



(51) International Patent Classification:

C12Q 1/6869 (2018.01) *C12Q 1/6806* (2018.01)
C40B 30/04 (2006.01) *C12Q 1/6809* (2018.01)
G16B 30/00 (2019.01) *C12N 15/113* (2010.01)
G16B 40/20 (2019.01)

(21) International Application Number:

PCT/US2024/061133

(22) International Filing Date:

19 December 2024 (19.12.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/612,257 19 December 2023 (19.12.2023) US

(71) Applicant: **THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY** [US/US];
Building 170, Third Floor, Main Quad, P.O. Box 20386,
Stanford, CA 94305-2038 (US).

(72) Inventors: **DAS, Rhiju**; 2435 Amherst Street, Palo Alto,
CA 94304 (US). **BLAIR, Hamish, Milson**; Building 170,

Third Floor, Main Quad, P.O. Box 20386, Stanford, CA
94305-2038 (US).

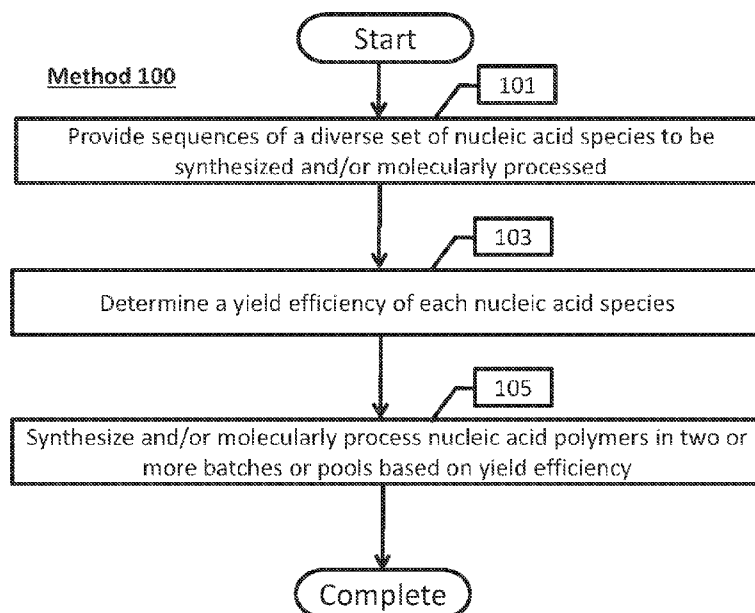
(74) Agent: **THOMAS, Charles, A.**; KPPB LLP, 3780 Kilroy
Airport Way, Suite 320, Long Beach, CA 90806 (US).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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,

(54) Title: SYSTEMS AND METHODS FOR ENHANCING SYNTHESIS OF NUCLEIC ACIDS

Fig. 1



(57) Abstract: Systems and methods for enhancing yield of synthesis and molecular processing of nucleic acid species are described. Nucleic acid species can be synthesized or molecularly processed in batches or pools. The nucleic acid species within a batch or pool have been determined to have a similar yield efficiency with the other nucleic acid species within its batch or pool.

DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT,
LU, LV, MC, ME, 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).

Published:

- *with international search report (Art. 21(3))*
- *with sequence listing part of description (Rule 5.2(a))*

SYSTEMS AND METHODS FOR ENHANCING SYNTHESIS OF NUCLEIC ACIDS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims benefit of U.S. Provisional Patent Application No. 63/612,257, filed December 19, 2023, entitled "Systems and Methods for Enhancing Synthesis of Nucleic Acids," the disclosure of which is hereby incorporated by reference in its entirety for all purposes.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0002] This invention was made with Government support under contract 2R35GM122579 awarded by the National Institutes of Health. The Government has certain rights in the invention.

TECHNICAL FIELD

[0003] The disclosure is generally directed toward systems and methods for enhancing synthesis of nucleic acids.

BACKGROUND

[0004] Bulk de novo synthesis of nucleic acids in which numerous sequences were synthesized in the same pool became popularized in the early 1990s when Affymetrix developed methods for spatially localized polymer synthesis (S. P. Fodor, et al., Science. 1991 Feb 15;251(4995):767-73; and A. C. Pease, et al., Proc Natl Acad Sci U S A. 1994 May 24;91(11):5022-6; the disclosures of which are incorporated herein by reference). Today, several methods are utilized to perform bulk pool synthesis, most typically using a phosphoramidite (or similar) protection protocol. Some techniques utilize mask-based photolithographic techniques to selectively deprotect photolabile nucleoside phosphoramidites. Other techniques utilize maskless procedures using programmable micromirror devices (e.g., S. Singh-Gasson, et al., Nat Biotechnol. 1999 Oct;17(10):974-

8, the disclosure of which is incorporated herein by reference). Some light-independent techniques include inkjet printing of nucleotides on an arrayed surface, selective electrochemical deprotection of nucleotides, and the use of microfluidics to control delivery of components for synthesis (see, e.g., R. J. Lipshutz, et al., *Nat Genet.* 1999 Jan;21(1 Suppl):20-4; T. R. Hughes, et al., *Nat Biotechnol.* 2001 Apr;19(4):342-7; and A. L. Ghindilis, et al., *Biosens Bioelectron.* 2007 Apr 15;22(9-10):1853-60; and C. C. Lee, et al., *Nucleic Acids Res.* 2010 May;38(8):2514-21; the disclosures of which is incorporated herein by reference).

SEQUENCE LISTING

[0005] This application hereby incorporates by reference the material of the electronic Sequence Listing filed concurrently herewith. The material in the electronic Sequence Listing is submitted as an XML file entitled "08940PCT.xml" created on December 19, 2024, which has a file size of 4.5 KB, and is hereby incorporated by reference in its entirety.

SUMMARY

[0006] Several embodiments are directed towards systems and methods for bulk synthesis of nucleic acids comprising a diversity of sequences. In several embodiments, a set of sequences to be synthesized is provided. In many embodiments, a yield efficiency for each sequence of the set of sequences of nucleic acid species to be synthesized is determined. In several embodiments, the nucleic acid species are synthesized in two or more separate pools. In many embodiments, each sequence of the set of sequences of nucleic acid species within each pool is determined based on its yield efficiency. In several embodiments, a yield efficiency for each sequence is determined utilizing data from a prior synthesis reaction. In many embodiments, a yield efficiency for each sequence is determined by performing a pilot synthesis reaction. In several embodiments, a yield efficiency for each sequence is determined by entering each sequence of the set of sequences of nucleic acid species to be synthesized into a trained predictive model.

[0007] In some aspects, the techniques described herein relate to a method to improve generation of diverse pools of nucleic species, including: providing a set of nucleic acid species to be synthesized, wherein each nucleic acid species within the set has a unique sequence; determining a yield efficiency for each nucleic acid sequence of the set of sequences of nucleic acid species to be synthesized; and generating the nucleic acid species in two or more batches, wherein each sequence of the set of sequences of nucleic acid species within each batch is determined based on its determined yield efficiency.

[0008] In some aspects, the techniques described herein relate to a method, wherein the step of synthesizing the nucleic acid species in two or more batches includes purification.

[0009] In some aspects, the techniques described herein relate to a method, wherein the step of synthesizing the nucleic acid species in two or more batches includes downstream processing.

[0010] In some aspects, the techniques described herein relate to a method, wherein the step of determining a yield efficiency for each sequence includes utilizing data from a prior synthesis reaction or data from a prior purification or data from prior downstream processing.

[0011] In some aspects, the techniques described herein relate to a method, wherein the step of determining a yield efficiency for each sequence includes performing a pilot synthesis reaction or a pilot purification or a pilot run of downstream processing.

[0012] In some aspects, the techniques described herein relate to a method, wherein the step of determining a yield efficiency for each sequence includes entering each sequence of the set of sequences of nucleic acid species to be synthesized into a trained predictive model.

[0013] In some aspects, the techniques described herein relate to a method, wherein the number of nucleic acid species in the set is at least 1000.

[0014] In some aspects, the techniques described herein relate to a method, wherein the number of nucleic acid species in the set is at least 100,000.

[0015] In some aspects, the techniques described herein relate to a method, wherein each nucleic acid species of the set is between 5 and 500 nucleosides.

[0016] In some aspects, the techniques described herein relate to a method, wherein each nucleic acid species of the set have a number of nucleosides not more than 500% of the shortest nucleic acid species of the set.

[0017] In some aspects, the techniques described herein relate to a method, wherein each nucleic acid species of the set have an equivalent number of nucleosides.

[0018] In some aspects, the techniques described herein relate to a method, wherein yield efficiency includes the efficiency of nucleic acid species chemical synthesis.

[0019] In some aspects, the techniques described herein relate to a method, wherein the chemical synthesis includes phosphoramidite synthesis.

[0020] In some aspects, the techniques described herein relate to a method, wherein the chemical synthesis includes multiplex microarray assembly.

[0021] In some aspects, the techniques described herein relate to a method, wherein yield efficiency includes the efficiency of one or more molecular processes.

[0022] In some aspects, the techniques described herein relate to a method, wherein the one or more molecular processes includes one or more of: replication, transcription, reverse transcription, translation, amplification, complement-based hybridization, cofactor binding, or introduction into cell.

[0023] In some aspects, the techniques described herein relate to a method, wherein the step of generating the nucleic acid species in two or more batches includes performing an independent chemical synthesis reaction for each batch of the two or more batches.

[0024] In some aspects, the techniques described herein relate to a method, wherein the step of generating the nucleic acid species in two or more batches includes performing dial-out PCR.

[0025] In some aspects, the techniques described herein relate to a method further including performing a downstream application, wherein the downstream application includes one or more of: antisense oligomer analysis, RNA-interference molecule analysis, short-hairpin RNA molecule analysis, miRNA molecule analysis, mutagenesis with guide-RNA molecules for CRISPR, aptamer analysis, in situ hybridization molecules, microarray analysis, mutated sequence analysis, assembly of genes, assembly of genomes, nucleic-acid based information storage, generation of protein medicines

encoded by nucleic acids, generation of nucleic acid medicines, generation of sequencing library, or hybridization capture.

[0026] In some aspects, the techniques described herein relate to a method further including performing mutational analysis profiling sequencing (MaP-seq).

