_67605	GTCTAGCCGGGTACT
16_69200	CATATTATGGGTCCGG
18_21	TGGGTGAGGTCATAGCTA
18_42	TGGTCCGCAGTGAATGGG
18_47	TAGAACGGTCGCTCACTG
18_48	AGGCCACTCCGTCTTTAC
18_6	ACTGTTTGACGCCTCCGT
18_71	AGCTTCCCGGTCGAATAC
20_100	TGTCCAATTAATGCTCGGCA
20_51	GTCTAGGACGCTCATTACCA
20_52	GACATAATCCTACCTGGAGC
20_54	TCCTGGCAGTGCGTAACTAA
20_56	GTCTAAGGGTGAAGACGAT
20_57	GCAGGCTTATCTGTCTCGCC
20_58	GGGTTGGCCCGGGTCTCGGA
20_59	CTTAACTATGCACGCACTAC
20_61	TGCTTTGATTACAGCGTCACG
20_64	TGGTAGTTTCACTCCGCATG
20_66	GATGATGCGTACGCTAATTC
20_68	CAGCTGAGGGATTGATTTAC
20_76	CATCCAACGTGCTTCCGTGC
20_78	CTGGCAAACACGGTCCATAA
20_82	TGACCCATACGATCGTCTTG
20_90	GTCCGGTCGCTTAGCCGTCC
20_92	AGACCTCATTAGACTCAAA
20_93	TCAACCACCGTTGGCCTGCG
20_94	CTTGAGTGAGTAGTCTCCAA
20_96	ATTAGAAATTATCGTTCATC
22_52	GCAGACGTTTGTCTGCGGTCCA
22_55	TGCATTCATCGAGGTGGATTGT
22_56	GCCCGCGGCTGCGTAAAGAAAC
22_60	TTGTGCACTCTACCCTAGACAA
22_61	CATTGCTCGGAAAGGTGCCTCG
22_63	TCCACCTTTCTGAATCCATGCT
22_67	AGTTTCAGCGAGATGTTTGAAG
22_69	TTGCACGCTAACGGTCCCTGTT
22_71	CATTGGTAGCTCCGACTTTCGC
22_77	AACGGTCCTTACTTCCGACTGA
22_78	TTTGGACCGGTATCAGTCGGCC
22_81	GCGGTTAAAGCCCGATCACTCC
22_82	TGAAAGCTGGCATGGTCCATAA
22_87	AGTGGCACCATCCTTCGGGATA
22_88	CCTCGACATGATCCAGTCTGCA
22_91	ACGTTTATTGCCTTCAAGTGGC
22_93	CGATAGTCCCGGTGCGTGGCGA
22_94	TTTGCCGATAAACAACCTACGA
22_98	ATCCACTTGGTCTGGGTCAATT
22_99	CAGGAGGCCTGCCAAGCGTGCT

24_100 ATGATCGGGTGAGCATCTGTTTAG
 24_55 CCGTTCGTTTCGGCAGATTGGTTCA
 24_56 GCCTTCCCATGTTAGGAAATGGGC
 24_58 GCCCGCACGGCTCCTTGGGCGAGG
 24_59 GTTTAGCGTGATTGGGTGAACCT
 24_60 GGGTTCGGCGGGCTGCGTCTTCCT
 24_61 CTGCCCTAGGTAGGCTCATCTTAG
 24_66 AGGTGCATCCTATTGGGTAGTTGG
 24_67 CAAGCAAGCTTGTTACTTCCAGTG
 24_68 GGCCGCCCTTCCAGGTTGCTTGCC
 24_74 AACGGGATCGAGATCCGACGCACT
 24_77 AGCAAACCTCTAATGCGTCCTCCT
 24_78 TGAGCCATTTGGGTGCGGACCACC
 24_82 GGGCAAATTAGTAGTACTGACAGG
 24_86 TATGGACTCAGTTCTATCGATTCT
 24_88 TCCCGATCAGTTCCTCGATGCAAT
 24_89 CCTCTTCCTCGTTTGAAGACCCAT
 24_92 GCTCGCAATCTTTACAATTCTTCC
 24_96 ATCGGCGTTAGATCCCATTGCCTG
 24_98 GCTCTGCTTCGAATAGGGTCTTCT
 28_100 CTATTGTGCGCCTTCTTGGATAGCCTT
 28_51 CCTTTGGCGGGATTTACCGTAAGTTCCC
 28_54 ACCTAGGTACCCACTTTAGGGAACTTTA
 28_58 GGGTTCGGGCGGGTGCGGCTTGCAGC
 28_59 GCTCACGGGTGATTCCGTACCCAAGGAA
 28_62 CTGCTGATCTGGTGTCTGCAACCACCTC
 28_65 CCCTTGCCAATTGATGAAAGGAGCTGAA
 28_68 ACCCTGGCGGGCAGACAACACTTTGAAG
 28_70 AGTGTACTCCTGGAGTGGCTTTGCTCCT
 28_71 TTACTTTGCCCGTCGTTCAACCATGCT
 28_73 TAATCTTGAAAGTTGTTTACCCTCCCTT
 28_79 AGCACCTAACAAGGATTGAAATGCCTG
 28_83 AAGCGTTCAGGGCGGTGATCCTAGTGGG
 28_86 TTGATTAGAAGTTAGTATGCCAAGTGCT
 28_87 GCAGTACCCGACTAGACCTCTTATCCTA
 28_93 CGCAAACATCGGTGCTACAAATCTTTC
 28_94 CATAGCGGGAGGATCTTTGCCCATTCCT
 28_96 ACTATTGGAGTTATTGATCCTATGCTGG
 28_98 TGTCAAGACATAGTTCAATGATAAAGGG
 28_99 CGTCCTTCAGTGAATCAAGCTCTTATTG
 32_57 CACTATAATCTACCCAGCTTGCCTAATTAGAA
 32_58 TTAGCAACTTGAACCAGGCATGAGGTAATTG
 32_62 GCTTCCGCCATGTAAGCTTGGGTTGACTGTTT
 32_66 CCATTTCTTGTTATCATTTGCTCAGCCTAAGTG
 32_67 TTTGAACTTGATCTAGCCACAGGCTGCTACCC
 32_81 TGGTGATAAGTGAAGTATTGCCATGGAATCCC
 36_100 ATTTGATATCCCGGGATTCTTACCCAGGTCGCTT
 36_52 TGGGCTGAAAGGGTACGTGCATCATGTCTTGTCATT
 36_66 GACATGGTCTATTTGTTATGTCTTCAACGCTGACGC
 36_69 GAGTGACAAGACTTTACCTGCCATAGCGTCTGGGTC
 36_80 AAAGCCGGCTCAGGAAGCTTTGGTCTACCACCTGAA

```

36_93      TGAGTGCCCTCAGCATGCTATGACTGTTTGATTTAG
40_51      GTCTCCCATGCAGTCCGCTGCAGGTTTCATGGGCCGATCCT
40_57      TTACAGTTGGCTGTCAGCCGACGGTATTAGGTCCTTGTA
40_77      TCCATGGGTGCGTCTGACCTAGACGATCACTGATCTGTAT
40_80      TCACCTCAGGTATCCCAGTTTGCCTTGGCTCGTAAATAAT
50_53      AACTAGGTTACCCTTGCTCGACATGACCCACGTACGCCTCAACTTCCACC
50_78      AGCCGACGGTGGAAGTTATTACGTCTTCGTAGTTTAATCAAGCTTGCTAA
50_92      TCTTTGAGCAGGTTACGCGTTGTAATGTTACTGGGAGTTTGACTCCAAT
50_96      CTACCAGTTACTCAGTCCATGCTGCCGAGTCAGGTAAGCGGGTTTCGGGA
70_57      ACCTCCAACTGGAAGGGTCTTCCTTGCGAATATGGTGATTAAACCGGCAGTAGGTCTAACGAA
70_70      TAGTCTGTTGGTTTCAATACGGCATCGCTCATGGTCTTGCCATCTCATCAAGGACTCCCTCAGGGTGCC
70_83      GTCAAGCGATAAGCATGCGGTATTCCTAGTTCAGACAAGTCCTACACTATGGTCGAATGGCTCCGCAACC
70_96      AGTTTAGTTGACTCGTGACGCTTCCCTTGCCTTAGTGGCGTGCCATGCCTGCACCGTCCGCGGTTAGGA