BRIEF DESCRIPTION OF THE DRAWINGS

[0027] The description and claims will be more fully understood with reference to the following figures and data graphs, which are presented as exemplary embodiments of the disclosure and should not be construed as a complete recitation of the scope of the disclosure.

[0028] Figure 1 provides a flow chart for generating a diverse batch of nucleic acid species.

[0029] Figure 2 provides a chart depicting the benefit of thoughtful separation of nucleic acid species for bulk generation. The number of sequences with final sequencing read number above a threshold of 1000, as a function of the way a library of 1 M sequences was split into two pools, where the oracle is a perfect oracle. Our experimental workflow involved reading out data with sequencing with final output of 1B reads, or 1000 reads on average.

[0030] Figure 3 provides a chart predicting the benefit of different split values for nucleic acid species for bulk generation. The number of sequences above a threshold of 1000, as a function of the way the library was split into two pools, where the oracle is a linear regression function. The vertical lines correspond to the best splits as predicted by the model.

[0031] Figure 4 provides a schematic of mutational profile sequencing (MaP-seq).

[0032] Figure 5 provides data graph showing of MaP-seq read-depth results when probing 2000 sequences and 1,000,000 sequences.

[0033] Figure 6 provides schematics for primer binding sequence tags to perform dial-out PCR. SEQ ID NOs: 1-4.

[0034] Figure 7 provides correlation of predictive model results with MaP-seq read-depth results.

[0035] Figure 8 provides performance and accuracy of predictive models for splitting nucleic acid species into pools.

[0036] Figure 9 provides MaP-seq read-depth results comparing no splitting of nucleotide species with splitting the nucleotide species into three pools.

DETAILED DESCRIPTION

[0037] Turning now to the drawings and data, systems and methods of the disclosure provide enhanced bulk de novo synthesis of nucleic acid species of diverse sequences. In several embodiments, a large diversity of sequences that are selected for synthesis are analyzed for their ability to be synthesized efficiently. In many embodiments, the sequences are separated into two or more groupings based on their efficiency to be synthesized. In these various embodiments, numerous diverse nucleic acid species are synthesized together based on their yield efficiency such that polymer sequences determined to have similar efficiency are synthesized together.

[0038] Current techniques for the synthesis of a nucleic acid sequence library rely on synthesizing all DNA molecules (or DNA templates for RNA) comprising a vast diversity in molecule sequences in one large batch. Bulk synthesis processes, however, do not ensure each sequence is synthesized in equal amounts. Some sequences synthesize more readily than others resulting in unequal concentrations of sequences. Even if sequence concentrations are near-equal at the initial bulk synthesis step, strong differences in the products of the sequences typically appear as the sequences are subjected downstream processes like PCR amplification, library cloning, various sequencing techniques, in vitro or in cellulo transcription, reverse transcription, introduction into cells, tissues, or organisms, translation, and intermediate and final purification steps. The resultant pool of nucleic acids ends up dominated by a few ‘winner’ sequences.

[0039] To illustrate the problem in an example, a common synthesis technique was utilized to generate 1,000,000 unique sequences for a sequencing library. The originally synthesized library was used in an experimental protocol for performing mutational profiling sequencing (MaP-Seq), which is a technique that can assess RNA secondary

structure dynamics. In this example, the protocol included transcription of RNA from the diversely synthesized DNA library, applying an experiment condition on the RNA (i.e., to determine how the experimental condition affects RNA secondary structure), reverse transcription of RNA, amplification of the reverse-transcribed RNA by polymerase chain reaction (PCR), and then sequencing via next-generation sequencing platforms. After these steps, the sequencing results indicated that the synthesized prepared sample utilized for sequencing reaction was dominated by 10% of the sequences of the diverse sequencing library. The other 90% of sequences (900,000 sequences) had poor, uninterpretable results and low read depth and thus the protocol resulted in significant reagent and financial waste.

[0040] Prior methodologies for generating pools of nucleic acids having great sequence diversity did not consider efficiency of synthesis and/or molecular processing. Instead, these methodologies kept the number of diverse species relatively small (e.g., less than 10,000 unique sequences) to obtain interpretable results on the majority of sequences. And if more sequence diversity was desired, multiple small batches (e.g., less than 10,000 unique sequences) were utilized. Furthermore, splitting of the nucleic acid species into multiple batches was randomized (or at least not based on computing an efficiency). Because of these issues, there is a need to develop protocols that enable generation of a pool of nucleic acid species with more sequence diversity (e.g., greater than 10,000 sequences). Furthermore, when a protocol would benefit from utilization of two or more pools of nucleic acid species, there is need to develop a means for intelligent splitting of the nucleic acid species for synthesis and/or molecular processing.

[0041] Herein, several embodiments of the disclosure are directed to systems and methods to generate pools of nucleic acids having great sequence diversity. In many embodiments, nucleic acid species are split into two or more pools based on yield efficiency of synthesis and/or molecular processing, which is dependent on its sequences of nucleobases. As demonstrated by the examples described herein, a computational model can be trained to intelligently split nucleic acid species into two or more pools based on nucleobase sequence the synthesis. In some particular implementations, a computational is trained utilizing data of diverse nucleic acid species pools and the ability

of the sequences to yield an interpretable result after one or more process steps of synthesis and molecular processing.

[0042] Several embodiments are directed to systems and methods of enhanced bulk synthesis and/or generation of pools of nucleic acid species comprising a diversity of sequences. In many embodiments, the efficiency to synthesize and/or molecularly process each unique sequence of nucleic acid species is determined. In several embodiments, synthesis and/or molecular processing is performed in two or more batches or pools, where each batch or pool is comprised of sequences determined to have a similar efficiency. Thus, a more equal concentration of nucleic acid polymer species with diverse sequences within a pool and/or assessed in an end-product result can be achieved.

[0043] Provided in Fig. 1 is an example of a method to enhance yield of bulk synthesis and/or molecular processing of nucleic acid species comprising a diversity of sequences. Method 100 can provide (101) sequences of a diverse set of nucleic acid species to be synthesized and/or molecularly processed, each nucleic acid species having a unique sequence. The nucleic acids polymer sequences can comprise deoxyribonucleotides, ribonucleotides, modified nucleotides, synthetic nucleotides, or any other base capable of being molecularly processed and/or synthesized utilizing nucleic acid synthesis protocol.

[0044] Molecular processing is any molecular technique performed on nucleic acids, especially techniques that have varied efficiencies based on sequence. Examples of molecular processes upon nucleic acids include (but are not limited to) replication, transcription, reverse transcription, translation, amplification, complement-based hybridization, cofactor binding, introduction into cell, etc. Several specific molecular techniques can be contemplated. For example, nucleic acids can be replicated and/or amplified by a number of techniques such as PCR, loop-mediated isothermal amplification (LAMP), rolling circle amplification (RCA), nucleic acid sequence-based amplification (NASBA), rolling circle replication (RCR), helicase-dependent amplification (HDA), strand displacement amplification (SDA), multiple displacement amplification (MDA), recombinase polymerase amplification (RPA), transcription-mediated

amplification (TMA), etc. Complement-based hybridization is any technique in which hybridization of one nucleic acid to another nucleic acid base on at least partial complementation, such as (for example) target sequence capture and primer binding. Cofactor binding is any technique that utilizes a non-nucleic acid species to directly or indirectly bind a nucleic acid, especially DNA-binding proteins and RNA-binding proteins but also techniques that utilize antigen-binding proteins that target nucleic acids, ions (e.g., Zn^{++} , Fe^{++}), organic molecules (e.g., caffeine, theophylline), and vitamins (e.g., vitamin C, vitamin D).

[0045] The number of nucleic acid species with unique sequences to be synthesized and/or molecularly processed can be any number larger than 1, but especially greater than 1000 (approximate species number that diversity effects on efficiency become problematic for synthesis and processing). In various embodiments, the number of unique sequences to be synthesized and is at least 2 unique sequences, at least 5 unique sequences, at least 10 unique sequences, at least 50 unique sequences, at least 100 unique sequences, at least 500 unique sequences, at least 1000 unique sequences, at least 5000 unique sequences, at least 10,000 unique sequences, at least 50,000 unique sequences, at least 100,000 unique sequences, at least 500,000 unique sequences, at least 1,000,000 unique sequences, at least 5,000,000 unique sequences, at least 10,000,000 unique sequences, at least 50,000,000 unique sequences, at least 100,000,000 unique sequences, at least 500,000,000 unique sequences, or at least 1,000,000,000 unique sequences.

[0046] The length of each nucleic acid species can generally be between 5 and 500 nucleosides, but longer sequences can be contemplated. In some embodiments, the length of each nucleic acid species is any length capable of being synthesized by the synthesis protocol being utilized, which may include other molecular processing. For example, phosphoramidite synthesis can generally synthesize nucleic acid species of about 200 nucleosides, which can be used in further extension, amplification, hybridization, etc. In various embodiments, the length of nucleic acid species is at least 5 nucleosides, at least 10 nucleosides, at least 20 nucleosides, at least 30 nucleosides, at least 40 nucleosides, at least 50 nucleosides, at least 60 nucleosides, at least 70

nucleosides, at least 80 nucleosides at least 90 nucleosides, at least 100 nucleosides, at least 110 nucleosides, at least 120 nucleosides, at least 130 nucleosides, at least 140 nucleosides, at least 150 nucleosides, at least 160 nucleosides, at least 170 nucleosides, at least 180 nucleosides, at least 190 nucleosides, at least 200 nucleosides, at least 300 nucleosides, at least 400 nucleosides, at least 500 nucleosides, at least 1000 nucleosides, at least 1500 nucleosides, at least 2000 nucleosides, at least 5000 nucleosides, at least 10000 nucleosides, at least 20000 nucleosides, at least 50000 nucleosides, at least 100,000 nucleosides, at least 200,000 nucleosides, at least 500,000 nucleosides, at least 1,000,000 nucleosides, at least 10,000,000 nucleosides, at least 100,000,000 nucleosides, at least 1,000,000,000 nucleosides, at least 10,000,000,000 nucleosides, at least 100,000,000,000 nucleosides, at least 1,000,000,000,000 nucleosides. In some implementations, each nucleic acid species to be synthesized and/or molecularly processed have an equivalent number of nucleosides. In some implementations, each nucleic acid species to be synthesized and/or molecularly processed have inequivalent lengths. In some implementations, each nucleic acid species to be synthesized and/or molecularly processed have similar number of nucleosides, such that all nucleic acid species have a length that is not more than a percentage than the shortest species. In some implementations, each nucleic acid species to be synthesized and/or molecularly processed has a number of nucleosides: not more than 10% of the shortest nucleic acid species, not more than 20% of the shortest nucleic acid species, not more than 30% of the shortest nucleic acid species, not more than 40% of the shortest nucleic acid species, not more than 50% of the shortest nucleic acid species, not more than 100% of the shortest nucleic acid species, not more than 200% of the shortest nucleic acid species, not more than 500% of the shortest nucleic acid species.