```

[0074] To balance the two priorities of inclusion of sufficient spike-in reads for QC and maximization of the number of reads from experimental samples, the spike-ins will be added with the aim to account for 1-5% of the total number of reads in a small RNA-seq library. Applicants expect that the different synthetic oligos provide sequence data at different efficiencies. Applicants start with a larger number of oligos than expected to be included in the final spike-in sets to allow for drop-out during the validation process. The selected sequences are synthesized as 5' phosphorylated RNA oligonucleotides to mimic the chemical characteristics of endogenous miRNAs. The synthesis is performed at the smallest commercially available scale (100nM); we estimate that this synthesis will provide enough spike-in oligo for over 1 billion RNA isolations from biofluid samples or small RNA sequencing libraries. **Table 4** also shows the number of oligos reserved for each of two distinct spike-in sets, one for the RNA isolation process and one for the RNA analysis process.

Table 4: Numbers of Oligos for Each Process

Oligo Length	RNA Isolation	Small RNA Library Production
16	3	3
18	3	3
20	10	10
22	10	10
24	10	10
28	10	10
32	3	3
36	3	3

40	2	2
50	2	2
70	2	2
Total	58	58

[0075] Applicants have designed and synthesized a spike-in oligonucleotide set that mimics the chemistry and base composition of biologically important exRNAs, while having sequences that are distinct from endogenous RNAs and genomes of humans and the other species represented in the Genboree exceRpt pipeline.

Example 2 – Validation of Small-RNA Oligonucleotide Spike-In Sets

A. Removal of over-represented oligos and construction of final spike-in sets.

Construction and Evaluation of Test pool

[0076] A small-scale equimolar test pool with all of the synthesized spike-in oligos is prepared and sequenced to identify oligos that are unsuitable for inclusion within subsequent pools due to overrepresentation in the resulting read counts, or “jackpotting,” underrepresented sequences may also be excluded. Not to be bound by theory, Applicants estimate that ~20% of the oligos do not pass this filter and are eliminated from the final spike-in sets.

[0077] The pooling and dilution strategy previously used by the Galas lab for development of a small set of spike-in oligos forms the basis of the strategy used here. In that strategy, separate pools were made for pre-RNA isolation spike-in and for direct spike-in during library generation. This allows for QC of both the RNA isolation and library preparation steps independently. For the direct library spike-in, a final count of $\sim 1 \times 10^{-19}$ moles of pooled oligo per library was added, and for the pre-RNA isolation spike-in, $\sim 3 \times 10^{-18}$ moles pooled oligo was diluted in the Qiazol lysis reagent used in the RNA extraction procedure. In libraries exRNA isolated from plasma or urine, the final spike-in reads accounted for <1% of the total. Here, with a higher target depth of 5% total sequence reads from spike-in oligos, the target number of total moles/pool increases 5x, so that the direct library spike-in pool increases to $\sim 5 \times 10^{-19}$ moles and the RNA isolation spike-in increases to $\sim 1.5 \times 10^{-17}$ moles. It

should be noted that because the original Galas pooling strategy utilized 14 oligos at <1% sequencing depth and this current strategy with 116 oligos with ~5% sequencing depth, each oligo within the pool, using the proposed strategy is anticipated to have ~60% of the sequencing depth as seen in the Galas strategy.

[0078] For the test pool, lyophilized RNA oligo will be resuspended to 100 μ M in RNA Storage Solution (RNAse free 1mM sodium citrate buffer pH 6.4, ThermoFisher, Carlsbad, CA). Only “low nucleic acid binding” tubes and RNA Storage Solution is used in the oligo dilution process. 10 μ L of each 100 μ M oligo will be diluted to 10 μ M in a 100 μ L final volume. This 1:10 dilution is repeated for a 100 μ L 1 μ M RNA oligo stock. The 10 μ M and 1 μ M stocks is used for subsequent oligo pooling operations. 1 μ L of each of these 1 μ M stocks is pooled by equal volume to yield a pool concentration of 1 μ M.