[0047] Method 100 can further determine (103) a yield efficiency of each nucleic species. Efficiency can consider one or more of: chemical synthesis, purification techniques, and/or molecular processing. In several embodiments, a yield efficiency is a metric of how efficient a particular nucleic acid species having sequence can be synthesized and/or molecular processed by a given protocol, which can comprise one or

more of: chemical synthesis, purification techniques, and/or molecular processing. In some embodiments, the yield efficiency is a metric of how efficiently a particular nucleic acid species having a sequence yields an interpretable result after performing a protocol to yield the result. For example, an interpretable result can be sequencing depth for a sequencing protocol, but any interpretable result affected by diverse pools of nucleic acid species can be assessed. Molecular processing can include replication, transcription, reverse transcription, translation, amplification, complement-based hybridization, cofactor binding, introduction into cell, etc. Generally, a relative efficiency is determined, where an efficiency metric is determined relative to other sequences of that are to be synthesized.

[0048] When determining yield efficiency of each nucleic species, it should be understood that a whole and/or partial synthesis and/or molecular processing protocol can be utilized. Yield efficiency of a protocol in some protocols, for example, is heavily influenced by one or two steps of the protocol. It may be beneficial streamline the efficiency assessment to those one or two steps. Accordingly, in some implementations, a yield efficiency is determined for one or more steps of protocol, but not the entire protocol. The one or more steps can include any of: the initial step, one or more intermediate steps, and the final step. If two or more steps are utilized to determine yield efficiency, the two or more steps are not required to be immediately sequential (e.g., the 2nd and 4th step can be utilized). In some implementations, the one or more steps of the protocol to determine yield efficiency includes the entire protocol from the initial step to end result. Yield efficiency assessment of entire protocols may be easier to determine as only the input and end result can be utilized to determine efficiency.

[0049] Various means can be utilized to determine yield efficiency. In some embodiments, a prior synthesis run (or portion thereof) is utilized to determine yield efficiency. In some embodiments, a pilot synthesis run (or portion thereof) is utilized to determine yield efficiency. In these embodiments, data of the prior and/or pilot run (or portion thereof) is utilized to determine which sequences are yielded efficiently. In some embodiments, the amount and/or quality of synthesis and/or molecular processing is determined for each sequence. In some embodiments, one or more thresholds and/or

clustering techniques can be utilized to group and/or cluster sequences into efficiency grouping.

[0050] In some embodiments, a predictive computational model is utilized to determine yield efficiency. A classifier and/or regressor can be trained to predict yield efficiency based for sequence. Accordingly, previously determined yield efficiency result of synthesis and/or molecular processing of nucleic acid sequences can be utilized to train a predictive model to learn which sequence patterns are more easily and more difficult to synthesize and/or molecularly process by a protocol (or a portion thereof). As noted, yield efficiency results can be based on any of one or more steps of the protocol. Accordingly, a yield efficiency result can be a final efficiency result or an intermediate efficiency result of the protocol. Once a computational model is trained, nucleic acid species sequences to be generated can be entered into the model as features to predict yield efficiency.

[0051] Any computational classifier and/or regressor can be utilized. Predictive computational models that can be utilized include (but are not limited to) linear regression, polynomial regression, Cox proportional hazards regression, multiple linear regression, ridge regression, logistic regression, Lasso, stepwise regression, principal component analysis, Bayesian inference, elastic net, random forest regression, and neural network. In some embodiments, various neural networks and/or sequential models are utilized to better handle various sequence sizes. For example, a combination of one or more of deep neural networks (DNN), convolutional neural networks (CNN), recurrent neural networks, transformer neural network, long short-term memory (LSTM) networks, kernel ridge regression (KRR), and/or gradient-boosted random forest decision trees can be utilized. In some embodiments, a foundation model or large language model is used to extract a latent embedding of nucleoside sequences. In some embodiments, an ensemble of computational models is utilized.

[0052] Method 100 can further synthesize (105) and/or molecularly process nucleic acid species in two or more batches or pools based on yield efficiency of the nucleic acid species. Nucleic acid species with lower efficiency can be synthesized in one pool and polymers with higher efficiency can be synthesized in another pool. Further, more than

two batches of synthesis can be performed, each batch grouping nucleic acid species with similar yield efficiencies. In some embodiments, one or more thresholds are utilized to separate nucleic acid species into the two or more batches or pools.

[0053] Upon synthesis and processing, the nucleic acid species can be utilized in a variety of downstream applications. In some embodiments, two or more batches or pools of synthesized and/or molecularly processed nucleic acid species are combined for downstream applications. Generally, any application that can utilize a diversity of nucleic acid species with unique sequences can be performed. Examples of applications include (but are not limited to) antisense oligomer analysis, RNA-interference molecule analysis, short-hairpin RNA molecule analysis, miRNA molecule analysis, mutagenesis with guide-RNA molecules for CRISPR, aptamer analysis, in situ hybridization molecules, microarray analysis, mutated sequence analysis, assembly of genes, assembly of genomes, nucleic-acid based information storage, generation of protein medicines encoded by nucleic acids, generation of nucleic acid medicines, generation of sequencing library, hybridization capture, etc.

Nucleic Acid Embodiments

[0054] In certain embodiments, the disclosure provides nucleic acids, which consist of linked nucleosides. Nucleic acids (RNA or DNA) may be unmodified or may be modified. Modified nucleic acids comprise at least one modification relative to unmodified RNA or DNA (i.e., comprise at least one modified nucleoside (comprising a modified sugar moiety and/or a modified nucleobase) and/or at least one modified internucleoside linkage).

[0055] Modified nucleosides can comprise a modified sugar moiety or a modified nucleobase or both a modified sugar moiety and a modified nucleobase.

[0056] In some embodiments, modified sugar moieties are non-bicyclic modified sugar moieties. In some embodiments, modified sugar moieties are bicyclic or tricyclic sugar moieties. In some embodiments, modified sugar moieties are sugar surrogates. Such sugar surrogates may comprise one or more substitutions corresponding to those of other types of modified sugar moieties.

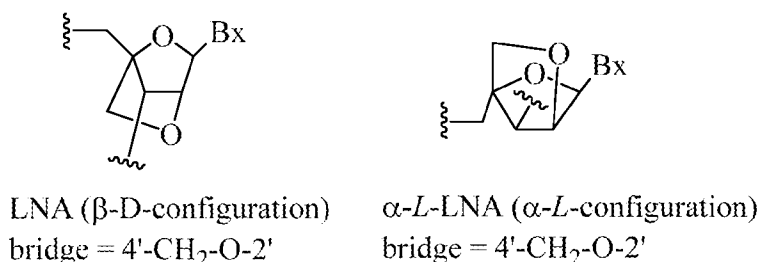
[0057] In some embodiments, modified sugar moieties are non-bicyclic modified sugar moieties comprising a furanosyl ring with one or more acyclic substituent, including but not limited to substituents at the 2', 4', and/or 5' positions. In some embodiments one or more acyclic substituent of non-bicyclic modified sugar moieties is branched. Examples of 2'-substituent groups suitable for non-bicyclic modified sugar moieties include but are not limited to: 2'-F, 2'-OCH₃ ("OMe" or "O-methyl"), and 2'-O(CH₂)₂OCH₃ ("MOE"). In some embodiments, 2'-substituent groups are selected from among: halo, allyl, amino, azido, SH, CN, OCN, CF₃, OCF₃, O-C₁-C₁₀ alkoxy, O-C₁-C₁₀ alkyl, O-C₁-C₁₀ substituted alkyl, S-alkyl, etc. Certain embodiments of these 2'-substituent groups can be further substituted with one or more substituent groups independently selected from among: hydroxyl, amino, alkoxy, carboxy, benzyl, phenyl, nitro (NO₂), thiol, thioalkoxy, thioalkyl, halogen, alkyl, aryl, alkenyl and alkynyl. Examples of 4'-substituent groups suitable for non-bicyclic modified sugar moieties include but are not limited to alkoxy (e.g., methoxy), alkyl, etc. Examples of 5'-substituent groups suitable for non-bicyclic modified sugar moieties include but are not limited to: 5'-methyl (R or S), 5'-vinyl, 5'-methoxy, etc.

[0058] Nucleosides comprising modified sugar moieties, such as non-bicyclic modified sugar moieties, may be referred to by the position(s) of the substitution(s) on the sugar moiety of the nucleoside. For example, nucleosides comprising 2'-substituted or 2'-modified sugar moieties are referred to as 2'-substituted nucleosides or 2'-modified nucleosides.

[0059] In some embodiments, modified sugar moieties comprise a bridging sugar substituent that forms a second ring resulting in a bicyclic sugar moiety. In such embodiments, the bicyclic sugar moiety comprises a bridge between the 4' and the 2' furanose ring atoms. Examples of such 4' to 2' bridging sugar substituents include but are not limited to: 4'-CH₂-2', 4'-(CH₂)₂-2', 4'-(CH₂)₃-2', 4'-CH₂-O-2' ("LNA"), 4'-CH₂-S-2', 4'-(CH₂)₂-O-2' ("ENA"), 4'-CH(CH₃)-O-2' (referred to as "constrained ethyl" or "cEt" when in the S configuration), 4'-CH₂-O-CH₂-2', 4'-CH₂-N(R)-2', etc.