[0079] Small RNA-seq libraries are generated from the test pool using the Illumina TruSeq, NEB NEBNext, Galas 4N, Erle 4N, and Clontech SMARTer small RNA-seq library preparation methods and sequenced. The resulting data are analyzed and highly overrepresented sequences are flagged and not included in construction of the final pools.

Construction of Final Pools

[0080] A maximum of $\frac{1}{2}$ of the available synthesis of any given oligo are used to generate the pools described below, reserving the remaining synthesis for generation of future pools.

Equimolar Pools

[0081] Two sets of equimolar oligo pools, with non-overlapping oligos, are prepared (Source Oligos listed on **Table 3**). It is our intention that **Equimolar SetA** be spiked into biofluids during the initial phase of the RNA extraction process and **Equimolar SetB** be spiked into the purified RNA at the beginning of the library preparation process.

Ratiometric Pools

[0082] It has been demonstrated that ratiometric standards can be used for accurate absolute quantification of target RNA molecules as observed with the Sequins, and to a lesser extent with the Zebrafish miRNA standards and the Exiqon standards.

Table 5 delineates the ratiometric pooling strategy.

Table 5: Oligo Pooling Strategy Option: Number of RNA molecules for each oligo size in pool

Oligo Lengt h	RNA Isolation			Oligo Lengt h	Library Generation		
	Equimolar SetA	Ratiometric SetA Mix1	Ratiometric SetA Mix2		Equimolar SetB	Ratiometric SetB Mix1	Ratiometric SetB Mix2
20	1.75E+08	1.00E+05	1.00E+11	20	5.18E+06	1.00E+03	1.00E+09
22	1.75E+08	1.06E+06	9.43E+10	22	5.18E+06	1.06E+04	9.43E+08
24	1.75E+08	4.20E+06	8.88E+10	24	5.18E+06	4.20E+04	8.88E+08
28	1.75E+08	1.12E+07	8.35E+10	28	5.18E+06	1.12E+05	8.35E+08
20	1.75E+08	2.39E+07	7.84E+10	20	5.18E+06	2.39E+05	7.84E+08
22	1.75E+08	4.44E+07	7.36E+10	22	5.18E+06	4.44E+05	7.36E+08
32	1.75E+08	7.51E+07	6.90E+10	32	5.18E+06	7.51E+05	6.90E+08
24	1.75E+08	1.18E+08	6.46E+10	24	5.18E+06	1.18E+06	6.46E+08
18	1.75E+08	1.77E+08	6.04E+10	18	5.18E+06	1.77E+06	6.04E+08
28	1.75E+08	2.53E+08	5.63E+10	28	5.18E+06	2.53E+06	5.63E+08
16	1.75E+08	3.49E+08	5.25E+10	16	5.18E+06	3.49E+06	5.25E+08
20	1.75E+08	4.70E+08	4.89E+10	20	5.18E+06	4.70E+06	4.89E+08
36	1.75E+08	6.17E+08	4.54E+10	36	5.18E+06	6.17E+06	4.54E+08
22	1.75E+08	7.94E+08	4.22E+10	22	5.18E+06	7.94E+06	4.22E+08
40	1.75E+08	1.00E+09	3.91E+10	40	5.18E+06	1.00E+07	3.91E+08
24	1.75E+08	1.25E+09	3.61E+10	24	5.18E+06	1.25E+07	3.61E+08
70	1.75E+08	1.54E+09	3.34E+10	70	5.18E+06	1.54E+07	3.34E+08
28	1.75E+08	1.87E+09	3.07E+10	28	5.18E+06	1.87E+07	3.07E+08
32	1.75E+08	2.24E+09	2.82E+10	32	5.18E+06	2.24E+07	2.82E+08
20	1.75E+08	2.67E+09	2.59E+10	20	5.18E+06	2.67E+07	2.59E+08
50	1.75E+08	3.15E+09	2.37E+10	50	5.18E+06	3.15E+07	2.37E+08
22	1.75E+08	3.69E+09	2.17E+10	22	5.18E+06	3.69E+07	2.17E+08
16	1.75E+08	4.30E+09	1.97E+10	16	5.18E+06	4.30E+07	1.97E+08
24	1.75E+08	4.97E+09	1.79E+10	24	5.18E+06	4.97E+07	1.79E+08
70	1.75E+08	5.71E+09	1.63E+10	70	5.18E+06	5.71E+07	1.63E+08
28	1.75E+08	6.52E+09	1.47E+10	28	5.18E+06	6.52E+07	1.47E+08
18	1.75E+08	7.42E+09	1.32E+10	18	5.18E+06	7.42E+07	1.32E+08
20	1.75E+08	8.39E+09	1.19E+10	20	5.18E+06	8.39E+07	1.19E+08
32	1.75E+08	9.46E+09	1.06E+10	32	5.18E+06	9.46E+07	1.06E+08
22	1.75E+08	1.06E+10	9.46E+09	22	5.18E+06	1.06E+08	9.46E+07
36	1.75E+08	1.19E+10	8.39E+09	36	5.18E+06	1.19E+08	8.39E+07
24	1.75E+08	1.32E+10	7.42E+09	24	5.18E+06	1.32E+08	7.42E+07
40	1.75E+08	1.47E+10	6.52E+09	40	5.18E+06	1.47E+08	6.52E+07
28	1.75E+08	1.63E+10	5.71E+09	28	5.18E+06	1.63E+08	5.71E+07
50	1.75E+08	1.79E+10	4.97E+09	50	5.18E+06	1.79E+08	4.97E+07
20	1.75E+08	1.97E+10	4.30E+09	20	5.18E+06	1.97E+08	4.30E+07
16	1.75E+08	2.17E+10	3.69E+09	16	5.18E+06	2.17E+08	3.69E+07
22	1.75E+08	2.37E+10	3.15E+09	22	5.18E+06	2.37E+08	3.15E+07
18	1.75E+08	2.59E+10	2.67E+09	18	5.18E+06	2.59E+08	2.67E+07
24	1.75E+08	2.82E+10	2.24E+09	24	5.18E+06	2.82E+08	2.24E+07
36	1.75E+08	3.07E+10	1.87E+09	36	5.18E+06	3.07E+08	1.87E+07
28	1.75E+08	3.34E+10	1.54E+09	28	5.18E+06	3.34E+08	1.54E+07
20	1.75E+08	3.61E+10	1.25E+09	20	5.18E+06	3.61E+08	1.25E+07
22	1.75E+08	3.91E+10	1.00E+09	22	5.18E+06	3.91E+08	1.00E+07

24	1.75E+08	4.22E+10	7.94E+08	24	5.18E+06	4.22E+08	7.94E+06
28	1.75E+08	4.54E+10	6.17E+08	28	5.18E+06	4.54E+08	6.17E+06
20	1.75E+08	4.89E+10	4.70E+08	20	5.18E+06	4.89E+08	4.70E+06
22	1.75E+08	5.25E+10	3.49E+08	22	5.18E+06	5.25E+08	3.49E+06
24	1.75E+08	5.63E+10	2.53E+08	24	5.18E+06	5.63E+08	2.53E+06
28	1.75E+08	6.04E+10	1.77E+08	28	5.18E+06	6.04E+08	1.77E+06
20	1.75E+08	6.46E+10	1.18E+08	20	5.18E+06	6.46E+08	1.18E+06
22	1.75E+08	6.90E+10	7.51E+07	22	5.18E+06	6.90E+08	7.51E+05
24	1.75E+08	7.36E+10	4.44E+07	24	5.18E+06	7.36E+08	4.44E+05
28	1.75E+08	7.84E+10	2.39E+07	28	5.18E+06	7.84E+08	2.39E+05
20	1.75E+08	8.35E+10	1.12E+07	20	5.18E+06	8.35E+08	1.12E+05
22	1.75E+08	8.88E+10	4.20E+06	22	5.18E+06	8.88E+08	4.20E+04
24	1.75E+08	9.43E+10	1.06E+06	24	5.18E+06	9.43E+08	1.06E+04
28	1.75E+08	1.00E+11	1.00E+05	28	5.18E+06	1.00E+09	1.00E+03

[0083] There are four ratiometric pools. For the oligos present in Equimolar SetA, two mixes are made in which the oligos are distributed in opposing molar dilution ladders across different oligos of different lengths; these are designated **Ratiometric SetA Mix1** and **Ratiometric SetA Mix2**. The same is done for the oligos present in Equimolar SetB and be designated **Ratiometric SetB Mix1** and **Ratiometric SetB Mix2**. Oligo size classes with a larger number of oligos/class (20 nt-28 nt) are included across a larger range of dilutions, while those with a smaller number of oligos/class (16 nt, 18 nt, 32 nt-70 nt) are included near the mean oligo concentration of the pool. Initial calculations predict that while there should be $\sim 4 \times 10^8$ - 4×10^{10} doses for the RNA extraction pools and 2×10^{10} - 2×10^{12} , the limiting factor for pool usage is likely sample loss on the walls of storage tubes, and freeze-thaw cycles.

B. Test NGS spike-in sets.

Evaluation of the Test Pool

[0084] As mentioned above, Applicants first pool all of the spike-in oligos at equal ratios, and prepare small RNA-seq libraries using two fixed adaptor sequence small RNA library preparation methods (the NEB NEBNext or Illumina Truseq small RNA methods) and two degenerate adaptor small RNA library preparation methods (the methods previously developed by the Erle and Galas labs) and one template-switching method (Clontech SMARTer small RNA). Libraries will be generated from eight serial 1:10 dilutions of this pool. This multiplexed library pool is quantified, balanced, and sequenced on a single MiSeq

PE150 (Illumina) run. Oligos that jackpot or which are not detected in libraries prepared using any method are excluded from subsequent pools.

[0085] An oligo was considered to be “jackpotting” when, in an equimolar spike in pool consisting only of spike in oligos, that oligo represented over 4% of the resultant reads in any of the tested NGS library preparation techniques. In such an experiment, an oligo was considered to have failed when it received <0.02% of the available reads. The library preparation kits tested were the NEB Next Small RNA kit (NEB), the TrueSeq Small RNA kit (Illumina), The Clontech SMARTer kit (Takara), and a “homebrew” 4N based method. The NEB and Illumina library preparation kits utilize similar chemistries for preparation of the small RNA material for deep sequencing, with the primary differences being in how each kit removes unwanted chemical side products. The “4N” method uses similar adaptor ligation methods as the NEB and Illumina kits, however, it includes a randomized adaptor that significantly reduces bias in annealing the adaptors to both small RNAs in solution as well as to the spike-in oligos. The Takara library preparation kit utilizes a ligation-free approach to NGS library preparation which is orthogonal to other available methods. As such, the spike in oligos have been tested with a representative and broad range of NGS library preparation techniques and chemistries, and would be expected to perform in a similar fashion to other kits not specifically tested here, as well as future NGS library preparation chemistries.

Evaluation of Final Equimolar and Ratiometric Pools

[0086] Each pool (**Equimolar SetA, Equimolar SetB, Ratiometric SetA Mix1, Ratiometric SetA Mix2, Ratiometric SetB Mix1** and **Ratiometric SetB Mix2**) is initially evaluated by generating small RNA-seq libraries using the NEBNext (by the Laurent group) and the 4N (by the Galas group) methods on the pure pools at 4-8 serial 1:10 dilutions and sequenced on a MiSeq. These experiments determine whether the read counts for the component oligos correspond well to the expected numbers based on the pooling ratios.

[0087] The three SetA pools (**Equimolar SetA, Ratiometric SetA Mix1, Ratiometric SetA Mix2**) will then be spiked individually into three biofluids (serum, plasma, and urine) for RNA Isolation using the miRNeasy micro kit at concentrations approximating 1%, 5%,

and 10% of the miRNA concentration in the biofluid sample. To the resulting RNA samples, the three SetB pools (**Equimolar SetB**, **Ratiometric SetB Mix1** and **Ratiometric SetB Mix2**) are spiked-in at concentrations approximating 1%, 5%, and 10% of the miRNA concentration in the RNA samples, in a corresponding fashion (e.g. the RNA samples from the biofluid samples spiked with **Equimolar SetA** are spiked with **Equimolar SetB**), and small RNA-seq libraries are generated using the NEBNext (by the Laurent group) and the 4N (by the Galas group) methods and sequenced on a HiSeq (Laurent) or NextSeq (Galas). These experiments determine whether spiking the pools into biofluids (for the SetA pools) or complex RNA mixtures (for the SetB pools) influences their performance.