[0060] In certain embodiments, bicyclic sugar moieties and nucleosides incorporating such bicyclic sugar moieties are further defined by isomeric configuration. For example,

an LNA nucleoside (described herein) may be in the α -L configuration or in the β -D configuration.



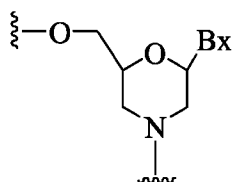
[0061] α -L-methyleneoxy (4'-CH₂-O-2') or α -L-LNA bicyclic nucleosides have been incorporated into antisense compounds that showed antisense activity (Frieden et al., *Nucleic Acids Research*, 2003, 21, 6365-6372, the disclosure of which is incorporated herein by reference). Herein, general descriptions of bicyclic nucleosides include both isomeric configurations. When the positions of specific bicyclic nucleosides (e.g., LNA or cEt) are identified in exemplified embodiments herein, they are in the β -D configuration, unless otherwise specified.

[0062] In some embodiments, modified sugar moieties comprise one or more non-bridging sugar substituent and one or more bridging sugar substituent (e.g., 5'-substituted and 4'-2' bridged sugars).

[0063] In some embodiments, modified sugar moieties are sugar surrogates. In certain such embodiments, the oxygen atom of the sugar moiety is replaced, e.g., with a sulfur, carbon or nitrogen atom. In such embodiments, modified sugar moieties can also comprise bridging and/or non-bridging substituents as described herein. For example, certain sugar surrogates comprise a 4'-sulfur atom and a substitution at the 2'-position and/or the 5' position.

[0064] In some embodiments, sugar surrogates comprise rings having other than 5 atoms. For example, in certain embodiments, a sugar surrogate comprises a six-membered tetrahydropyran ("THP"). Such tetrahydropyrans may be further modified or substituted. Nucleosides comprising such modified tetrahydropyrans include but are not limited to hexitol nucleic acid ("HNA"), anitol nucleic acid ("ANA"), manitol nucleic acid ("MNA") (see, e.g., Leumann, C.J. *Bioorg. & Med. Chem.* 2002, 10, 841-854, the disclosure of which is incorporated herein by reference).

[0065] In some embodiments, sugar surrogates comprise rings having more than 5 atoms and more than one heteroatom. For example, nucleosides comprising morpholino sugar moieties and their use in nucleic acids have been reported (see, e.g., Braasch et al., *Biochemistry*, 2002, 41, 4503-4510, the disclosure of which is incorporated herein by reference). As used here, the term “morpholino” means a sugar surrogate having the following structure:



[0066] In some embodiments, sugar surrogates comprise acyclic moieties. Examples of nucleosides and nucleic acids comprising such acyclic sugar surrogates include but are not limited to: peptide nucleic acid (“PNA”), acyclic butyl nucleic acid (see, e.g., Kumar et al., *Org. Biomol. Chem.*, 2013, 11, 5853-5865, the disclosure of which is incorporated herein by reference).

[0067] Many other bicyclic and tricyclic sugar and sugar surrogate ring systems are known in the art that can be used in modified nucleosides).

[0068] In some embodiments, modified nucleic acids comprise one or more nucleosides comprising a modified nucleobase (also referred to as nucleobase analogs). In some embodiments, modified nucleic acids comprise one or more nucleosides that do not comprise a nucleobase, referred to as an abasic nucleoside.

[0069] In various embodiments, modified nucleobases are selected from: 5-substituted pyrimidines, 6-azapyrimidines, alkyl or alkynyl substituted pyrimidines, alkyl substituted purines, and N-2, N-6 and O-6 substituted purines. In various embodiments, modified nucleobases are selected from: 2-aminopropyladenine, 5-hydroxymethyl cytosine, xanthine, hypoxanthine, 2-aminoadenine, 6-N-methylguanine, 6-N-methyladenine, 2-propyladenine, 2-thiouracil, 2-thiothymine and 2-thiocytosine, 5-propynyl ($-C\equiv C-CH_3$) uracil, 5-propynylcytosine, 6-azouracil, 6-azocytosine, 6-azothymine, 5-ribosyluracil (pseudouracil), 4-thiouracil, 8-halo, 8-amino, 8-thiol, 8-thioalkyl, 8-hydroxyl, 8-aza and other 8-substituted purines, 5-halo, particularly 5-bromo, 5-trifluoromethyl, 5-halouracil, and 5-halocytosine, 7-methylguanine, 7-methyladenine, 2-

F-adenine, 2-aminoadenine, 7-deazaguanine, 7-deazaadenine, 3-deazaguanine, 3-deazaadenine, 6-N-benzoyladenine, 2-N-isobutyrylguanine, 4-N-benzoylcytosine, 4-N-benzoyluracil, 5-methyl 4-N-benzoylcytosine, 5-methyl 4-N-benzoyluracil, universal bases, hydrophobic bases, promiscuous bases, size-expanded bases, and fluorinated bases. Further modified nucleobases include tricyclic pyrimidines, such as 1,3-diazaphenoxazine-2-one, 1,3-diazaphenothiazine-2-one and 9-(2-aminoethoxy)-1,3-diazaphenoxazine-2-one (G-clamp). Modified nucleobases may also include those in which the purine or pyrimidine base is replaced with other heterocycles, for example 7-deaza-adenine, 7-deazaguanosine, 2-aminopyridine and 2-pyridone. Further nucleobases include those disclosed in Sanghvi, Y.S., Chapter 15, *Antisense Research and Applications*, Crooke, S.T. and Lebleu, B., Eds., CRC Press, 1993, 273-288; and those disclosed in Chapters 6 and 15, *Antisense Drug Technology*, Crooke S.T., Ed., CRC Press, 2008, 163-166 and 442-443; the disclosures of which are incorporated herein by reference.

[0070] In some embodiments, nucleosides of modified nucleic acids may be linked together using any internucleoside linkage. The two main classes of internucleoside linking groups are defined by the presence or absence of a phosphorus atom. Representative phosphorus-containing internucleoside linkages include but are not limited to phosphates, which contain a phosphodiester bond ("P=O") (also referred to as unmodified or naturally occurring linkages), phosphotriesters, methylphosphonates, phosphoramidates, and phosphorothioates ("P=S"), and phosphorodithioates ("HS-P=S"). Representative non-phosphorus containing internucleoside linking groups include but are not limited to methylenemethylimino (-CH₂-N(CH₃)-O-CH₂-), thiodiester, thionocarbamate (-O-C(=O)(NH)-S-); siloxane (-O-SiH₂-O-); and N,N'-dimethylhydrazine (-CH₂-N(CH₃)-N(CH₃)-). Modified internucleoside linkages, compared to naturally occurring phosphate linkages, can be used to alter, typically increase, nuclease resistance of the nucleic acid. In some embodiments, internucleoside linkages having a chiral atom can be prepared as a racemic mixture, or as separate enantiomers. Representative chiral internucleoside linkages include but are not limited to alkylphosphonates and phosphorothioates. Methods of preparation of phosphorous-

containing and non-phosphorous-containing internucleoside linkages are well known to those skilled in the art.

[0071] In some embodiments, a nucleic acid comprises a neutral internucleoside linkage, such as (for example) phosphotriesters, methylphosphonates, MMI (3'-CH₂-N(CH₃)-O-5'), amide-3 (3'-CH₂-C(=O)-N(H)-5'), amide-4 (3'-CH₂-N(H)-C(=O)-5'), formacetal (3'-O-CH₂-O-5'), methoxypropyl, and thioformacetal (3'-S-CH₂-O-5'). Further neutral internucleoside linkages include nonionic linkages comprising siloxane (dialkylsiloxane), carboxylate ester, carboxamide, sulfide, sulfonate ester and amides. Further neutral internucleoside linkages include nonionic linkages comprising mixed N, O, S and CH₂ component parts.

[0072] In various embodiments, nucleic acids (including modified nucleic acids) can have any of a variety of ranges of lengths. In some embodiments, a nucleic acid consists of X to Y linked nucleosides, where X represents the fewest number of nucleosides in the range and Y represents the largest number nucleosides in the range. In certain such embodiments, X and Y are each an integer between 5 and 250; provided that X≤Y.

[0073] In some embodiments, nucleic acids (including modified nucleic acids) are circular. Circular nucleic acids (CNAs) are covalently closed polymeric molecules and can comprise any nucleobases as described herein. CNAs can be covalently closed in any manner. In some embodiments, the 5'-terminus is covalently linked with the 3'-terminus. In some embodiments, the 5'-terminus and/or 3'-terminus is covalently linked with an internal nucleobase or adduct extending from an internal nucleobase. In some embodiments, a CNA (or a portion thereof) has functional activity, such as any functional activity described herein (e.g., ASO activity or RNA structure mimicry).

[0074] In some embodiments, nucleic acids (unmodified or modified nucleic acids) are further described by their nucleobase sequence. A nucleobase sequence can be defined by Watson-crick base pairing such that certain modifications may yield a different nucleobase chemical formula yet maintain (or even improve) base pairing, complementation, and/or formation of secondary and tertiary structures.

[0075] In some embodiments, nucleic acids have a nucleobase sequence that is configured to bind to a target nucleic acid, such as for use in a hybridization assay.

Accordingly, in some embodiments, a nucleic acid comprises a sequence that is at least partially complementary to a sequence of a target nucleic acid. In various embodiments, the nucleobase sequence of a region or an entire length of a nucleic acid is at least 50% complementary to a sequence of a target nucleic acid, at least 60% complementary to a sequence of a target nucleic acid, at least 70% complementary to a sequence of a target nucleic acid, at least 80% complementary to a sequence of a target nucleic acid, at least 90% complementary to a sequence of a target nucleic acid, at least 95% complementary to a sequence of a target nucleic acid, at least 99% complementary to a sequence of a target nucleic acid, or 100% complementary to a sequence of a target nucleic acid.

[0076] In some embodiments, nucleic acids have a nucleobase sequence that is configured to form a particular secondary structure, such as an aptamer. Accordingly, in some embodiments, a nucleic acid comprises a sequence that is at least partially identical to a sequence of a particular nucleic acid structure. In various embodiments, the nucleobase sequence of a region or an entire length of a nucleic acid is at least 50% identical to a sequence of a particular nucleic acid structure, at least 60% identical to a sequence of a particular nucleic acid structure, at least 70% identical to a sequence of a particular nucleic acid structure, at least 80% identical to a sequence of a particular nucleic acid structure, at least 90% identical to a sequence of a particular nucleic acid structure, at least 95% identical to a sequence of a particular nucleic acid structure, at least 99% identical to a sequence of a particular nucleic acid structure, or 100% identical to a sequence of a particular nucleic acid structure.

[0077] Any method for synthesizing nucleic acids can be utilized, including chemical synthesis methods, in vitro synthesis methods, and in vivo synthesis methods. In some implementations, nucleic acids are synthesized via a nucleic acid synthesizer utilizing a synthesis protocol (e.g., phosphoramidite synthesis). In some implementations, an in vitro polymerase reaction is performed (e.g., polymerase chain reaction; e.g., in vitro pol III RNA synthesis). In some implementations, nucleic acids are expressed using an in vivo expression system (e.g., *E. coli* expression system). In any synthesis reaction, the resulting nucleic acid synthesis products can be purified and prepared for downstream use (e.g., as a therapeutic).

[0078] Splitting or dividing nucleic acids into two or more pools can be done by any appropriate method. In some embodiments, division of synthesis and/or molecular processing is performed by performing two or more batch or pool synthesis protocols, where each nucleic species is to be synthesized within a batch or a pool based on a determination of yield efficiency as described herein (e.g., see Fig. 1). In some implementations, controlled-pore glass (CPG) media is utilized for bulk nucleic acid species synthesis. In some implementations, multiplex microarray assembly is utilized for bulk nucleic acid species synthesis (see, e.g., S. Kosuri and G. M. Church, *Nat Methods*. 2014 May;11(5):499-507, the disclosure of which is hereby incorporated by reference). In some embodiments, division of synthesis and/or molecular processing is performed by molecular process for separating nucleic acid species into pools. One example of a molecular process for separating nucleic acids is dial-out PCR (see, e.g., J. J. Schwartz, et al., *at Methods*. 2012 Sep;9(9):913-5, the disclosure of which is hereby incorporated by reference).

[0079] Dial-out PCR is a method that tags nucleic acid with a sequence tag, and then separates the molecules based on the tag received. To begin, nucleic acid species comprising sequences of interest can be synthesized in bulk, which can be synthesized (for example) using of a multiplex microarray assembly system. Each sequence can be synthesized with a sequence tag at ends of each fragment. Nucleic acid species molecules to be processed in a specific batch or pool can all share the same sequence tag. As described herein, the nucleic acid species having similar efficiency can be batched or pooled, and thus can have the same sequence tag for dial-out PCR. For retrieval, primers complementary to the tags of desired sequences are utilized for dial-out PCR, amplifying only the nucleic acid species of a particular batch or pool. Multiple dial-out PCRs can be performed for each batch or pool, amplifying only the nucleic acid species of a particular batch or pool via the added tags.

[0080] One example of a downstream application that can be performed is mutational profiling sequencing (MaP-seq), which is a method to assess secondary conformation and/or protein binding of nucleic acid species, such as ssDNA or RNA. Generally, for a MaP-seq method to be successful, it is ideal to have several hundred copies (e.g., greater

than 200 copies) for each sequence that is being assessed. And for a high throughput method having greater than 200,000 unique species, a process capable to synthesize and amplify at least 200 copies of 200,000 unique nucleic acid is necessary, such as those described herein. In fact, the methods described herein allow for very high-throughput processing, as millions of copies of millions of unique species can be synthesized and amplified by separation into two or more batches or pools (which can be subsequently combined to yield one large pool).

[0081] To perform MaP-seq, an experimental condition is applied to a pool of diverse nucleic acid species to determine the effect of the experimental condition. In some implementations, the experimental condition is to determine the effect of the condition on secondary structure. In some implementations, the experimental condition is to determine the effect of the condition on binding of RNA proteins. A chemical can be utilized can be added to convert unprotected and unpaired nucleosides. For example, dimethyl sulfate can be utilized to methylate unprotected and unpaired adenines and cytosines. Other chemical compounds that can be utilized include methyl iodide and dimethyl carbonate A MaP-seq protocol will further utilize a reverse transcriptase to yield complement. The reverse transcriptase (or polymerase if ssDNA) can have an error rate when encountering methylated adenines and cytosines, resulting in incorrect incorporation of incorrect sequences at methylated sites. Deep sequencing is performed to identify which adenines and cytosines are protected or paired, allowing identification of secondary structure of and/or protein binding to the nucleic acid species.

EXAMPLES

[0082] Biological data support the systems and methods of treatment for enhancing bulk synthesis and/or molecular processing of diverse nucleic acid species. In the following examples, data describing computational analysis of nucleic acid synthesis are provided, indicating that the systems and methods described herein can be utilized to improve yield of bulk synthesis of nucleic acid species.

THOUGHTFUL PRE-SPLITTING OF DNA SYNTHESIS POOLS TO RESOLVE THE UNIFORMITY PROBLEM IN HIGH THROUGHPUT MOLECULAR BIOLOGY EXPERIMENTS

Background

[0083] The synthesis of large numbers of RNA and DNA molecules is relevant to biological, biotechnical and medical research, ranging from CRISPR screens to designing RNA and protein medicines to DNA storage. Current techniques for the synthesis of a nucleic acid sequence library rely on synthesizing all DNA molecules (or DNA templates for RNA) in one large batch, based on a computer file with the desired sequences. However, this synthesis process does not guarantee that every sequence is synthesized in equal amounts; indeed, some sequences synthesize much more readily than others. In addition, this synthesis process is ignorant of down-stream steps like PCR amplification, library cloning, transcription in vitro or in cellulo, and/or reverse transcription. Due to biases in synthesis and down-stream steps shared by most modern protocols, the ending pool is dominated by a few ‘winner’ sequences.

[0084] In assessments performed with RNA pools, these steps leave up to 90% of the original library with too few molecules to be probed by, e.g., Illumina sequencing. In a recent workflow seeking high throughput data for training artificial intelligence models for RNA structure, for example, we sought to synthesize 1 million sequences, but achieved high quality data for 10%, leaving 900,000 sequences unprobed – an enormous financial waste. This same non-uniformity affects CRISPR screens where non-uniform transcription and reverse transcription of guide RNA libraries lead to dropout of precious data, often with more than half the pool lost.

[0085] Here we propose a concept to mitigate issues with library imbalance that plague all modern molecular biology protocols. We noted that for large synthesis runs, commercial synthesis providers typically split into two or more pools, but this split is done randomly. We propose that thoughtfully splitting the data is able to yield a much greater number of sequences that have a useful number of molecules synthesized, making the process of library synthesis more uniform and increasing the effective amount of data generated by the process.

Description and Proof of Concept

[0086] Suppose we had an oracle that could, for a given nucleic acid sequence, predict the relative number of reads it would yield when synthesized. With such an oracle, one could pass to it each sequence in the library, and get an understanding of which sequences synthesize in large amounts, and which are likely to not yield great numbers of molecules. This information can then be used to split the library into two separate pools, each of which are synthesized independently. By placing the difficult-to-synthesize sequences in their own pool, we increase the chances that they yield good numbers of molecules, since now they are no longer competing with the sequences with synthesize readily.

[0087] We first demonstrate this in the case of a perfect oracle. In Fig. 2, we show the effect of splitting the library into two pools, one which contains all sequences that yielded fewer than T molecules, where T is some threshold, and then all remaining sequences. We vary T on the x-axis, and on the y-axis we record the number of sequences that yielded more than a given number of molecules, in this case 1,000.

[0088] The data used to generate this plot comes from an actual RNA experimental run with a library of 1 million sequences, and that yielded sequencing information for around 1 billion total molecules, where we hoped to achieve 1000 reads for each molecule. The red horizontal line plots the actual number of sequences in the run that had more than 1000 molecules synthesized. As we can see, if the library had been split with such an oracle so that all sequences with less than 6,000 predicted molecules were placed in their own separate pool, the number of sequences with an appreciable number of molecules would have been increased. Indeed, at the optimal split threshold, the number of such sequences would have been double, which would have the effect of doubling the effective amount of data with almost no extra cost.

[0089] Such an oracle could come in many forms. One form would be to use experimental verification. The desired library could be synthesized on a smaller 'pilot' scale, and the number of molecules per sequence in this small-scale experiment would then be used as the oracle to split the full run. The results of this would be similar to in Figure 1, since the ratios of molecules generated will remain the same as the total number

of molecules synthesized is varied. Even if the pilot experiment cost as much as the final experiment, this doubling in cost could still be economical if the loss in useable data incurred in the first experiment is worse than 50%.

[0090] If data on past runs is available, then this can instead be used as the oracle, removing the need to perform this experiment. Another method which avoids experimental analysis is to train a machine learning algorithm to predict the relative number of reads. As a proof of concept, we have replicated Fig. 2, but where the oracle is instead a linear regression model trained on the same dataset. In this case, we split the dataset not by the true reads, but by the predicted reads. The results are in Fig. 3. The two vertical lines correspond to the two best splits, as predicted by the model. We see that, even with a simple linear model, we can perform either at par or better than a random split, with the first best split yielding around 10,000 additional sequences above the threshold of 1,000. A more complex sequential model, such as a transformer or LSTM, may be able to achieve a much more accurate prediction of the reads, and hence an even greater increase in the yield over a random split. Training a larger model will require larger data sets, which could be collected by synthesis companies or other service companies who make their models available to customers in exchange for their data.

[0091] One important insight that led to this result is that a simple linear regression model that tries to predict the number of reads is not effective. Instead, the model predicts the logarithm of the number of reads. This aids the model in fitting to the data, since the data spans multiple scales which otherwise cannot be fit well by a linear model. Indeed, the model needs only to accurately predict the order of magnitude of molecules per sequence in order to generate a good split of the library.