[0088] Applicants determine that it would be informative, the same experiment is also performed with swapping the SetA and SetB pools (i.e., **Equimolar SetB**, **Ratiometric SetB Mix1** and **Ratiometric SetB Mix2** are spiked into the biofluids and **Equimolar SetA**, **Ratiometric SetA Mix1**, and **Ratiometric SetA Mix2** are spiked into the purified RNA).

[0089] Applicants test the ratiometric pools to determine whether they can be used to normalize two serum samples that are expected to have different levels of certain miRNAs. **Ratiometric SetA Mix1** is spiked into a non-pregnant female serum sample and **Ratiometric SetA Mix2** is spiked into a pregnant female serum sample. RNA will be isolated using the miRNeasy micro kit. **Ratiometric SetB Mix1** is spiked into the non-pregnant female RNA sample and **Ratiometric SetB Mix2** is spiked into the pregnant female RNA sample. Small RNAseq libraries are generated using the NEBNext (by the Laurent group) and the 4N (by the Galas group) methods and sequenced on a HiSeq (Laurent) or NextSeq (Galas). These experiments demonstrate that we can use the ratiometric pools to normalize data from different samples.

[0090] Applicants develop a rigorously validated series of spike-in small RNA sets, made from the spike-in RNA oligos designed and synthesized under a separate proposal (“**Design and Synthesis of Small RNA Oligonucleotide Spike-ins**”). Not to be bound by theory, Applicants believe that the results are critical, in that they can generate tools that can be easily adopted by both highly experienced and less experienced laboratories whose experiments include exRNA isolation and/or analysis.

Example 3 – Normalization using the validated spike-in sets

[0091] Using the known concentration of spike-ins added during RNA isolation, the total molar amount and concentration of miRNAs (or other types of RNA) in a sample was estimated in this experiment. **Figure 5** shows the results of an experiment in which a large volume human plasma sample was divided into aliquots of different volumes (from 100uL-450uL) and spiked with a range of concentrations of a set of spike-in oligonucleotides (1×10^{-17} – 1×10^{-18} moles per 100uL of plasma). Extracellular RNA was then isolated from each spike sample and subjected to small RNA sequencing. The read count of spike-in RNAs in the range of 20-24nt long (the target range for size selection in these libraries) was compared to the total miRNA read count in each library. Using the known concentration of spike-ins added during RNA isolation, along with the number of spike-in oligonucleotides in the pool and the volume of plasma used for RNA isolation, the absolute concentration of miRNA in the plasma is calculated. Since the source of plasma was the same for all of the libraries, the finding that the estimated input miRNA concentration is relatively consistent across samples shows that the use of the spike-ins to estimate the miRNA concentration is robust to variations in sample input volume and spike-in concentration.

Equivalents

[0092] Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this technology belongs.

[0093] The present technology illustratively described herein may suitably be practiced in the absence of any element or elements, limitation or limitations, not specifically disclosed herein. Thus, for example, the terms “comprising,” “including,” “containing,” *etc.* shall be read expansively and without limitation. Additionally, the terms and expressions employed herein have been used as terms of description and not of limitation, and there is no intention in the use of such terms and expressions of excluding any equivalents of the features shown and described or portions thereof, but it is recognized that various modifications are possible within the scope of the present technology claimed.

[0094] Thus, it should be understood that the materials, methods, and examples provided here are representative of preferred aspects, are exemplary, and are not intended as limitations on the scope of the present technology.

[0095] The present technology has been described broadly and generically herein. Each of the narrower species and sub-generic groupings falling within the generic disclosure also form part of the present technology. This includes the generic description of the present technology with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is specifically recited herein.

[0096] In addition, where features or aspects of the present technology are described in terms of Markush groups, those skilled in the art will recognize that the present technology is also thereby described in terms of any individual member or subgroup of members of the Markush group.

[0097] All publications, patent applications, patents, and other references mentioned herein are expressly incorporated by reference in their entirety, to the same extent as if each were incorporated by reference individually. In case of conflict, the present specification, including definitions, will control.

[0098] Other aspects are set forth within the following claims.

References

- 1 Locati, M. D. et al. Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization. *Nucleic acids research* 43, e89, doi:10.1093/nar/gkv303 (2015).
- 2 Hardwick, S. A. et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nature methods* 13, 792-798, doi:10.1038/nmeth.3958 (2016).
- 3 Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. *Genome research* 21, 1543-1551, doi:10.1101/gr.121095.111 (2011).
- 4 Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* 42, D68-73, doi:10.1093/nar/gkt1181 (2014).

WHAT IS CLAIMED IS:

1. A set of spike-in RNAs comprising one or more artificial short RNA sequences that (1) have a base composition that mimics that of endogenous human miRNAs; (2) have broad sequence diversity; 3) cover a range of sequence lengths of between about 16 to 70 nucleotides; and/or (4) do NOT share sequence identity with known endogenous sequences in human, animal, plant, fungus, bacterial, and/or viral genomes.
2. The set of spike-in RNAs, wherein the one or more, optionally about 10 to 100, artificial short RNA sequences are selected from the sequences in **Table 3**, optionally having the same length.
3. A method of analyzing extracellular RNAs in a sample comprising introducing the set of spike-in RNAs from any one of claims 1 or 2 to the sample comprising small extracellular RNAs and small RNAs.
4. A method of normalizing data relating to extracellular RNAs and small RNAs between one or more samples comprising introducing the set of spike-in RNAs from any one of claims 1 or 2 to the one or more samples comprising small extracellular RNAs and/or small RNAs.
5. A method of detecting one or more of the set of spike-in RNAs from any one of claims 1 or 2 using qRT-PCR, microarray, or next-generation sequencing.