[0092] We also note that for most modern protocols, the steps that generate final sequencing reads are often shared across diverse protocols. In our example above, we synthesized RNA libraries for a structure mapping assay, but the same steps of PCR, in vitro transcription, reverse transcription, and further amplification are shared for CRISPR screens which are also based on RNAs. For synthesis vendors who wish to provide a useful product, providing an AI oracle that is trained for "RNA" will be useful across all RNA research and development. Similarly, an AI oracle for "E. coli", "human", "yeast", etc,

which would takes into account in-cell transcription and translation rates in those organisms based on a few high-throughput experiments would allow for future researchers to automatically get thoughtfully pre-split pools for their applications.

[0093] This splitting idea can with no effort be extended to splitting into more than two pools, which would yield an even greater increase in the number of effective datapoints with only a small amount of extra effort on the side of the synthesis company and on the side of the experimenter. Indeed, as noted above, DNA pools are already delivered as (randomly split) subpools, so a 'thoughtful split' has no added cost to synthesis companies except handling the additional files and separate quality control of subpools (a negligible cost). Researchers who receive the subpools can choose to process their subpools in parallel experiments on the subpools, which can then be sequenced in separate lanes in sequencing flowcells, or barcoded and pooled for running in single lanes on sequencing flowcells. If parallel experiments on subpools are undesirable, researchers can pre-mix the subpools, but increase the concentration of each subpool in inverse proportion to the predicted read number from the oracle.

IMPROVED MAP-SEQ UNIFORMITY VIA SEQUENCE DROPOUT PREDICTION

MaP-Seq at scale

[0094] MaP-Seq experiments exploit the fact that certain chemical modifications of RNA induce mutations or deletions when reverse transcribed by the right enzyme, which can then be detected via sequencing (Fig. 4). The distribution of the location of these mutations sufficiently correlates with the underlying structure to prove useful for structure prediction.

[0095] Exploratory experiments have shown that millions of unique RNA molecules can be probed simultaneously. While a boon for deep learning applications, it comes at a significant cost to uniformity. While the read depths when probing two thousand sequences showed about 10 dropouts, the read depths when probing one million sequences had nearly 200,000 dropouts (Fig. 5).

Dropout prediction and reduction

[0096] The mechanism driving the dropout of sequences during chemical probing is unknown, but varying enzymatic affinities of the different sequences is a working hypothesis. Under this hypothesis, it is natural to try to split the high-affinity and low-affinity sequences into separate experiments, or sub-libraries, lowering the affinity delta and increasing the uniformity. The requisite sequences for each experiment can be extracted via adding primer-binding sequences to the ends of sequences for dial-out PCR (Fig. 6). The dial-out primer-binding sequences added to the end of each sequence can be determined by predicting expected sequence depth.

[0097] Three models were trained for the task of predicting read depth:

1. An LSTM without access to structural information,
2. An LSTM with access to the predicted secondary structure, and
3. A fine-tuned copy of the RibonanzaNet transformer, which had previously been pre-trained on chemical probing profiles.

For more on the RibonanzaNet transformer, see. S. He, et al., bioRxiv [Preprint]. 2024 Jun 11:2024.02.24.581671, the disclosure of which is hereby incorporated by reference.

[0098] The models were trained on the log-read depths associated with a training set of 700,000 sequences and a validation set of 300,000 sequences. The simple heuristic of counting the GC-content of the longest stem was employed as a benchmark.

Benchmarking

[0099] Five diverse unseen libraries were used for testing (Fig. 7). In all tests, the finetuned copy of RibonanzaNet edged out the competitors. Some libraries exhibit dropout with a structural dependence, whereas others do not.

Sub-library testing

[0100] The fine-tuned RibonanzaNet model was used to split a sequence library into three sub-libraries, which were subsequently chemically probed in three separate experiments. Fig. 8 shows the performance of the model in predicting the true read depths, and a confusion matrix indicating the accuracy of the model in choosing the ideal

sub-libraries.

[0101] When compared against a control experiment with no splitting of the libraries, the separation of the library into three sub-libraries led to the rescue of the majority of sequences which dropped out in the control experiment (Fig. 9). The split experiment also has a median read depth twice that of the control.

WHAT IS CLAIMED IS:

1. A method to improve generation of diverse pools of nucleic species, comprising:
providing a set of nucleic acid species to be synthesized, wherein each nucleic acid species within the set has a unique sequence;
determining a yield efficiency for each nucleic acid sequence of the set of sequences of nucleic acid species to be synthesized; and
generating the nucleic acid species in two or more batches, wherein each sequence of the set of sequences of nucleic acid species within each batch is determined based on its determined yield efficiency.
2. The method of claim 1, wherein the step of synthesizing the nucleic acid species in two or more batches comprises purification.
3. The method of claim 1, wherein the step of synthesizing the nucleic acid species in two or more batches comprises downstream processing.
4. The method of claim 1, 2, or 3, wherein the step of determining a yield efficiency for each sequence comprises utilizing data from a prior synthesis reaction or data from a prior purification or data from prior downstream processing.
5. The method of any one of claims 1-4, wherein the step of determining a yield efficiency for each sequence comprises performing a pilot synthesis reaction or a pilot purification or a pilot run of downstream processing.
6. The method of any one of claims 1-5, wherein the step of determining a yield efficiency for each sequence comprises entering each sequence of the set of sequences of nucleic acid species to be synthesized into a trained predictive model.
7. The method of any one of claims 1-6, wherein the number of nucleic acid species in the set is at least 1000.

8. The method of claim 7, wherein the number of nucleic acid species in the set is at least 100,000.
9. The method of any one of claims 1-8, wherein each nucleic acid species of the set is between 5 and 500 nucleosides.
10. The method of any one of claims 1-9, wherein each nucleic acid species of the set have a number of nucleosides not more than 500% of the shortest nucleic acid species of the set.
11. The method of any one of claims 1-10, wherein each nucleic acid species of the set have an equivalent number of nucleosides.
12. The method of any one of claims 1-11, wherein yield efficiency comprises the efficiency of nucleic acid species chemical synthesis.
13. The method of claim 12, wherein the chemical synthesis comprises phosphoramidite synthesis.
14. The method of claim 13, wherein the chemical synthesis comprises multiplex microarray assembly.
15. The method of any one of claims 1-14, wherein yield efficiency comprises the efficiency of one or more molecular processes.
16. The method of claim 15, wherein the one or more molecular processes comprises one or more of: replication, transcription, reverse transcription, translation, amplification, complement-based hybridization, cofactor binding, or introduction into cell.

17. The method of any one of claims 1-16, wherein the step of generating the nucleic acid species in two or more batches comprises performing an independent chemical synthesis reaction for each batch of the two or more batches.

18. The method of any one of claims 1-17, wherein the step of generating the nucleic acid species in two or more batches comprises performing dial-out PCR.

19. The method of any one of claims 1-18 further comprising performing a downstream application, wherein the downstream application comprises one or more of: antisense oligomer analysis, RNA-interference molecule analysis, short-hairpin RNA molecule analysis, miRNA molecule analysis, mutagenesis with guide-RNA molecules for CRISPR, aptamer analysis, in situ hybridization molecules, microarray analysis, mutated sequence analysis, assembly of genes, assembly of genomes, nucleic-acid based information storage, generation of protein medicines encoded by nucleic acids, generation of nucleic acid medicines, generation of sequencing library, or hybridization capture.

20. The method of any one of claims 1-19 further comprising performing mutational analysis profiling sequencing (MaP-seq).