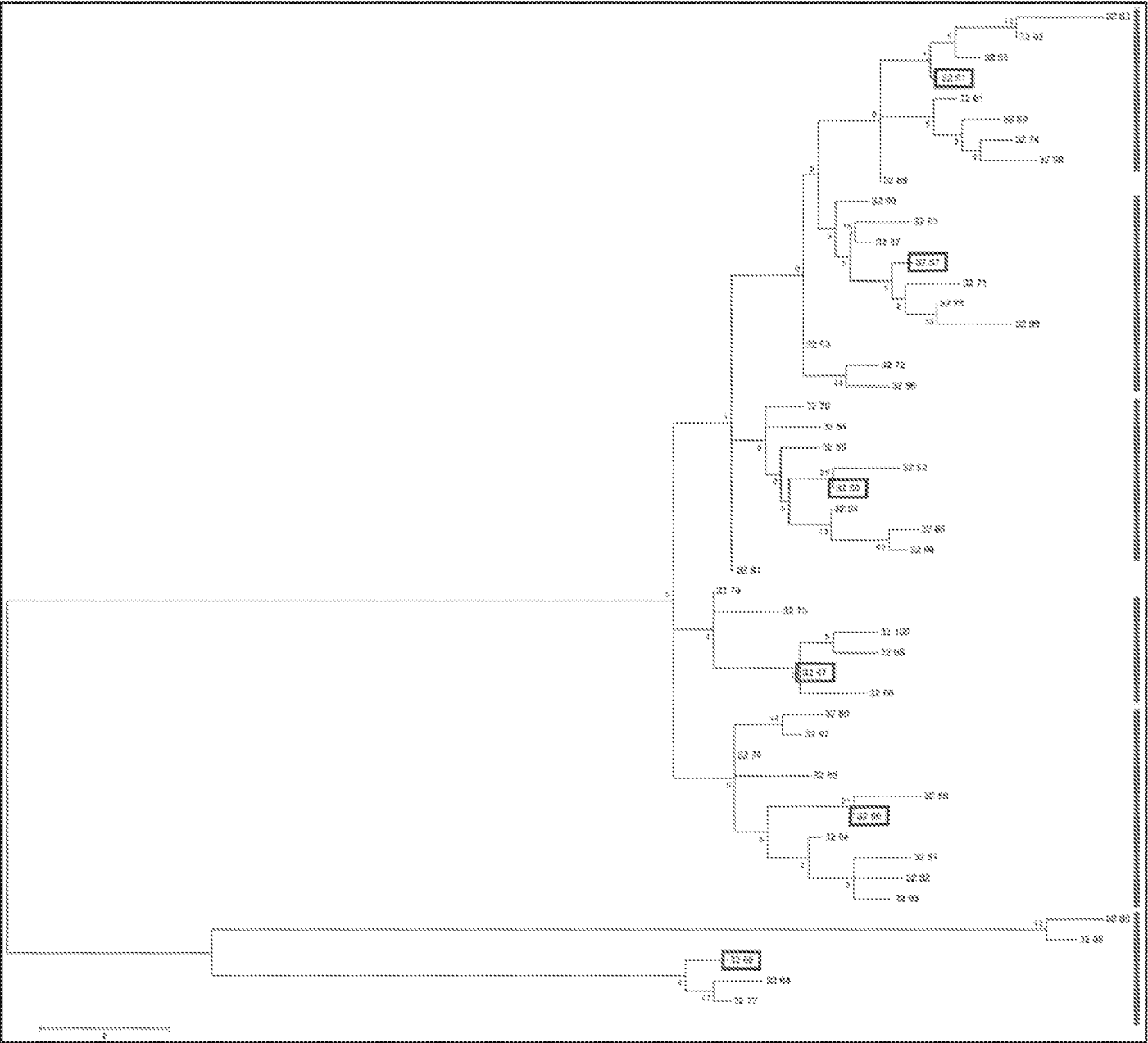


FIGURE 1

name	Sequence	Length	Pval	logP	min free energy (kcal/mol)	Im	uM (by Nanodrop)
seq20A	ACACCCUACUSAUUGUGG	20	A	6448.8	-4.5	52.7	13.1
seq20C	GAUUCGAUCAGCGUGCAGUA	20	A	6348.8	6	47.1	12.1
seq22B	AGGUUCCGSAUAAUAAGACCC	22	A	7134.3	-3.4	58.4	12.2
seq22D	GAAGCAGAUUCGACAUUAU	22	A	7048.2	-1	45.3	12.0
seq22E	GAUGUACGACUUUGGAUCCAA	22	A	7206.3	6	50.5	12.2
seq22G	UASCAUAUJCAGCCUGAGAAC	22	A	7128.3	6	55.7	12.9
seq25A	AUUCGGAUACGCCAGUACUUGUGG	25	A	8038.8	-1.7	52.7	12.1
seq25B	CUAUACGACAGUCUGUAUCUAGUCU	25	A	8038.8	-2.1	55.7	10.8
seq30A	CUAGAUUGAACUUGCCUCCGCCGAUAUUAUUG	30	A	9831.7	8	84.8	11.8
seq30C	UUAUGGUUACAGUACUUCUGGCGUAAAGAUUC	30	A	9831.7	-4.4	51.6	11.3
seq45A	AUGGUCUAUAGGAGUCUGCGCGGUAUUGGUACAGACACUCAC	40	A	12876	-3.8	65	12.2
seq45B	UCUUCUUAAGUAGCGACAGAUUCUACGACGACUAUUAUGCCUU	40	A	12876	-8.8	63.1	12.5
seq50B	CAGUGGGAGUAGCGGUUCAGCUGGCGUUGCUUUAACAAUAAGCAGGUCUU	50	A	15060	-12.5	63.7	12.0
seq50D	UGCAUUGGAGUGCAGUUGGAGACAAACACUACUGCGCCAUUAAGUGGAGUUC	50	A	15060	-9.5	63	11.8
seq50H	UAGUGAGUAGUUGUGUAAGG	20	B	5448.8	6	48	12.1
seq51B1	UCUUCUAGUAGACUGCGAUAC	20	B	6348.8	6	56.4	11.8
seq52A	AUAUUGGUAUCUGCGGUCCGCCAA	22	B	7037.2	-6.4	56.8	11.8
seq52C	CAUUGGGGAACUAUGUGGAA	22	B	7106.2	-9.3	48.1	13.1
seq52F	UAACUGCGUAAGCGGAUUGUGCG	22	B	7902.2	-2.5	58.6	12.1
seq52H	UGAUACGGAUUGUAUACUGGAC	22	B	7122.2	6	52.3	12.3
seq52C	GAUUCGACUCCGUGUAUAUAUUCG	25	B	8038.8	-2.4	55.8	12.2
seq55D	UCUGUUAUGGGGAGGCGAUUAUA	25	B	8038.8	-1.2	57.6	12.0
seq50B	AUGGCGUAUUGCGCGUUAUAACUGGAACU	30	B	9831.7	-2.3	55.6	10.8
seq51B5	GAUUCUAGAACGUUAGCGGAGUGUAAGCCAA	30	B	9830.7	-5.6	53.9	11.2
seq40C	UUNUUAUUCAUUGUGCAGCGUUAUAUUGUGGUAUACCGA	40	B	12876	-5.5	63.3	12.5
seq40D	UACAGUAGGAUACUUGUGGCAUCCAAUGGAGUGGCUUG	40	B	12876	-10	67.3	11.7
seq50A	AUUBUUAACUAUUCUAGCGGAGUUAUGGCUACACGAGAUUCCAGUAGUG	50	B	15060	-6	84.6	12.5
seq50A	UGAUAUAGGAGCGCCUCCAGGCCAUUAUUCUUAUCCAUAGUAGCGCCUUAUA	50	B	15060	-14.1	61.5	12.8

3/5

Name	Pool	Pool A only	Pool B only	Pool A+B	U01 plasma + 20 antibios pool A	U01 plasma + 1 antibios pool A	U01 plasma	IR plasma + 20 antibios pool A	IR plasma + 1 antibios pool A	IR plasma
seq21A	A	8,078	5	4,050	0	4	0	49	0	23
seq20G	A	352,083	23	193,184	1,310	122	5	1,390	308	13
seq22B	A	1,791,388	232	856,270	3,032	598	24	8,799	1,547	185
seq22D	A	25,774	0	13,630	74	8	0	102	5	0
seq22E	A	769,547	20	381,510	3,231	420	13	8,990	1,252	18
seq22G	A	1,011,885	3	547,169	3,543	408	29	10,281	1,073	26
seq25A	A	560,584	28	292,531	8,023	772	8	18,830	1,098	36
seq25B	A	990,190	303	487,350	5,848	733	21	17,889	2,136	38
seq30A	A	841,125	423	426,216	21,863	2,772	29	63,667	7,072	56
seq30C	A	679,062	51	379,220	13,426	1,753	34	48,504	4,139	31
seq40A	A	97,049	42	53,630	8,042	774	5	20,303	2,363	16
seq40B	A	17,376	14	8,576	3,982	493	10	14,281	1,536	7
seq50B	A	5,176	1	1,502	40	9	0	101	11	0
seq50D	A	5,537	14	3,252	108	5	0	218	40	2
seq20J	B	45	738,307	350,780	57	2	2	15	23	1
seq20O	B	7	156,975	77,731	9	0	0	1	3	0
seq22A	B	7	32,599	15,929	1	0	0	0	4	0
seq22C	B	29	908,916	410,277	56	0	3	9	31	4
seq22F	B	105	241,794	120,960	18	1	0	2	5	0
seq22H	B	34	745,910	334,541	58	0	1	5	26	0
seq25C	B	22	269,044	130,952	15	0	0	2	5	0
seq25D	B	44	721,352	299,007	40	0	1	10	25	0
seq30B	B	62	1,352,844	678,469	59	0	0	11	22	0
seq30D	B	137	734,805	345,073	49	1	1	5	15	0
seq40C	B	8	8,965	3,126	0	0	0	0	0	0
seq40E	B	1	7,327	3,287	2	0	0	0	0	0
seq50A	B	19	15,850	8,256	3	0	0	0	5	0
seq50C	B	0	4,640	2,396	1	0	0	0	1	0
Total reads		7,335,710	6,502,398	6,877,452	6,304,227	7,827,270	6,730,108	8,589,130	8,140,456	8,056,994
calibrator (%)		94	95	95	1	0	0	3	0	0
reads used for alignment (%)		2	2	2	89	90	89	71	72	79
genome (%)		1	0	0	12	82	80	48	47	61
mRNA sense (%)		0.0	0.0	0.0	19.3	12.5	15.1	3.1	2.5	4.2