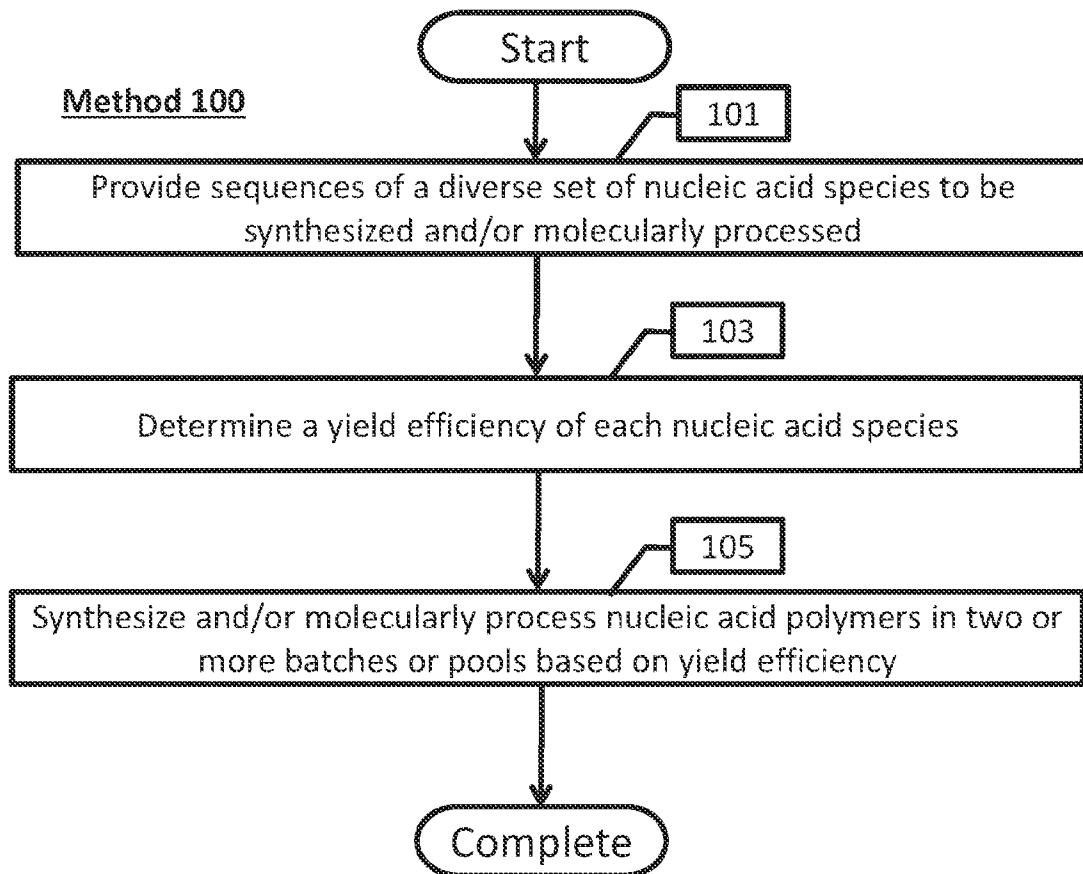
Fig. 1

Fig. 2

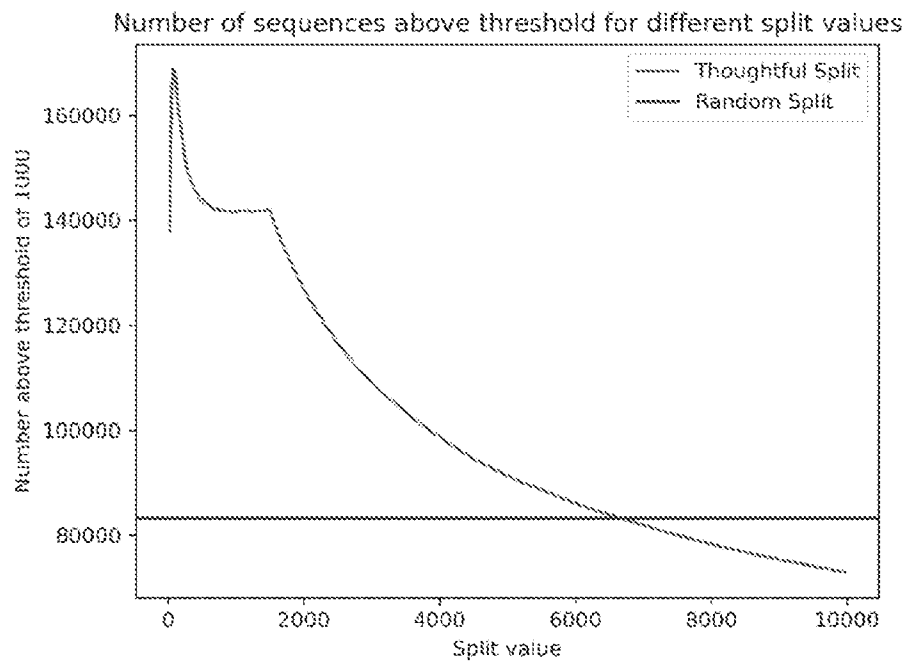


Fig. 3

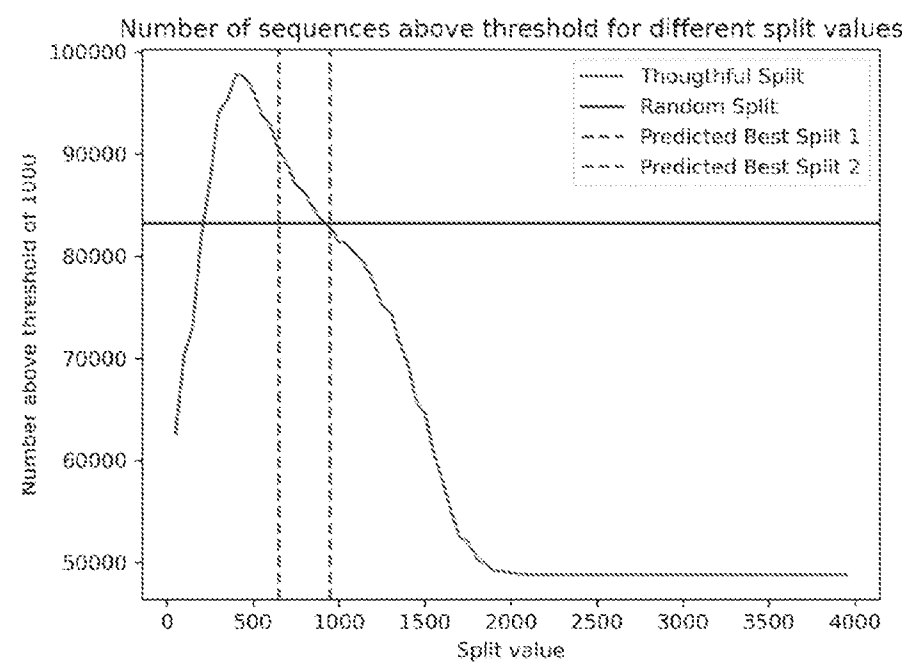


Fig. 4

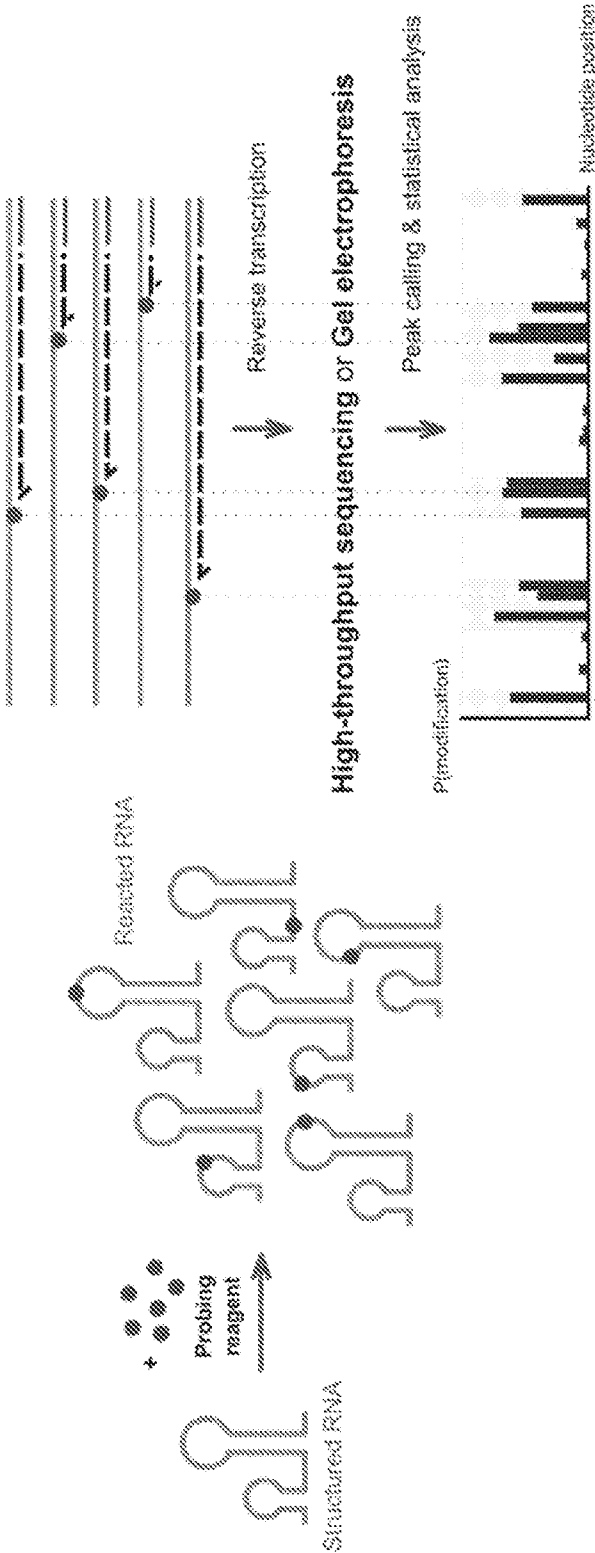


Fig. 5

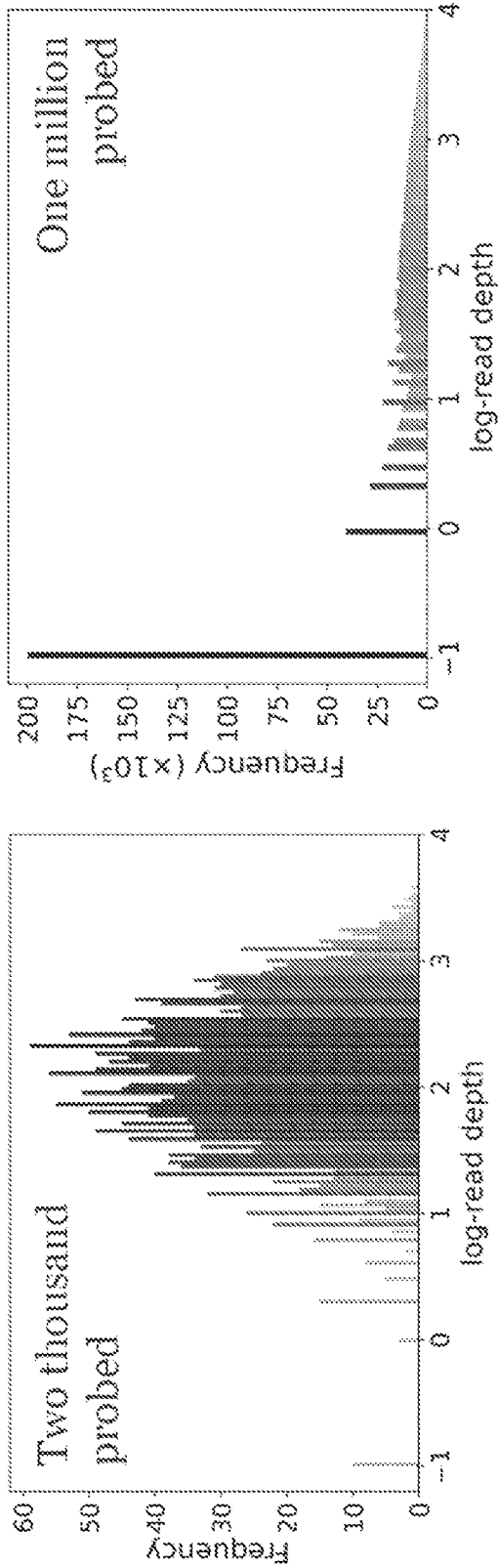


Fig. 6

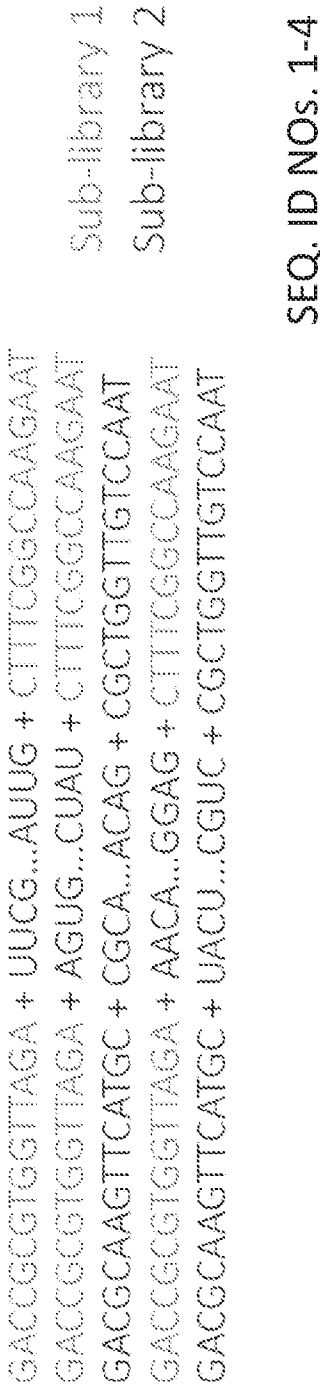


Fig. 7

Sequence Library			Correlation with log-read depth			
Size	Length	Type	Longest-Stem GC Content	LSTM Without Structure	LSTM With Structure	Fine-tuned RibonanzaNet
2,496	226	Natural	0.24	N/A	N/A	0.76
1,000,000	207	Mixed	−0.54	0.74	0.76	0.79
2,000	307	Synthetic	−0.72	0.22	0.89	0.92
6,000	457	Synthetic	−0.28	0.29	0.53	0.63
14,049	207	Synthetic	−0.31	N/A	N/A	0.51

Fig. 8

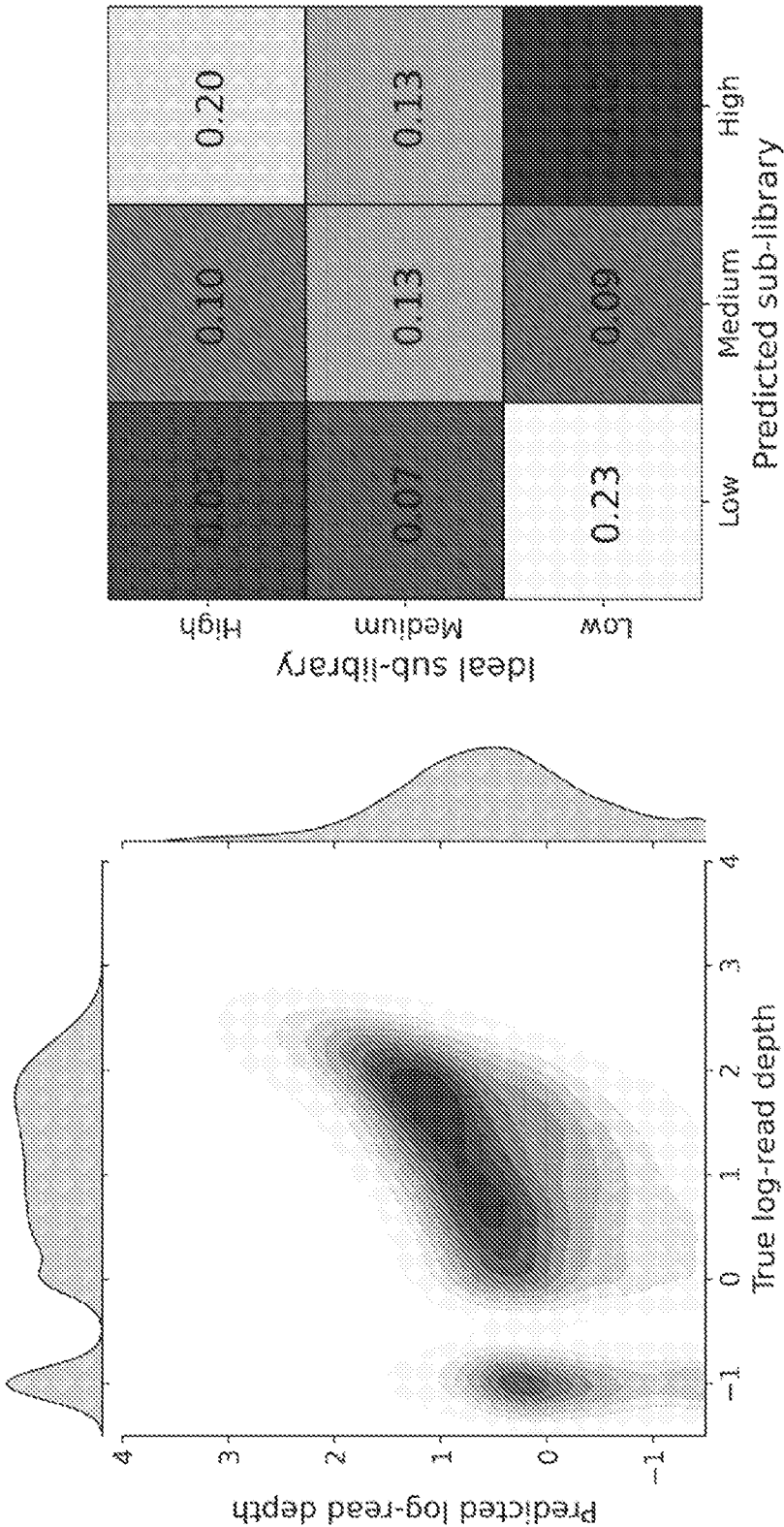
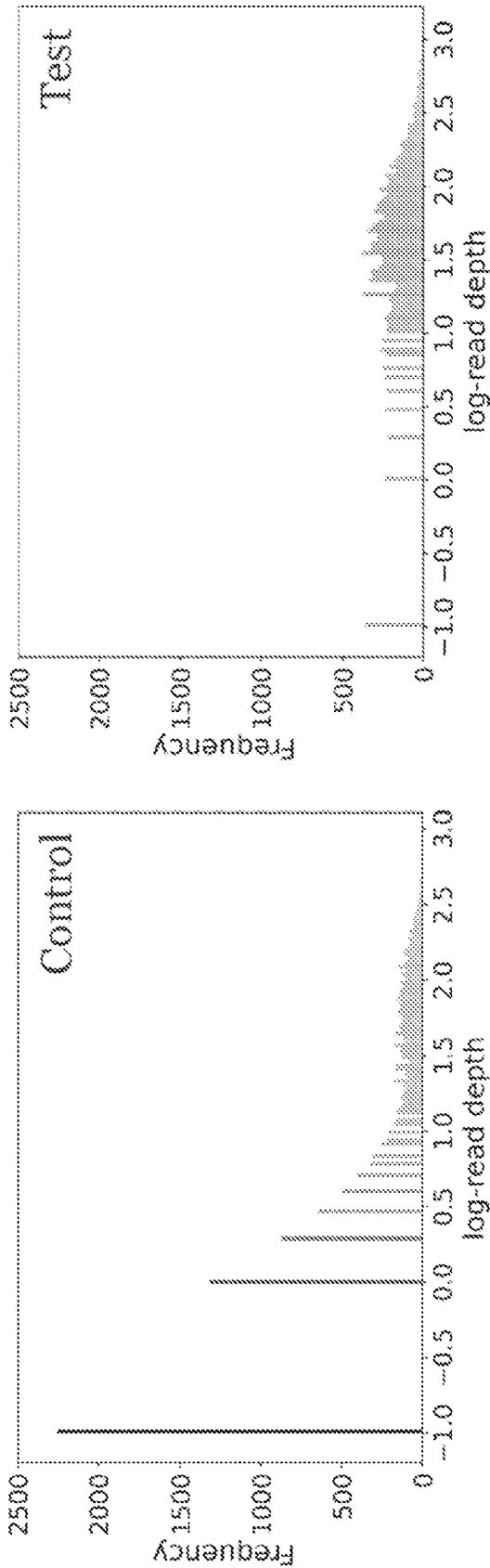


Fig. 9



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/061133

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
 - a. ☒ forming part of the international application as filed.
 - b. ☐ furnished subsequent to the international filing date for the purposes of international search (Rule 13ter.1(a)),
☐ accompanied by a statement to the effect that the sequence listing does not go beyond the disclosure in the international application as filed.
2. ☐ With regard to any nucleotide and/or amino acid sequence disclosed in the international application, this report has been established to the extent that a meaningful search could be carried out without a WIPO Standard ST.26 compliant sequence listing.
3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/061133

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.: **5-20**
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/061133

A. CLASSIFICATION OF SUBJECT MATTERIPC: **C12Q 1/6869** (2025.01); **C40B 30/04** (2025.01); **G16B 30/00** (2025.01); **G16B 40/20** (2025.01); **C12Q 1/6806** (2025.01); **C12Q 1/6809** (2025.01); **C12N 15/113** (2025.01)CPC: **C12Q 1/6869**; **C12Q 1/6806**; **C40B 30/04**; **G16B 40/20**; **C12Q 1/6809**; **G16B 30/00**; **C12N 15/113**

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.
A	WO 2023/028618 A1 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 02 March 2023 (02.03.2023) entire document	1-4
X	Rangan et al. RNA structure landscape of <i>S. cerevisiae</i> introns. bioRxiv. 23 July 2022. Pgs. 1-38, [retrieved on 07.02.2025]. Retrieved from the internet: <URL: https://doi.org/10.1101/2022.07.22.501175 > entire document	1-4
A	ANDREASSON et al. "Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular switches" PNAS, USA. 26 April 2022. Vol. 119, No. 18:e2112979119, Pgs. 1-7 entire document	1-4



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

"D" document cited by the applicant in the international application

"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 February 2025 (07.02.2025)

Date of mailing of the international search report

14 February 2025 (14.02.2025)

Name and mailing address of the ISA/US

COMMISSIONER FOR PATENTS
MAIL STOP PCT, ATTN: ISA/US
P.O. Box 1450
Alexandria, VA 22313-1450
UNITED STATES OF AMERICA

Facsimile No. **571-273-8300**

Authorized officer

TAINA MATOSTelephone No. **571-272-4300**