FIGURE 3

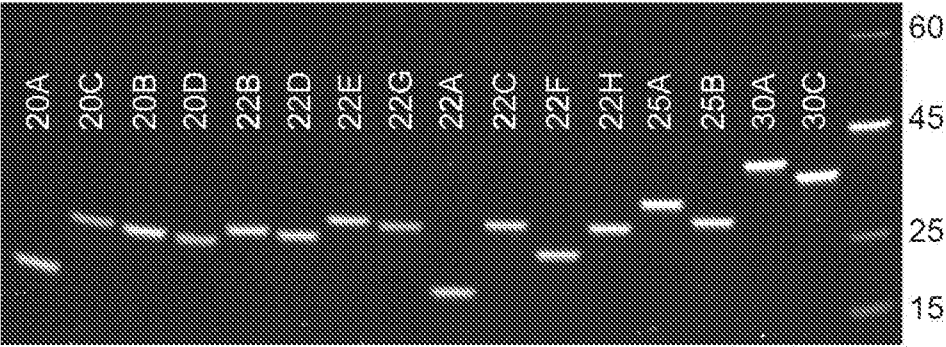


FIGURE 4

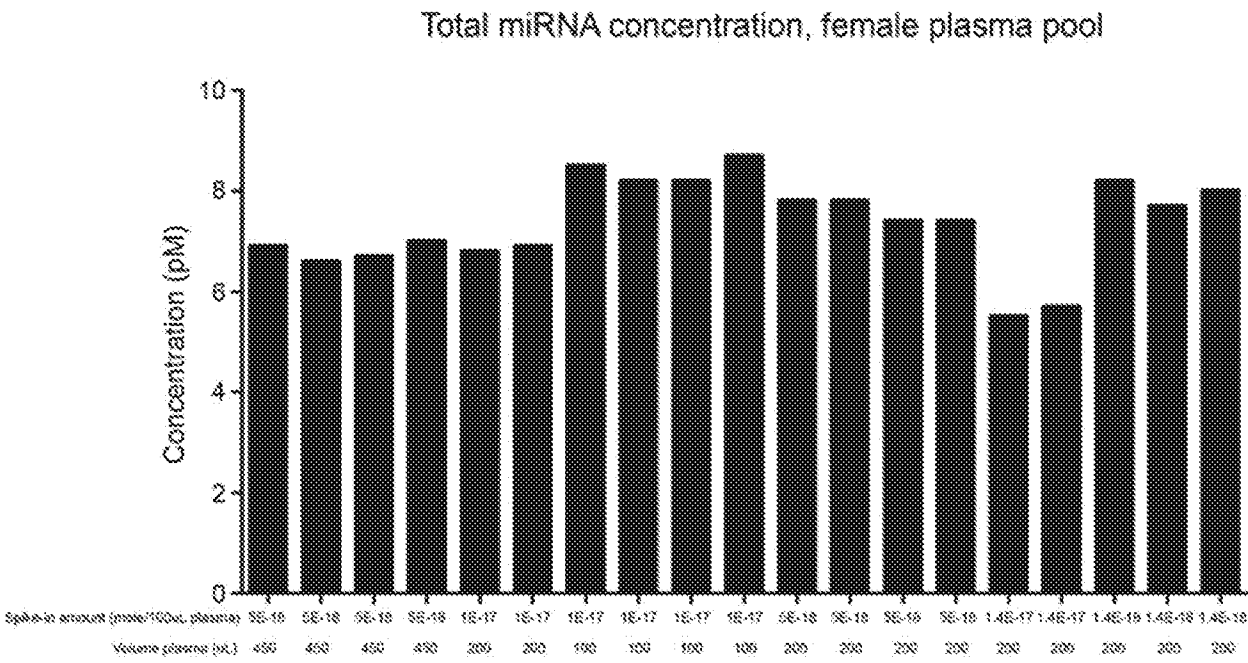


FIGURE 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 18/23191

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C12Q 1/68, G06F 19/22 (2018.01)

CPC - C12Q 1/6874, C12Q 1/6869, C12Q 1/689, G06F 19/22, C12Q 2600/166

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X -- Y -- A	Locati et al., "Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization" Nucleic Acids Research, 2015, Vol. 43, No. 14 e89, page 1-10; abstract, pg 2, col 2, para 4, pg 3, col 1, para 1, pg 5, col 2, para 1, pg 6, col 1, para 2 to col 2, para 1, Fig. 3C	1, 5/1 --- (3-4)/1 --- 2, (3-5)/2
Y	US 2015/0011413 A1 (hereinafter 'Zhang') 08 January 2015 (08.01.2015) para [0054]	(3-4)/1

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

07 August 2018

Date of mailing of the international search report

24 AUG 2018

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents

P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 18/23191

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I+: Claims 1-5, drawn to a set of spike-in RNAs. The set will be searched to the extent that the spike-in RNA encompasses the first oligonucleotide listed in Table 3 (TTGTGCCCTAATCCGT). It is believed that claims 1-5 encompass this first named invention, and thus these claims will be searched without fee to the extent that they encompass the spike-in RNA comprising (TTGTGCCCTAATCCGT). Additional spike-in RNA(s) will be searched upon the payment of additional fees. Applicants must specify the claims that encompass any additionally elected spike-in RNA(s). Applicants must further indicate, if applicable, the claims which encompass the first named invention, if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched. An exemplary election would be a spike-in RNA comprising (ACTCACGGCCGTCTTA).

- Please see extra sheet for continuation -

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-5 limited to the spike-in RNA comprising oligo 16_102089 (TTGTGCCCTAATCCGT)

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

Continuation of:

Box NO. III. Observations where unity of invention is lacking

The inventions listed as Groups I+ do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

No technical features are shared among the spike-in RNA sequences of Groups I+ and, accordingly, these groups lack unity a priori.

Additionally, even if Groups I+ were considered to share the technical features of claims 1 and 2, these shared technical features are previously disclosed by the article "Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization" by Locati et al. (hereinafter "Locati") (Nucleic Acids Research, 2015, Vol. 43, No. 14 e89).

Locati teaches a set of spike-in RNAs comprising one or more artificial short RNA sequences (abstract, The external reference for data-normalization (ERDN) spike-in set, consisting of 19 oligoribonucleotides, was developed for sample ? to - sample normalization in differential-expression analysis of sRNA-seq data.) that

(1) have a base composition that mimics that of endogenous human miRNAs (pg 2, col 2, para 4, A list of candidate random RNA sequences of the desired length was generated using the following constraints: maximum homopolymer length of two nucleotides, GC content between 40 and 60%, .delta.G for hairpin formation > -0.5 kcal/mol at 37C.); (2) have broad sequence diversity; 3) cover a range of sequence lengths of between about 16 to 70 nucleotides (Fig. 3C, Oligoribonucleotide sequences of the 25-mer ERDN spike-in set.); and/or (4) do NOT share sequence identity with known endogenous sequences in human, animal, plant, fungus, bacterial, and/or viral genomes (pg 2, col 2, para 4, an E-value >10 when aligned against the NCBI nucleotide collection (nr/nt: <http://blast.ncbi.nlm.nih.gov>) using BLASTN.).

Locati teaches a set of spike-in RNAs, comprising the one or more, optionally about 10 to 100, artificial short RNA sequences (Fig. 3C, Oligoribonucleotide sequences of the 25-mer ERDN spike-in set.).

As said technical features were known in the art at the time of the invention, these cannot be considered special technical features that would otherwise unify the groups.

Groups I+ therefore lack unity under PCT Rule 13 because they do not share a same or corresponding special technical feature.